# Migration, culture, and inequalities in algorithmically-mediated societies

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science of
Saarland University

by
Carolina Coimbra Vieira

Saarbrücken
2025

| | |
|---|---|
| **Date of Colloquium:** | April 27, 2026 |
| **Dean of Faculty:** | Univ.-Prof. Dr. Jan Reineke |
| | |
| **Reporter:** | |
| **First Reviewer:** | Prof. Dr. Krishna P. Gummadi |
| **Second Reviewer:** | Prof. Dr. Emilio Zagheni |
| **Third Reviewer:** | Prof. Dr. Ingmar Weber |
| **Fourth Reviewer:** | Prof. Dr. Meeyoung Cha |
| **Chair of the Examination Board:** | Prof. Dr. Isabel Valera |
| **Scientific Assistant:** | Dr. Johnnatan Messias Peixoto Afonso |

**Declaration of original authorship**

I hereby declare that this dissertation is my own original work except where otherwise indicated. All data or concepts drawn directly or indirectly from other sources have been correctly acknowledged. This dissertation has not been submitted in its present or similar form to any other academic institution either in Germany or abroad for the award of any other degree.

**Eidesstattliche Versicherung**

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, 25.10.2025
Carolina Coimbra Vieira (signed)

# Abstract

The widespread use of online platforms has transformed how people interact, express opinions, and access information. As a byproduct of these digital interactions, users continuously generate large-scale digital trace data, opening up unprecedented opportunities for social science research. In this thesis, digital trace data from Facebook, Twitter (now X), TikTok, and Wikipedia are used to address key questions ranging from cross-country cultural similarity to large-scale human migration, patterns of gender and social inequalities, and online behavior on algorithmically mediated platforms. Across six studies, it demonstrates how digital trace data can complement or overcome the limitations of traditional data sources, which are often slow, costly, or unavailable. The first two chapters link culture and migration by introducing a novel measure of cross-national cultural similarity based on Facebook users' interests and showing that this measure improves predictions of international migration flows beyond standard economic and geographic factors. The third chapter proposes a novel use of Wikipedia page view data to detect information-seeking behavior during crises, showing that readership patterns can serve as a near real-time proxy for forced migration, as demonstrated during the 2022 Ukrainian refugee crisis. Addressing global inequality, the fourth chapter uses Facebook data to examine gender balance among users interested in Science, Technology, Engineering, and Mathematics (STEM) fields worldwide and provides a detailed analysis of gender disparities across age and education groups in Brazil. The fifth chapter investigates social vulnerability by analyzing Twitter data to characterize missing children in Guatemala, offering insights that complement scarce and delayed official statistics. Finally, the sixth chapter evaluates user watching behavior on short-form video platforms by analyzing donation-based TikTok data, revealing the limited predictability of watching behavior and highlighting methodological challenges arising from increasing platform data access restrictions. Collectively, these studies demonstrate how digital trace data can advance the measurement and understanding of complex social phenomena, particularly in regions of the Global South and in crisis contexts where conventional data are sparse or delayed. By integrating computational methods with theoretical perspectives from the social sciences, this thesis advances the field of computational social science and highlights the societal value of using digital trace data for research.

# Zusammenfassung

Die weit verbreitete Nutzung von Online-Plattformen hat die Art und Weise verändert, wie Menschen interagieren, Meinungen äußern und auf Informationen zugreifen. Als Nebenprodukt dieser digitalen Interaktionen generieren NutzerInnen kontinuierlich umfangreiche digitale Spuren (in Englisch, "digital traces"), die der sozialwissenschaftlichen Forschung beispiellose Möglichkeiten eröffnen. In dieser Arbeit werden digitale Spuren aus Facebook, Twitter (jetzt X), TikTok und Wikipedia verwendet, um wichtige Fragen zu untersuchen, die von länderübergreifenden kulturellen Ähnlichkeiten über groß angelegte Migrationen bis hin zu Mustern geschlechtsspezifischer und sozialer Ungleichheiten sowie Online-Verhalten auf algorithmisch vermittelten Plattformen reichen. In sechs Studien wird gezeigt, wie digitale Spuren Daten die Grenzen traditioneller Datenquellen, die oft langsam, kostspielig oder nicht verfügbar sind, ergänzen oder überwinden können. Die ersten beiden Kapitel verbinden Kultur und Migration, indem sie eine neuartige Messgröße für die länderübergreifende kulturelle Ähnlichkeit auf der Grundlage der Interessen von Facebook-NutzerInnen einführen und zeigen, dass diese Messgröße die Vorhersagen zu internationalen Migrationsströmen über die üblichen wirtschaftlichen und geografischen Faktoren hinaus verbessert. Das dritte Kapitel schlägt eine neuartige Verwendung von Wikipedia-Seitenaufrufdaten vor, um das Informationssuchverhalten in Krisenzeiten zu erfassen, und zeigt, dass Lesermuster als nahezu Echtzeit-Proxy für Zwangsmigration dienen können, wie während der ukrainischen Flüchtlingskrise 2022 demonstriert wurde. Das vierte Kapitel befasst sich mit globaler Ungleichheit und untersucht anhand von Facebook-Daten das Geschlechterverhältnis unter NutzerInnen, die sich weltweit für die Bereiche Wissenschaft, Technologie, Ingenieurwesen und Mathematik (STEM) interessieren. Außerdem enthält es eine detaillierte Analyse der geschlechtsspezifischen Unterschiede in verschiedenen Alters- und Bildungsgruppen in Brasilien. Das fünfte Kapitel untersucht soziale Vulnerabilität durch die Analyse von Twitter-Daten, um vermisste Kinder in Guatemala zu charakterisieren, und liefert Erkenntnisse, die die spärlichen und verzögerten offiziellen Statistiken ergänzen. Schließlich bewertet das sechste Kapitel das Verhalten der NutzerInnen auf Kurzvideo-Plattformen. Mithilfe der Analyse spendenbasierter TikTok-Daten werden die begrenzte Vorhersagbarkeit des Verhaltens aufgezeigt und methodische Herausforderungen hervorgehoben, die sich aus den zunehmenden Zugangsbeschränkungen zu Plattformdaten ergeben. Zusammen zeigen diese Studien, wie die Analyse von digitalen Spuren die Messung und das Verständnis komplexer sozialer Phänomene verbessern können, insbesondere in Regionen des Globalen Südens und in Krisensituationen, in denen herkömmliche Daten spärlich oder verzögert verfügbar sind. Durch die Integration von computergestützten Methoden mit theoretischen Perspektiven aus den Sozialwissenschaften treibt diese Arbeit das Gebiet der computergestützten Sozialwissenschaften voran und unterstreicht den gesellschaftlichen Wert der Nutzung digitaler Spuren für die Forschung.

# Publications

**Parts of this thesis have appeared in the following publications:**

- "Forced Migration and Information Seeking Behavior on Wikipedia: Insights from the Ukrainian Refugee Crisis". **Vieira, Carolina C.**; Sanlitürk, Ebru; Zagheni, Emilio. In *Proceedings of the 20$^{th}$ International AAAI Conference on Web and Social Media (ICWSM)*, Los Angeles, USA, May 2026.

- "Exploring the Limits of Predicting User Watching Behavior with Short-Form Videos on TikTok". **Vieira, Carolina C.**; Mousavi, Sepehr; Ayalon, Oshrat; Dash, Abhisek; Gummadi, Krishna P.; Zannettou, Savvas. In *Companion Proceedings of the 18$^{th}$ ACM Web Science Conference (WebSci)*, Braunschweig, Germany, May 2026.

- "Characterizing Global Gender Gaps in STEM using Facebook data". **Vieira, Carolina C.**; Vasconcelos, Marisa. In *Proceedings of the 20$^{th}$ International Society of Scientometrics and Informetrics Conference (ISSI)*, Yerevan, Armenia, June 2025.

- "The Value of Cultural Similarity for Predicting Migration: Evidence from Food and Drink Interests in Digital Trace Data". **Vieira, Carolina C.**; Lohmann, Sophie; Zagheni, Emilio. In *Population and Development Review*, Wiley. Volume 50, Issue 1, Pages 149-176, March 2024.

- "Desaparecidxs: characterizing the population of missing children using Twitter". **Vieira, Carolina C.**; Alburez-Gutierrez, Diego; Nepomuceno, Marilia R.; Theile, Tom. In *Proceedings of the 14$^{th}$ ACM Web Science Conference (WebSci)*, Barcelona, Spain, June 2022.

- "The interplay of migration and cultural similarity between countries: Evidence from Facebook data on food and drink interests". **Vieira, Carolina C.**; Lohmann, Sophie; Zagheni, Emilio; de Melo, Pedro O. V.; Benevenuto, Fabrício; Ribeiro, Filipe N. In *PLoS ONE*, Plos. 17(2): e0262947, Pages 1–21, February 2022.

- "Using Facebook Ads Data to Assess Gender Balance in STEM: Evidence from Brazil". **Vieira, Carolina C.**; Vasconcelos, Marisa. In *Companion Proceedings of the 30$^{th}$ The Web Conference (WWW)*, Ljubljana, Slovenia, April 2021.

**Additional publications during the PhD:**

- "Beyond Sentiment Analysis with ChatGPT: Classifying Authors' Perspectives on Russian Topics". **Vieira, Carolina C.**; Chechik, Elena; Di Césare, Victoria. In *Proceedings of the 20th International Society of Scientometrics and Informetrics Conference (ISSI)*, Yerevan, Armenia, June 2025.
- "Constructing Social Vulnerability Indexes with Increased Data and Machine Learning Highlight the Importance of Wealth Across Global Contexts". Zhao, Yuan; Paul, Ronak; Reid, Sean; **Vieira, Carolina C.**; Wolfe, Chris; Zhang, Yan; Chunara, R. In *Social Indicators Research*, Springer. Volume 175, Pages 639–657, July 2024.
- "Using Facebook and LinkedIn data to study international mobility". **Vieira, Carolina C.**; Fatehkia, Massoomali; Garimella, Kiran; Weber, Ingmar; Zagheni, Emilio. In: Salah, A. A.; Eren Korkmaz, E.; Bircan, T. (Eds.): *Data science for migration and mobility*. Oxford: Oxford University Press. November 2022.
- "Evaluating Digital Polarization in Multi-Party Systems: Evidence from the German Bundestag". Chin, Amber; **Vieira, Carolina C.**; Kim, Jisu. In *Proceedings of the 14th ACM Web Science Conference (WebSci)*, Barcelona, Spain, June 2022.

To Gucci (my dog, not the brand)

# Acknowledgements

This chapter of my life, spent during my PhD, has been the most challenging and rewarding journey of both my academic and, perhaps, personal life. This would certainly not have been possible without the support, encouragement, and inspiration of many people.

First, I was fortunate to have not just one, but two great supervisors: Emilio Zagheni and Krishna Gummadi. With their different personalities but equally brilliant minds, they complemented each other and guided me throughout this process. It has truly been a pleasure and a privilege to have them as my supervisors.

I am also immensely grateful to my mentors, collaborators, and co-authors. From my early scientific life in Brazil to the end of my PhD in Germany, I thank all the professors, mentors, collaborators, and co-authors who encouraged me to pursue this path. A special thanks to Sophie Lohmann and Ebru Şanlıtürk, with whom I learned and grew a lot. Thank you for being such great mentors and friends! To all my colleagues at MPIDR, MPI-SWS, and Saarland University, thank you for all the support, discussions, and idea exchanges. More than that, to the MPIDR crew I interacted with regularly, whether in the office, at the Department of Digital and Computational Demography (DCD), or beyond, thank you for your support, friendship, and for creating such an enjoyable and welcoming atmosphere!

My gratitude also extends to the institutional support of Saarland University, MPIDR, and MPI-SWS. At MPI-SWS, I extend my thanks to the staff, especially the secretarial team, for their assistance and efforts during my stays. At MPIDR, I am also grateful to the staff members who make the institute such a special place: receptionists, press, IT, administration and secretarial team. Thank you for everything you do, often behind the scenes, to make research possible.

This PhD journey also took me to many scientific events. I am grateful for the opportunities to present my work, exchange ideas, and learn from others. I am also deeply thankful to Emilio Zagheni for giving me the chance to mentor students, serve on selection committees, and teach workshops and summer schools, sharing knowledge while learning just as much in return.

On a personal note, I have been incredibly lucky to be surrounded by people who brought joy, love, and energy into this journey. I feel very grateful to call many of my colleagues friends. Special thanks to Donata, Maria, and Su, who started this journey with me in 2020 and shared many academic and personal moments. Thank you for your friendship, support, and all the moments we shared!

To my Brazilian friends at MPIDR (whether Brazilian by blood or by heart) and to my office mates, thank you for bringing even more excitement to the workplace and my life! Special thanks to Maria Laura and Amanda for being my safe space, for all the patience, support, amazing Brazilian food, and, of course, the countless special moments we shared!

To Mathis, a special thank you for all the support, love, and understanding. My gratitude also extends to your family for bringing me joy, warmth, and a sense of belonging while I was far from my own.

To my family, who have supported me with unconditional love and patience, I am forever grateful. Your belief in me, even from miles away, has been my strongest source of strength. Thank you for all the calls, prayers, and encouraging words! *Obrigada por tudo! Amo vocês!*

Last but most importantly, to Gucci. Thank you for accompanying me everywhere, for bringing joy wherever we go, for the warm welcomes after each business trip, and for your everyday company. Your gentle snores next to me as I wrote these pages were a comforting reminder that I was never alone.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

Over the past decades, the widespread use of online platforms, such as social networks (Adamic and Adar, 2003; Boyd and Ellison, 2007), has fundamentally transformed the way people interact, form social connections, express opinions, and consume information (Bakshy et al., 2015; Benevenuto et al., 2009; Kwak et al., 2010; Myers et al., 2014; Vosoughi et al., 2018). Currently, we live in an algorithmically-mediated society, where algorithms play a central role in shaping human interactions (Gillespie, 2014; Wagner et al., 2021). While algorithms influence user behavior by curating content, personalizing recommendations, and optimizing engagement (Tang et al., 2013; Zou et al., 2019), user behavior simultaneously shapes algorithms, as models continuously adjust based on interactions, clicks, likes, and watch patterns (Cinelli et al., 2021). In this process, users generate vast amounts of digital trace data, offering unprecedented opportunities for social science studies (Lazer et al., 2009).

Traditionally, social science studies have relied heavily on survey-based approaches (Wright et al., 2010). Despite being essential for research, these traditional data sources and methods often face significant limitations, particularly the high costs and long timeframes required to design and implement surveys. In contrast, digital trace data offer real-time, large-scale, and passively collected behavioral information, characterized by high precision and fine-grained detail. As a result, researchers are increasingly leveraging these novel data sources to complement traditional approaches and to generate new insights into social phenomena (Salganik, 2019).

Digital trace data have increasingly been used to gain insights across a wide range of interdisciplinary topics within Computational Social Science. In particular, in the context of this thesis, these data have been applied to studies on culture, migration, gender inequalities, and online behavior. In particular, within cultural studies, digital trace data have been used to capture cultural aspects of countries and to examine cultural diffusion across borders through users' preferences on social media (Obradovich et al., 2020; Vieira et al., 2020). In migration studies, researchers have employed digital trace data to assess migration intentions, both planned, such as for labor markets (Böhme et al., 2020; Perrotta et al., 2022; Vieira et al., 2022b), and unplanned, such as those arising from crises (Anastasiadou et al., 2024; Sanliturk and Billari, 2024). Digital trace data have also been used to estimate migration stocks (Zagheni et al., 2017) and to nowcast or forecast migration flows (Chi et al., 2025). In the context of vulnerable populations, these data

have served to nowcast and forecast displacement after crises (Alexander et al., 2019; Leasure et al., 2023; Rufener et al., 2024), as well as to assess the socio-economic situation of vulnerable groups (Fatehkia et al., 2022). Within inequality research, researchers have drawn on digital trace data to examine gender disparities, particularly using Facebook data to assess gender gaps across different contexts (Fatehkia et al., 2018; Garcia et al., 2018; Jacobs et al., 2024). Finally, digital trace data have also become a subject of study in their own right, with research investigating online behaviors and engagement patterns on viral platforms, such as TikTok (Vombatkere et al., 2024; Zannettou et al., 2024).

Building on the literature across multiple disciplines, this thesis leverages digital trace data to study key societal issues, with a particular focus on culture, migration, gender inequalities, vulnerability, and user engagement. It develops measures of cross-country cultural similarity from online preferences and demonstrates how these measures improve predictions of migration flows. It also introduces an innovative use of Wikipedia data to examine information-seeking behavior during forced migration, focusing on the Ukrainian refugee crisis of 2022. In the area of inequalities, it analyzes Facebook data to assess global gender gaps in Science, Technology, Engineering, and Mathematics (STEM), and further provides an in-depth exploration of gender disparities across age and education groups within Brazil. Research on vulnerability uses Twitter data to characterize the population of missing children in Guatemala, showing how social media data can complement official records in contexts where traditional statistics are scarce. Finally, as platforms increasingly restrict API access, the thesis highlights the potential of alternative approaches such as data donation, drawing on user-donated TikTok data to evaluate the predictability of user watching behavior on short-form video platforms.

These studies demonstrate how digital trace data can advance scientific understanding of society while providing societal value. As a contribution to the discipline, this thesis bridges the fields of social and computational sciences, advances the development of computational social science, and illustrates the potential of interdisciplinary approaches. The contributions of each chapter, as well as the thesis's contributions to the areas studied, are explored in greater detail in the following section outlining the thesis contributions.

## 1.1 Overview of thesis contributions

This thesis makes significant contributions to the study of culture, migration, gender inequalities, and user behavior on algorithmically mediated platforms. The research contributions of each chapter are summarized below.

### 1.1.1 Measuring Cross-country Cultural Similarity: Evidence from Facebook (Vieira et al., 2022c)

The first chapter proposes a scalable methodology to measure cultural similarity between countries using data from the Facebook Advertising Platform, leveraging users' preferences for food and drink as cultural markers. This approach enables the study of the relationship between migration and cultural similarity, providing one of the first large-scale analyses in this area.

The main contributions are the following:

▷ We propose a methodology to measure cultural similarity between countries using Facebook data, providing a cost-effective and scalable alternative to traditional approaches.

▷ We validate the proposed Facebook-based cultural similarity measures against traditional survey-based measures from the World Values Survey (WVS) and against online data from Foursquare.

▷ We develop non-symmetric measures of cultural similarity that better capture asymmetries in cultural adoption across countries.

▷ We present substantive results on cultural similarity across countries, using food and drink as transparent and interpretable cultural markers.

▷ We investigate the relationship between migration and cultural similarity in food and drink preferences using digital trace data.

▷ To promote transparency and facilitate reproducibility, all measures derived from the Facebook data, along with the code to replicate analyses and generate figures, are publicly available (Vieira et al., 2022c).

This chapter is based on Vieira et al. (2022c). The author contributions for this publication are as follows: **Vieira, C.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Lohmann, S.: Conceptualization, Writing – review & editing; Zagheni, E.: Conceptualization, Writing – review & editing; de Melo, P.O.V.: Conceptualization, Methodology, Writing – review & editing; Benevenuto, F.: Conceptualization, Methodology, Writing – review & editing; Ribeiro, F.N.: Conceptualization, Data curation, Writing – review & editing

### 1.1.2 Evaluating the Impact of Cultural Similarity on Migration Prediction (Vieira et al., 2024)

The second chapter assesses the role of culture in shaping international migration flows. Building on the well-established gravity model of migration, we expand the analysis of cultural effects by incorporating measures of cultural similarity derived from surveys (i.e., WVS) and online data sources (i.e., Facebook and Foursquare). In particular, we test the effect of including the Facebook cultural similarity measure based on food and drink interests, developed in Chapter 2. We show that this Facebook-based measure provides predictive power for migration flows comparable to

classic cultural variables, such as shared language and shared history, while also capturing more rapid cultural changes during migration crises.

The main contributions are the following:

▷ We extend the gravity model of migration by incorporating both traditional measures of cultural similarity and those based on food and drink interests derived from social media data, with particular focus on Facebook data.

▷ We provide a stringent test of the Facebook measures' incremental effects, showing that this measure of cultural similarity explains migration flows in a manner comparable to standard predictors such as shared language and shared history.

▷ We highlight the potential of Facebook data as a timely, cost-effective, and scalable approach for capturing cultural change more quickly than traditional surveys, particularly useful for studying migration dynamics in near real time, especially during crises.

▷ To promote transparency and ensure reproducibility, we make all code used in the analysis publicly available (Vieira et al., 2024).

This chapter is based on Vieira et al. (2024). The author contributions for this publication are as follows: **Vieira, C.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Lohmann, S.: Conceptualization, Writing – review & editing; Zagheni, E.: Conceptualization, Writing – review & editing

### 1.1.3 Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia (Vieira et al., 2026b)

The third chapter examines the relationship between online information-seeking behavior and forced migration flows, focusing on the case of Ukrainian refugees following Russia's invasion of Ukraine in 2022. We use Wikipedia Pageview data to capture information-seeking patterns about cities as a proxy for potential destinations. As a proxy for the origin, we analyze views across different language editions, since Wikipedia does not provide the geographic location of pageviews. Our analysis shows that Wikipedia readership reflects Ukrainian refugee movements across Europe, particularly into Poland and Germany, and reveals a temporal association between border crossings into Poland and spikes in Wikipedia views. Specifically, refugee crossings into Poland precede increases in views of Ukrainian-language Wikipedia articles about Polish cities, suggesting that information-seeking surges after displacement. This pattern highlights the reactive nature of information needs during forced migration, in contrast to the pre-departure planning typical of regular labor migration. Moreover, while official applications for protection often lag by weeks, Wikipedia activity responds almost immediately after border crossings, positioning it as a potential near real-time indicator of emerging migration patterns during crises.

The main contributions are the following:

▷ We introduce Wikipedia Pageviews as a novel, timely, and publicly available data source for analyzing information-seeking behavior in the context of forced migration.

▷ We quantify the increase in Wikipedia views in the context of forced migration, showing that for articles about Polish cities hosting large numbers of Ukrainian refugees, pageviews in Ukrainian rise by at least 200% compared to the previous year.
▷ We provide empirical evidence that refugee flows during the Ukrainian crisis are closely associated with increases in Wikipedia views of destination-related articles, both across countries and within cities.
▷ We show that information-seeking during forced migration follows a reactive pattern, with spikes in Wikipedia activity typically occurring after refugee arrivals by about one week, in contrast to the preparatory information-seeking patterns characteristic of traditional labor migration.
▷ We demonstrate that Wikipedia activity responds more immediately than official refugee registration data, highlighting its potential as a near real-time indicator and early-warning system for migration monitoring during crises.
▷ To promote transparency and facilitate scientific reproducibility, we make the code used in our analysis publicly available (Vieira et al., 2026b).

This chapter is based on Vieira et al. (2026b). The author contributions for this publication are as follows: **Vieira, C.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Şanlıtürk, E.: Conceptualization, Methodology, Writing – review & editing; Zagheni, E.: Conceptualization, Writing – review & editing

### 1.1.4 Mapping Global Gender Balance in STEM: Evidence from Facebook (Vieira and Vasconcelos, 2021, 2025)

The fourth chapter presents a large-scale analysis of the global STEM gender gap on Facebook, using data on users' interests across majors. We leverage digital trace data from the Facebook Advertising Platform to examine gender disparities in STEM and non-STEM interests across 198 countries. Our analysis highlights global patterns, regional variations, and demographic-specific gender differences. Additionally, we present a case study on Brazil, one of the largest Facebook markets, to explore STEM gender balance at a more granular level across states, age groups, and education levels. Although the Facebook population is biased toward women, the proportion of men interested in STEM majors is higher than that of women. Within STEM fields, distinct patterns emerge: Life Sciences and Math/Physical Sciences show female dominance, whereas Environmental Science, Technology, and Engineering remain male-dominated. We also examine the impact of educational level and age on interest in majors, showing that in Brazil the gender gap in STEM increases with women's educational attainment and age, consistent with official national data.

The main contributions are the following:
▷ We apply a methodology to measure gender balance using Facebook users' interests across STEM and non-STEM fields, providing a scalable, timely, and cost-effective alternative to assess

the gender gap.
▷ We analyze the gender gap in Facebook users' interests across 198 countries, including an in-depth analysis for Brazil, offering new insights into global and regional variations in STEM and non-STEM gender balance across 142 majors.
▷ We compare Facebook-based gender gap estimates with the Global Gender Gap Report, providing estimates for 48 countries not included in the report and thereby expanding the global evidence base on gender inequality in education and career aspirations.
▷ We show that within STEM fields, interest patterns differ: women show higher interest in Life Sciences and Mathematics, in contrast to male-dominated interests in Engineering and Technology.
▷ We conduct a detailed case study of Brazil, analyzing gender balance across states, age groups, and education levels, and comparing our Facebook-based estimates with national survey data, demonstrating the method's applicability to Global South contexts often overlooked by traditional approaches.
▷ We demonstrate that Facebook interests can serve as a general measure of interest, which can be leveraged to promote STEM majors among female audiences, in particular among those already showing interest in STEM-related topics.

This chapter is based on Vieira and Vasconcelos (2021) and Vieira and Vasconcelos (2025). The author contributions for these publications are as follows: **Vieira, C.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Vasconcelos, M.: Conceptualization, Writing – review & editing

### 1.1.5 Characterizing the Population of Missing Children in Guatemala: Evidence from Twitter (Vieira et al., 2022a)

The fifth chapter characterizes the demographic composition of the population of missing children in Guatemala using data collected from Twitter. We focus on Guatemala, a country that emerged from a bloody 36-year civil war in 1996 and, over the last decades, has experienced increasing levels of violence related to gang activity and drug trafficking. We systematically collect individual-level data from the official Twitter account of *Alerta Alba-Keneth*, including text extracted from images, to provide a near real-time view of missing children. This chapter presents the first detailed demographic characterization of missing children in Guatemala, including age, sex, and geographic distribution, for the period 2018–2020. We also compare the Twitter data with official police records, showing that the latter often undercounts cases and omits data from certain periods. This work highlights the potential of social media data to complement traditional sources for studying societal challenges that are difficult to measure through official statistics alone.

The main contributions are the following:
▷ We present a novel methodology for collecting data on missing children using Twitter, in-

cluding systematic data collection and image text extraction techniques, enabling near real-time monitoring.
▷ We provide the first detailed demographic profile of missing children in Guatemala (2018–2020), analyzing age, sex, and geographic distribution over time.
▷ We demonstrate how combining social media data with official sources can overcome limitations in traditional datasets and provide more timely and comprehensive insights on missing populations.

This chapter is based on Vieira et al. (2022a). The author contributions for this publication are as follows: **Vieira, C.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Alburez-Gutierrez, D.: Conceptualization, Visualization, Writing – review & editing; Nepomuceno, M.R.: Conceptualization, Writing – review & editing; Theile, T.: Visualization, Writing – review & editing

### 1.1.6 Investigating the Predictability of User-watching Behavior on TikTok via Data Donation (Vieira et al., 2026a)

The sixth chapter investigates the predictability of TikTok users' watching behavior, focusing on whether users watch short-format videos until the end. This work is motivated by previous studies showing that, despite increasing personalization on TikTok, the proportion of videos watched until the end remains relatively stable over time. To explore this phenomenon, we conducted a controlled experiment with 80 participants on Zoom, using a curated playlist of over 250 TikTok videos. Through data donation, we collected granular digital trace data capturing users' interactions with the playlist, along with demographic information and video metadata. We analyzed the predictability of user watching behavior and identified the most important features for classifying whether a user would watch a video until the end. To validate our findings, we compared the experimental results with real-world TikTok data, demonstrating the generalizability of the patterns observed. Furthermore, we evaluated matrix factorization–based recommendation models on the experimental dataset and benchmarked their performance against real-world datasets, highlighting the inherent challenges of recommending short-format videos compared to more established items such as movies.

The main contributions are the following:
▷ We design and conduct a controlled experiment to study TikTok users' engagement with short-format videos, providing a consistent dataset of user-watching behavior across a curated playlist.
▷ We demonstrate that video metadata are key predictors of user-watching behavior, with video duration emerging as the most important factor, whereas user demographics play only a marginal role.
▷ We show that recommending short-format videos is substantially more challenging than for traditional items, such as movies, and that user-watching behavior for very short videos (up to 13 seconds) is inherently stochastic.

▷ We validate our experimental findings using a real-world TikTok data, highlighting the limits of predictability in user engagement on short-format video platforms.
▷ We provide actionable insights for recommendation systems, content creators, marketers, and content moderators regarding short-format video engagement.
▷ To promote transparency and enable reproducibility, the code and datasets used in this analysis are publicly available (Vieira et al., 2026a).

This chapter is based on Vieira et al. (2026a). The author contributions for this publication are as follows: The author contributions for this publication are as follows: **Vieira, C.C.**: Conceptualization, Methodology, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review & editing; Mousavi, S.: Data curation, Writing – review & editing; Remaining co-authors: Conceptualization, Writing – review & editing

## 1.2 Thesis outline

In summary, this thesis examines the use of digital trace data to advance research on migration, culture, inequality, and online behavior in algorithmically mediated platforms. Drawing on data from multiple platforms – Facebook, Twitter (now X), Wikipedia, and TikTok – the studies demonstrate both the opportunities and challenges of using such data to address questions that are often difficult to tackle with traditional sources. The chapters collectively address four thematic areas: (i) culture, through the development of a measure of cultural similarity derived from Facebook data; (ii) migration, by applying cultural similarity measures to study international mobility and examining the relationship between migration and information-seeking on Wikipedia during crises; (iii) inequalities and vulnerable populations, through analyses of gender gaps in STEM-related interests across countries and within Brazil, as well as the characterization of missing children in Guatemala; and (iv) online behavior, by investigating user engagement with TikTok videos. Together, the studies highlight digital trace data as a valuable complement to traditional sources for studying society, while also expanding research beyond the Global North. The final thematic area also reflects on evolving modes of data access, underscoring the potential of data donation as a complementary approach in light of increasingly restricted social media APIs.

Specifically, this thesis is organized as follows:

**Chapter 2** proposes the use of passively collected digital traces from the Facebook Advertising Platform (Facebook Ads) and focuses on food and drink as markers of a country's culture to develop a measure of cultural similarity. The methodology provides measures of similarity between countries using both symmetric and asymmetric approaches. The proposed measures of cultural similarity, despite focusing solely on food and drink as cultural markers, captures a substantial portion of the variability observed in traditional data sources, such as the World Values Survey (WVS). Finally, the chapter assesses the association between migration and cultural similarity between countries by comparing this measure of cultural similarity with international

migration data. In most countries, larger immigrant populations are associated with greater similarity in food and drink preferences between the country of origin and the destination country. These results suggest that immigrants contribute to transmitting the culture of their home countries to new environments.

**Chapter 3** investigates the relationship between the cultural similarity derived from Facebook data, presented in the first chapter, and human migration across countries. This chapter illustrates the impact of incorporating measures of cultural similarity, based on food and drink interests, into gravity models for predicting migration. The results indicate that, despite their limitations, the new measures of cultural similarity derived from Facebook data improve the predictive power of traditional gravity models and have a predictive capacity comparable to traditional variables used in the literature, such as shared language and history. Finally, while some variables such as shared language, history, and geographic distance are static and symmetric, cultural attributes from daily life are sensitive to changes in the environment and can be represented as an asymmetric measure of similarity between countries, adding value to models of migration, from both the substantive and predictive perspectives.

**Chapter 4** addresses the challenge of studying information-seeking behavior during forced migration, where timely and granular data are often scarce. Focusing on the Ukrainian refugee crisis following Russia's invasion of Ukraine in 2022, it investigates the relationship between refugee flows and online information-seeking. The study leverages Wikipedia Pageviews as a novel source to capture refugees' information needs, using language editions as proxies for refugee origin and Wikipedia articles about European cities as proxies for potential destinations. By analyzing daily article views across these language editions, it uncovers correlations between Wikipedia readership patterns and the distribution of Ukrainian refugees in Poland and Germany, as well as the temporal alignment between border crossings and views on Wikipedia. The results show that increases in refugee arrivals Granger-cause surges in Wikipedia article views, underscoring the reactive nature of information seeking during forced migration. This contrasts with labor migration, where information-seeking typically precedes movement. Moreover, while official temporary protection applications often lag refugee arrivals by several weeks, Wikipedia activity responds almost immediately after border crossings. These findings position Wikipedia as a valuable, near real-time indicator of refugee movements and a potential early-warning system in crisis contexts, while also highlighting the methodological limitations of applying digital trace data to forced migration research.

**Chapter 5** presents a large-scale analysis of the global STEM gender gap using Facebook Ads data, offering a complementary approach to traditional surveys that are often costly and restricted to developed countries. By curating interests associated with 142 STEM and non-STEM college majors across 198 countries, the study quantifies gender balance related to interests in disciplines. The findings reveal that, despite the platform's bias toward women, female Facebook users dominate non-STEM interests, while STEM interests remain largely male-dominated. Within STEM, however, distinct patterns emerge: Life Sciences and Mathematics attract more women,

whereas Engineering and Technology are strongly male-oriented. A comparison with the 2021 Global Gender Gap Report shows that Facebook-based estimates are highly correlated with official statistics and can extend coverage to countries not included in traditional surveys. As a case study, the chapter examines Brazil, analyzing gender balance across states, majors, and demographic groups. Consistent with official national data, the results show that women's interest in STEM decreases with age and higher levels of education, highlighting persistent inequalities in the Brazilian context.

**Chapter 6** addresses the challenge of studying missing children, a phenomenon of high societal relevance. Focusing on Guatemala, a country affected by poverty, violence, and migration pressures, it investigates the demographic composition of missing children during the 2018–2020 period. The study leverages Twitter data as a novel source, using the official Alerta Alba-Keneth account as a proxy for reports of missing children and applying image processing techniques to extract structured information from posted alerts. By complementing official police records, these digital traces provide detailed information on disappearances and enable the first systematic description of missing children in Guatemala by age, sex, and geographic distribution. The results show that more than half of reported cases involve children under 18, with girls and younger children particularly overrepresented, and reveal spatial patterns linked to urban centers. These findings highlight how social media data can complement scarce official statistics to shed light on vulnerable populations. Moreover, while the study focuses on Guatemala, the methodology can be applied to other contexts, demonstrating the broader potential of digital trace data for studying populations that may otherwise be difficult to observe using only traditional sources.

In terms of data collection techniques, the first three chapters present studies that primarily rely on APIs for data access. However, as social media platforms increasingly restrict data availability, alternative approaches such as data donation, facilitated by GDPR, provide new opportunities for data collection. The last study explores these new opportunities raised from data donation to study users' online behavior, specifically regarding engagement with online content.

**Chapter 7** highlights the potential of data donation for research while investigating video recommendation and the predictability of user engagement on TikTok, measured by whether users watch videos until the end. The chapter first focuses on user-watching behavior, framing the likelihood that a given user will watch a video until the end as a classification task. It leverages an experimental dataset collected through controlled interactions with curated TikTok playlists, complemented by demographic information from participants, and validates the findings against real-world TikTok data. The results show that video metadata, particularly video duration, are the primary predictors of user-watching behavior, while demographic attributes contribute only marginally. The chapter then evaluates the predictability of video recommendations by applying matrix factorization, revealing that short-format videos are substantially more challenging to recommend than traditional items such as movies. Within the TikTok experimental dataset, very short videos (up to 13 seconds) exhibit even higher errors compared to longer videos. Overall,

these findings highlight both the inherent unpredictability of user engagement on short-format video platforms and the challenges of modeling and recommending content in this context.

**Chapter 8** concludes the thesis by synthesizing findings across all chapters, demonstrating how digital trace data contribute to research on culture, migration, inequalities, and online behavior in algorithmically mediated platforms. It highlights methodological innovations, such as the use of data donation, limitations, and outlines future research directions in the field of computational social science.

CHAPTER 2

# Measuring Cross-country Cultural Similarity: Evidence from Facebook

## 2.1 Introduction

Açaí bowls are a very popular food in the United States. According to *The Daily Meal* media outlet, açaí bowls were the trendiest breakfast of 2020. [1] How can we explain this when Brazil, the home country of açaí, is thousands of miles away from the US?

Despite far distances and differences between cultures, many populations share cultural preferences. Cultures are influenced by economic, political, and demographic changes, with migration being one of the main drivers of cultural changes across countries (Cooper and Denner, 1998). However, it has been difficult to measure the association between culture and international migration because cultural data that can be compared across multiple countries are scarce.

In this study, we propose a scalable methodology to measure cultural similarity between countries by using data from social media. Our proposed measure of cultural similarity also considers other characteristics such as international data availability, time, cost, and asymmetry (e.g. country $A$ may adopt features of country $B$'s culture, but country $B$ may not adopt features of country $A$'s culture), which have hampered previous efforts to study cultural similarity. Moreover, because migration plays a crucial role in shaping cultural similarity between countries, as an additional contribution we show how our approach can be used to quantify the association between cultural similarity and international migration across countries.

The relationship between migration and culture is bidirectional: cultural fit is one of the most important factors that people consider before moving to a new destination country (Caragliu et al., 2013) and migrants transmit cultural elements from their origin country to their destination country and back (Mesoudi, 2018). However, most of the studies analyzing cultural changes due to migration are restricted to one or a few countries or, when considering an international perspective of the effect of migration on cultural dynamics, the cultural distance measures used are by construction symmetric (Rapoport et al., 2020). Since migration is neither homogeneous

[1] https://www.thedailymeal.com/news/trendiest-breakfast-of-2020/070920

across countries nor symmetric, we develop non-symmetric measures of cultural similarity across numerous countries to more accurately represent processes of international cultural exchange.

Moreover, other factors complicate the comparison between cultural similarity and migration. First, to measure cultural similarity between countries, we need a measure of cultural similarity or cultural distance (Ghemawat, 2001). Cultural distance measures refer to operational parameters that can be used as proxies for cultural dimensions and allow estimating scores to gauge the extent to which countries differ culturally (Tung and Verbeke, 2010). The cultural dimensions used to measure culture can vary depending on the focus of the research (Mohr et al., 2020).

Operationally, culture has been traditionally measured via sampling surveys (Kwantes and Glazer, 2017) in which the survey responses are used to characterize cultural aspects of a country. Several studies, such as Schwartz's value survey (Schwartz, 1994), the World Values Survey (Inglehart and Welzel, 2010), and Hofstede's cultural characteristics (Hofstede, 1983) [2] try to identify cultural dimensions related to values on which countries tend to differ. Moreover, some methods derived from surveys propose evaluating the relative distance between countries (De Santis et al., 2016; Gupta et al., 2002; Mucciardi and De Santis, 2017; Muthukrishna et al., 2020).

However, the study of culture can also focus on aspects of our daily life by considering cultural objects such as the clothes we wear, the music we listen to, and the food we eat (Kwantes and Glazer, 2017; Recchi and Favell, 2019). Food studies is an established interdisciplinary field that recognizes the centrality of food for cultural practices and cultural identity (Ashley et al., 2004; De Solier and Duruz, 2013). Several studies have explored how food communicates our culture and the mechanisms by which we relate food to our cultural identities, whereas others revealed that people and groups can be discriminated against on the basis of their food and cultural habits (Almerico, 2014; Boutaud et al., 2016; Sibal, 2018). Similarly, Almerico (2014) presented an interdisciplinary study that documents the intricate relationships between food, culture, and society from a sociological perspective. Typical food from a country, for example, can be used to approximate cultural distance by characterizing the preferences for similar food in other countries. Cantarero et al. (2013) show that cultural identity greatly influences food choices by performing a qualitative and quantitative analysis in the Comunidad Autónoma de Aragón, Spain. They found that people prefer to consume foods that are symbolically associated with their own culture to reinforce their sense of belonging. In summary, these previous studies are important to create a strong connection between the culture of a country and food, our cultural marker to measure cultural similarity between countries.

Such studies have typically been based on surveys, but surveys have important limitations. For example, in addition to measurement error (Groves and Lyberg, 2010), results may suffer from various biases (Suchman, 1962) like social desirability bias, question order bias, and acquiescence bias. Furthermore, surveys are costly and require a long time to run. To overcome part of these limitations, we propose an approach that relies on passively-collected data from social media.

[2] *https://www.hofstede-insights.com/models/national-culture/*

Social media platforms provide complementary tools that can be used to measure cultural preferences and compare regions via passively-collected data (You et al., 2017). As one of the first studies to address this question by using online data sources, Silva et al. (2014) identified cultural boundaries and similarities across populations by clustering them based on the analysis of food and drink habits. However, their analyses of culinary habits around the world were limited to Foursquare check-ins, which considered only 101 interests and, consequently, underestimate the true breadth of users' interests and of cultural variation. Moreover, Foursquare is not widely used and demographically biased, [3] especially because Silva et al. (2014) only gathered data from Foursquare users who were also Twitter users and decided to cross-post their Foursquare check-ins to Twitter.

Other studies have used data provided by the Facebook Advertising Platform. With more than 2.7 billion worldwide users, [4] Facebook captures a larger and more diverse population than other social media. A growing literature has successfully used Facebook Ads data to study many different topics such as migration (Dubois et al., 2018; Spyratos et al., 2018; Zagheni et al., 2017), the relationships between immigrant communities (Herdağdelen et al., 2016), and cultural assimilation between migrants (Stewart et al., 2019). We therefore see high potential in applying Facebook Ads data to the study of cultural similarity regarding international food and drink preferences.

In a first study in this area, Vieira et al. (2020) examined the similarity between selected countries and Brazil based on their population's interests in typical Brazilian food. However, the results were limited to foods listed on Wikipedia. This limits the potential list of food and, importantly, some countries do not have a Wikipedia page dedicated to list their typical food. Here, we propose a more scalable, data-driven methodology to identify the most popular foods and drinks in each country from a list containing *more than 200,000 interests* on Facebook Ads (Speicher et al., 2018).

More recently, Obradovich et al. (2020) examined cross-national cultural differences across nearly 60,000 topic dimensions from Facebook Ads. They also validated their work by comparing the cultural distance calculated from their measurement with traditional survey-based measures. However, the authors used different types of features from Facebook Ads, which is a 'black-box' model, to represent countries in terms of culture. In contrast, we propose a methodology to select the most important cultural attributes, in this case, related to food and drink for each country from a data set containing more than 200,000 interests of Facebook users. In this case, we choose to compare countries in terms of fewer, but explicitly known attributes selected from a very large data set through a data-driven approach. In other words, interests that are not relevant to any of the countries are disregarded in order to reduce feature sparsity. Finally, we additionally present

[3] ***https://www.statista.com/statistics/814726/share-of-us-internet-users-who-use-foursquare-by-age/***

[4] ***https://www.facebook.com/iq/insights-to-go/2740m-facebook-monthly-active-users-were-2740m-as-of-september-30/***

the first work exploring a non-symmetric measure of cultural similarity and further assess the association between migration and food- and drink-based cultural similarity across countries.

We expand the methodology proposed by Vieira et al. (2020) to (i) propose a scalable approach based on data from the Facebook Advertising Platform (Facebook Ads) to obtain proxies for culture to measure cultural similarity between countries; (ii) develop measures of cultural similarity to compare countries according to their population's interests revealed by Facebook; (iii) assess to which extent cultural similarity between countries is associated with migration.

Culture encompasses many aspects, including preferences for music, art, and food. Although in this paper we focus on food and drink as markers of culture (Recchi and Favell, 2019), we offer a methodological contribution that can be applied to other cultural markers. Moreover, we provide a substantive contribution of results for the case of food and drink. Finally, we hypothesize a link between migration and food and drink preferences revealed by social media. We provide initial results that lead to further avenues for research in this area at the intersection of social computing, demography, and sociology.

## 2.2 Data

In this section, we present the data and methodology proposed in this work. Data collection was performed in compliance with the terms of services of the websites from which data was collected. All data used in this study are openly available, and the data sources are specified in each subsection, including the data from Facebook, openly available through Facebook's Marketing Application Programming Interface (API). [5]

Because food is recognized as a central cultural marker (Ashley et al., 2004; De Solier and Duruz, 2013; Recchi and Favell, 2019) and there is a great variety of popular local food and drink in each country, we decided to use the data categorized by Facebook Ads as related to *Food and drink* from the data set collected via *snowball* (Speicher et al., 2018) containing most of the interests available on Facebook Ads in 2019.

Following Silva et al. (2014)'s prior work in this area, we selected a subset of 16 countries (Argentina, Australia, Brazil, Chile, Great Britain, France, Indonesia, Japan, South Korea, Malaysia, Mexico, Russia, Singapore, Spain, Turkey, and the United States). The analysis is limited to the subset of countries we consider, but following the methodology proposed, it is possible to extend the analysis to include any other country with a Facebook audience.

Regarding privacy, our work uses only aggregated data and we do not gather nor link any personal information to any particular user. The methodology adopted to gather the data as well as the audience of each food and drink over Facebook users from each country through the

[5] https://developers.facebook.com/docs/marketing-apis

Facebook Advertising Platform is described in the following section, where we also present other data sets that we use to assess our proposed methodology.

### 2.2.1 Facebook Ads data

The Facebook Marketing API allows marketers and researchers to obtain an estimate of the number of monthly active users for a proposed advertisement that matches given input criteria (Kosinski et al., 2015). For that, the platform provides a list of demographic attributes, such as age, gender, home location, and interests that can be customized by the advertiser as the input query. Thus, after specifying the target audience, and before the ad is launched, advertisers are provided with the size of the audience that matches the target specifications.

Attributes like age, gender, and location are explicitly declared by the users in their profiles, whereas interests can be either declared by the user or inferred by Facebook based on user activities such as posting or interacting with contents (e.g., liking content, sharing content, or updating one's status). Facebook users generate traces of their preferences in multiple domains. The interests are categorized in one of the following categories: *People*, *Education*, *News and entertainment*, *Travel, places and events*, *Food and drink*, *Business and industry*, *Technology*, *Hobbies and activities*, *Sports and outdoors*, *Lifestyle and culture*, *Shopping and fashion*, *Fitness and wellness*, *Family and relationships* (Dubois et al., 2018) such as music (Stewart et al., 2019) and food (Vieira et al., 2020).

Once not all the 258,366 interests collected via *snowball* (Speicher et al., 2018) are related to food and drink, we selected 9,309 categorized by Facebook Ads as referring to *Food and drink* interests. However, only 996 interests related to *Food and drink* have an audience higher than 1,000 in each one of the countries considered. Note that 1,000 is the minimum value given by the Facebook Ads API, meaning that the true value can be anywhere between 0 and 1,000. We adopted this cutoff to reduce data sparsity and include only interests that can be meaningfully compared across all countries in our sample. Because the Facebook audience in each country is in the order of millions, this threshold is further low enough that popular interests relevant to our analyses would not be erroneously excluded. Moreover, from those 996 interests, many included restaurants (e.g., Burger King), kitchen utensils (e.g., spoon), and specific brands (e.g., Pepsi). To consider only interests which represent the names of food, drinks, ingredients, or dishes, we manually validated the data set by removing other interests such as brand names. At the end of this process, 728 interests were considered in the following analysis.

The Facebook Ads Platform does not include detailed information on how interests are estimated, and the population of Facebook users is known to be biased concerning gender, age, and other socio-demographic characteristics (Araujo et al., 2017; Ribeiro et al., 2020). However, many studies have already shown that this platform can be used as a good proxy to study population's characteristics (Grow et al., 2021). Also, previous studies showed that the subset of interests used in our work is representative (Speicher et al., 2018).

Regarding privacy, our work uses only aggregated data and we do not gather nor link any personal information to any particular user. We complied with the terms of service of Facebook's Marketing API. [6] In particular, the data collected is anonymous (12. j.) and we did not build or augment any user profiles (5. b. ii.). Moreover, we did not perform, or facilitate or support others in performing, any of the following prohibited practices (3. a.) listed on the new Facebook Platform Terms the new Developer Policies effective since August 31, 2020. Due to legal requirements regarding the publication of the raw data collected from the Facebook Advertising Platform data, the minimal data underlying the results of this study are available for academic purposes upon request. All the measures derived from the Facebook Ads data and the code to analyze the data sets and generate the figures are available in a public web repository. [7]

### 2.2.2 Survey data

Because we are interested in measuring cultural similarity between countries by using Facebook Ads data, it is important to compare our results with previous works that try to address similar questions. Inglehart and Welzel proposed a cultural map of the world based on one of the most prominent surveys, the World Values Surveys data. We used the data from the Inglehart-Welzel cultural map using the wave 7 (2017-2020) of the WVS (Inglehart and Welzel, 2005). [8]

The World Values Surveys data set looks at several cultural dimensions such as religion, politics, economics, and lifestyle. However, Inglehart and Welzel assert that there are two major dimensions of cross-cultural variation in the world: Traditional values (which emphasize the importance of religion, parent-child ties, deference to authority, and traditional family values) versus Secular-rational values (represented by societies which place less emphasis on religion, traditional family values and authority); and Survival values (emphasis on economic and physical security) versus Self-expression values (high priority to environmental protection, equality, and rising demands for participation in decision-making in economic and political life). Based on these two dimensions, they proposed an international cultural map where the location of each country is given by the scores on these two dimensions. Figure 2.4a shows the reproduction of the World Value Survey cultural map 2020 considering the selected countries.

Even if not focused on food and drink, these data can be used for comparison purposes, providing an idea of how food- and drink-based cultural markers are correlated with the value-based cultural markers estimated by the WVS. In addition, because we are using only food and drink as cultural markers, we can measure how similar our results are in comparison with one of the most complete data sources based on surveys.

[6] https://developers.facebook.com/policy/#marketingapi

[7] https://github.com/carolcoimbra/cultural-similarity-fb

[8] https://www.worldvaluessurvey.org/WVSEventsShow.jsp?ID=428

### 2.2.3 Migration data

We also compare our results with migration data to understand the mechanisms behind the cultural similarity expressed by the Facebook users' interests in popular food and drink from different parts of the world. The migration data refer to the total international migrant stock in 2019 by destination and origin provided by the United Nations (DESA, 2019). [9] These data include estimates of how many immigrants from each of one of the 232 countries and areas of the world were living in each of these 232 countries and areas of the world in 2019. The estimates are based on official statistics on the foreign-born or the foreign population.

## 2.3 Methodology

In this section, we present the methodology that we developed to measure the cultural similarity between countries by exploring Facebook audiences interested in popular food and drink across countries. First, we focus on selecting a subset of popular food and drink for each country. Then, we use those interests related to food and drink to create a representation of each country as a vector to calculate the similarity between countries.

### 2.3.1 Popular food and drink

In order to select the subset of popular food and drink, we collected the Facebook audience interested in each food and drink. To consider only the most popular food and drink in each country, we selected the top food and drink according to their popularity among Facebook users living in the country.

Concretely, in our context, the popularity of a food or a drink among the Facebook users living in the country is measured as the proportion of users on Facebook living in each country interested in that food or drink. Equation 2.1 shows how we can obtain this proportion. $audience_c(i)$ represents the number of Facebook users in country $c$ interested in food or drink $i$, while $audience(i)$ represents the total audience interested in the same food or drink on Facebook as a whole. Notice that $audience(i)$ can also be written as a sum of the Facebook audiences interested in $i$ for each country, so $audience(i) = \sum_{c^*} audience_{c^*}(i)$.

$$audience_{i-norm} = \frac{audience_c(i)}{audience(i)} \tag{2.1}$$

Finally, according to Equation 2.1, for each country, we select the top 50 popular foods and drinks based on the proportion of the Facebook audience interested in each food and drink. For

[9] https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp

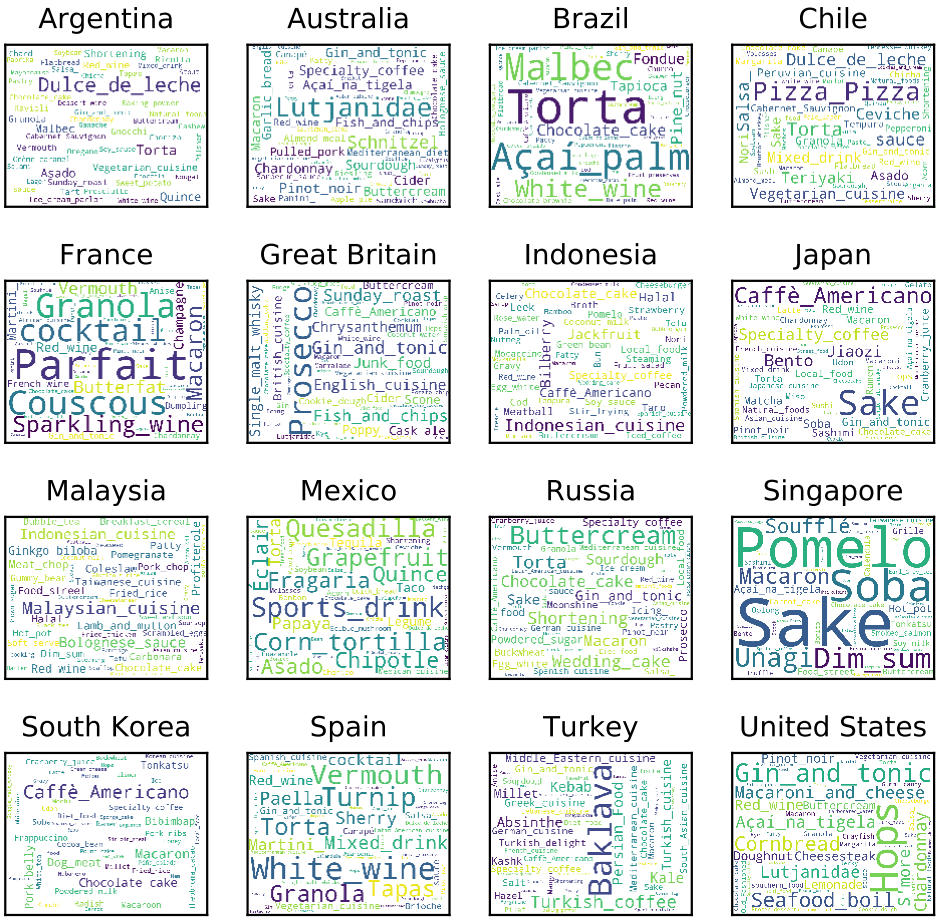
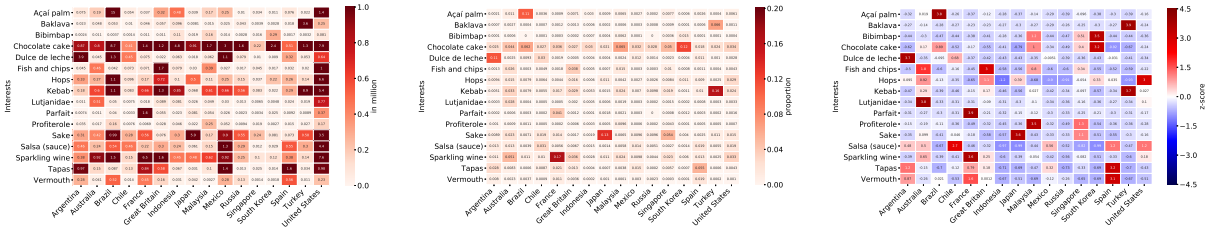



**Figure 2.1:** Word clouds showing the names of the 50 food and drink with the largest proportion of the audience in each country, based on data from the Facebook Advertisement Platform. The size of the words is proportional to the audience interested in the food and drink in the country according to Equation 2.1. The colors do not have substantive meaning: they are used only to differentiate the words.

the rest of the paper, we consider only the top foods and drinks selected. Figure 2.1 shows a word cloud representing the top 50 foods and drinks selected for each country. The words differ in color (colors were chosen randomly) and the size represents the proportion of Facebook users from each country interested in each food and drink according to Equation 2.1.

### 2.3.2 Vector representation

The Cultural Similarity between countries is measured via a vector, where each position in the vector corresponds to the size of the Facebook audience interested in a food or a drink. Because the audience can also vary greatly across countries, to make a fair comparison between interests



**(a)** Absolute audience on Facebook

**(b)** Audience normalized using Equation 2.2

**(c)** Audience z-score normalized using Equation 2.3

**Figure 2.2:** Descriptive statistics on Facebook audience size across countries interested in an illustrative, randomly selected sample of food and drink.

in these countries, we need to normalize the audience for each interest by the estimated Facebook population of each country.

We use Equation 2.2 to normalize the audience for each interest by the Facebook population in each country. Given the number of Facebook users in country $c$ ($FBUsers_c$), the proportion of the audience $A_c(i)$ in $c$ who is interested in food or drink $i$ is given by:

$$A_c(i) = \frac{audience_c(i)}{FBUsers_c} \tag{2.2}$$

For illustrative purposes, Figure 2.2a shows the absolute number of Facebook users living in each one of the 16 countries interested in 16 randomly-selected interests. Similarly, Figure 2.2b shows the same subset of food and drink and the *proportion* of the Facebook audience interested in them for each country. For instance, let us consider the US, which has 230 million Facebook users. Because the number of Facebook users living in the US interested in Kebab is 5.4 million (see Figure 2.2a), applying Equation 2.2, 2.4% of the American population on Facebook is interested in Kebab (see Figure 2.2b). We applied the same operation to all the other foods and drinks.

However, the distribution of foods and drinks is highly unbalanced and some are vastly more popular than others. This difference can impact the similarity measurement between countries because the measure would be disproportionately driven by these few, most popular foods and drinks. For instance, the difference between foods and drinks that have a small proportion of interest in each country will be almost zero, whereas the difference between the most popular food and drink are more likely to be large in relative terms. Note that row *Chocolate cake* and column *United States* in Figure 2.2a dominate the whole matrix. Observe in Figure 2.2b that this problem is only partially solved by computing the proportions, as the row *Chocolate cake* continues to be have the highest values for all countries. To give the same importance to all food and drink, we normalize and smooth these distributions by their z-scores as shown in Equation 2.3.

$$\textit{z-score}\ (A_c(i)) = \frac{A_c(i) - E[A(i)]}{\sigma(A(i))} \tag{2.3}$$

where $E[A(i)]$ is the mean of the $A_p(i)$ across countries, whereas $\sigma$ refers to their standard deviation. In short, the mean is subtracted from the score of each interest and divided by its standard deviation. As a result, each value now represents how many standards deviations an interest in a certain country deviates positively or negatively from the mean. Figure 2.2c shows the same heatmap as Figure 2.2b, but considering the z-score normalization. As expected, we observe that the distribution is heterogeneous and does not seem to exhibit a few dominant interests in all the countries. Considering the example in Figure 2.2c, we can assume that each column represents each country as a vector in terms of their population interests in those foods and drinks.

### 2.3.3 Cultural similarity

The Cultural Similarity ($CS$) between two countries is defined in terms of the Cosine distance between countries with respect to the subset of popular interests in one country. Equation 2.4 defines this formally: $c_1$ and $c_2$ represent two countries, $d_{c_1}$ is the subset of interests that are popular in country $c_1$ that we are considering to generate the vectors $\mathbf{v_{d_{c_1}}}(c_1)$ and $\mathbf{v_{d_{c_1}}}(c_2)$. These vectors have elements that represent the level of interest in the country considered for the subset of food and drink that are most popular in country $c_1$.

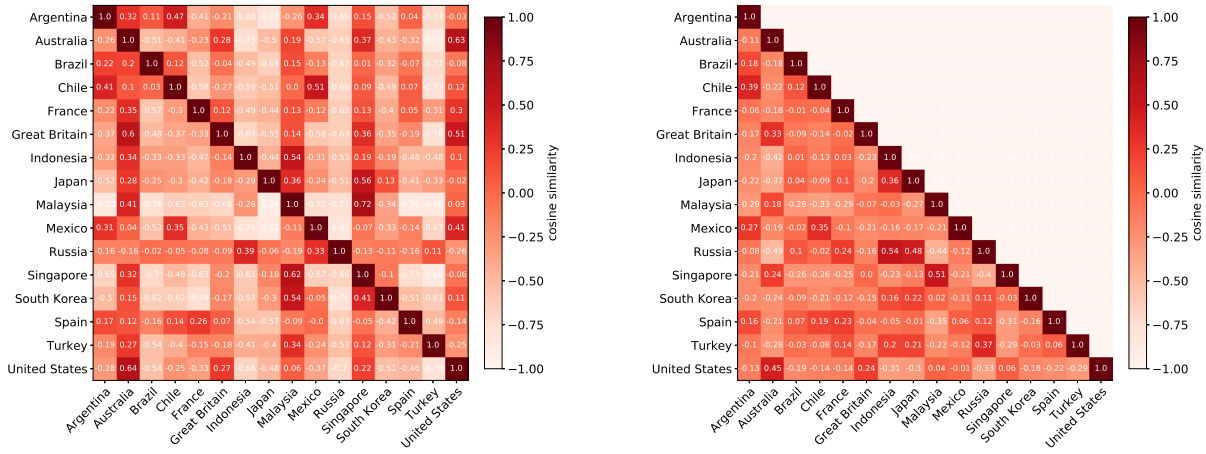
$$CS_A(c_1, c_2||d_{c_1}) = 1 - cosine\ dist(\mathbf{v_{d_{c_1}}}(c_1), \mathbf{v_{d_{c_1}}}(c_2)) \tag{2.4}$$

Note that the measure of cultural similarity between $c_1$ and $c_2$ is measured in terms of the $c_1$ top popular foods and drinks whereas the similarity between $c_2$ and $c_1$ is measured in terms of the $c_2$ top popular foods and drinks. In this case, because the Cultural Similarity between $c_1$ and $c_2$ is different from the Cultural Similarity between $c_2$ and $c_1$, this measure is not symmetric ($CS_A(c_1, c_2||d_{c_1}) \neq CS_A(c_2, c_1||d_{c_2})$). Because of the asymmetry, the measure is named *Asymmetric Cultural Similarity* ($CS_A$). However, we could also measure the similarity between two countries considering a fixed set of interests for both countries. In this case, we can refer to the measure as *Symmetric Cultural Similarity* ($CS_S$) as the Equation 2.4 can be re-written as shown by Equation 2.5:

$$CS_S(c_1, c_2||d^*) = CS_B(c_1, c_2||d^*) = 1 - cosine\ dist(\mathbf{v_{d^*}}(c_1), \mathbf{v_{d^*}}(c_2)) \tag{2.5}$$

Equation 2.5 shows that the Symmetric Cultural Similarity between two countries $c_1$ and $c_2$ corresponds to a measure of the Cultural Similarity between them considering all top foods and drinks occurring across all countries, $d^*$. Because the subset of interests is fixed, the Symmetric Cultural Similarity between $c_1$ and $c_2$ is equal to the Symmetric Cultural Similarity between $c_2$ and $c_1$.

To create a representation via vector with a fixed size for each country while keeping the number of attributes manageable, we consider the top 50 popular foods and drinks for each




**(a)** Asymmetric Cultural Similarity

**(b)** Symmetric Cultural Similarity

**Figure 2.3:** Cross-country cultural similarities. Each cell corresponds to the cultural similarity between the country in the row and the country in the column. In (a) countries are represented as a vector of 50 dimensions considering the top 50 food and drink of the country in the row. In (b) countries are represented as a vector of 394 dimensions considering the top 50 food and drink in each country.

country. Because the top 50 popular food and drink of several countries overlap because one food or drink can be popular in more than one country, we examined 394 unique interests. In other words, considering the z-scored normalization, we create a representation via vector for each country in terms of their Facebook users' preferences regarding the 394 most popular foods and drinks selected. Then, for each country, we measure the Symmetric Cultural Similarity by applying the measure given by Equation 2.5.

## 2.4 Results

In this section, we present the main results regarding our proposed measures of cultural similarity, the comparison between cultural similarity and the WVS data, and the association between cultural similarity and migration data.

### 2.4.1 Patterns of cultural similarity

Figure 2.3a shows the Cultural Similarity between each representation considering the top 50 popular foods and drinks from the country represented by the rows. Notice that because for each row we are considering a different subset of food and drink, the matrix is non-symmetric. As explained before, this measure is therefore named Asymmetric Cultural Similarity.

Notice that the Asymmetric Cultural Similarity shows that some countries are closer to others in terms of their own popular foods and drinks but are more distant from each other when the foreign country's popular foods and drinks are considered. Some countries such as Indonesia, Japan, Russia, and Turkey are more distant from most of the others in terms of their interest in foreign food and drink. Other countries such as Australia, Great Britain, Malaysia, Singapore, and the United States seem to be more eclectic in their interest in foreign food and drink. Note that Indonesia, Japan, Russia, and Turkey have relatively few immigrants, whereas Australia, Great Britain, and the United States have many. We therefore show that countries with fewer immigrants have more idiosyncratic tastes in food and drink, possibly due to limited exposure to foreign cuisines. Australia, Great Britain, and the United States also appeared the most similar to each others' food preferences. The exception was the United States which showed that American Facebook users were also highly interested in Mexican food and drink. The similarity between the US and Mexico could be related to the number of Mexican immigrants living there. Since the US is one of the most preferred destinations by immigrants worldwide, we notice that the similarity between the US and the other countries is high when considering interests in foreign food and drink by Facebook users living in the US. However, the opposite is not true - these other countries appeared less interested in American food and drink than the the other way around. Migration can be one reason why some countries' cultures are more similar to others. However, language and geographic distance also seem to correlate with the cultural similarity. For instance, the English-speaking countries (Australia, Great Britain, and the US), the Spanish-speaking countries (Argentina, Chile, Mexico, and Spain), and the Latin American countries (Argentina, Brazil, Chile) each seemed to be similar to each other. Finally, the matrix in Figure 2.3a is constructed in a way that countries with high column sums represent more culturally diverse countries with preferences for varied foreign foods, whereas the countries with high row sums represent more culturally influential countries whose food is popular in various other countries.

Instead of asymmetric measures, we next examine symmetric measures of cultural similarity. Figure 2.3b shows the Symmetric Cultural Similarity between each pair of countries represented via vector. Note that now the matrix is symmetric and we can name the measure Symmetric Cultural Similarity. In this case we are comparing the countries in terms of their preferences for foods and drinks that are popular around the world. In contrast to the measure of Asymmetric Cultural Similarity, the Symmetric Cultural Similarity provides a broad measure of cultural similarity between countries. For instance, Australia and Indonesia and Australia and Russia are some of the most distant countries from each other. Moreover, Indonesia is the most similar country to Russia. As observed for the Asymmetric Cultural Similarity, language and geographic distance are again associated with low Cultural Similarity because Asian, Latin American, English-speaking, and Spanish-speaking countries, respectively, are culturally close to one another.

### 2.4.2 Comparison with the WVS

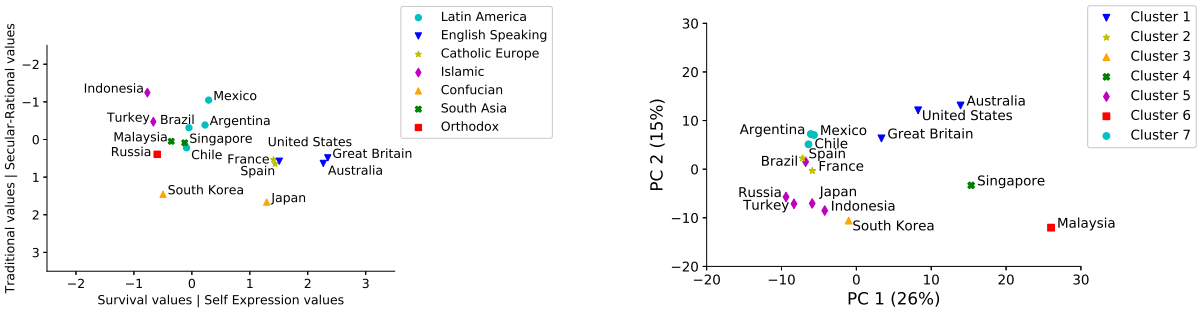
We also contrast our results on Cultural Similarity with other cultural measures based on the World Value Survey (WVS) data. Figure 2.4a shows the World Value Survey cultural map (Inglehart and Welzel, 2010) considering two major dimensions of cross-cultural variation according the WVS data. The map shows each country as part of one of the seven clusters classified by the WVS according to language, religion, or geographic location (Figure 2.4a: Latin America, English Speaking, Catholic Europe, Islamic, Confucian, South Asia, and Orthodox). For simplification, we replicate the WVS cultural map showing only the countries we analyze in this paper. Figure 2.4b shows the cultural map created by using our Facebook-based cultural vector representation of each country. Because each country is represented by a vector of 394 dimensions, we apply Principal Component Analysis (PCA) over the data to represent the data in two dimensions. The map shows the location of each country in terms of the first and second principal components as well as the variance in the data explained by each one (41% total). Note that we applied the PCA algorithm only to visualize and compare our data the WVS cultural map.

To identify clusters in our Symmetric Cultural Similarity map we use the k-means algorithm, a widely-used clustering technique, to group countries with similar representations over the 394 popular foods and drinks. Each color/symbol represents a cluster obtained by the *k-means* algorithm. We set the parameter $k = 7$ to follow the same number of clusters defined by Inglehart and Welzel (2010).

Observe that the algorithm correctly identified the English-speaking countries, according to the WVS cultural map, as being part of one cluster (cluster 0) as well as Catholic Europe (France and Spain) as part of another (cluster 1). The Latin American countries (Argentina, Chile, and Mexico) was also in the same cluster (cluster 6), except for Brazil which was included in cluster 4 with Russia, Turkey, Japan, and Indonesia. Finally, South Korea (cluster 2), Singapore (cluster 3), and Malaysia (cluster 5) formed part of different clusters. In general, if we observe the map in Figure 2.4b, the Latin American countries and Catholic Europe were close to each other. However, they were not in the same cluster because we are not identifying the clusters based only on this two-dimensional representation, but based on the 394-dimensional representation of popular foods and drinks.

If we compare the cultural maps in Figure 2.4a and Figure 2.4b we observe that the Latin American countries seemed to be more similar to the European countries and Indonesia, Japan, Russia, and Turkey seemed to cluster together in terms of their Facebook users' interests in popular food and drink. However, we can also observe that Singapore and Malaysia were distant from the other countries, which means that the Facebook users from those countries do not share similar interests in food and drink with other countries.

Finally, by using the PCA of the our Symmetric Cultural Similarity Map, as shown in Figure 2.4b, we formally document the differences between the WVS and our methodology. First, we create the *Symmetric Cultural Similarity ranking* for each target country by sorting the list of




**(a)** The Inglehart-Welzel World Cultural Map - WVS 7 (2020)

**(b)** PCA of the Symmetric Cultural Similarity Map - Facebook data

**Figure 2.4:** Cultural similarity maps derived from traditional survey data (WVS) and digital trace data (Facebook). Panel (a) shows the Inglehart–Welzel World Cultural Map based on World Values Survey (WVS) data. Panel (b) presents a comparable principal component projection derived from Facebook interest data, clustered using k-means (k = 7) to mirror the number of cultural clusters defined in the WVS map.

countries according to their cosine similarity to the target country. Similarly, the *WVS ranking* for a given country is sorted by the most similar or, in other words, the least distant countries according to their cosine distance in the WVS cultural map. Then, we compute the Jaccard similarity between these two ranks to see if the most similar countries to a specific country are the same with both approaches.

When considering only the top 5 most similar countries to each country, the Jaccard similarity is approximately 0.22. The similarity improves when we consider the top 10 most similar countries (Jaccard similarity = 0.5), which means that our results are relatively close to the WVS in terms of identifying similarity of cultural interests, even though our measure is based only on food and drink popularity.

Similarly, we compared our results with Silva et al. (2014) to validate our results and ensure comparability with prior research. Similarly, to the *WVS ranking*, we created the *Foursquare ranking* for a given country sorted by the most similar countries according to Silva et al. (2014)'s cultural map based on food and drink preferences using Foursquare data. The Jaccard similarity between the *Symmetric Cultural Similarity ranking* and the *Foursquare ranking* is approximately 0.44 and 0.59 when we consider, respectively, the top 5 and the top 10 most similar countries to each country. We also compare the Jaccard similarity between the *WVS ranking* and the *Foursquare ranking*. When considering the top 5 and the top 10 most similar countries to each country, the Jaccard similarity is respectively, 0.14 and 0.4. This result indicates that the results obtained from our cultural map improves upon that of prior work in this area. Importantly, these comparisons are limited by the time difference: the data presented by Silva et al. (2020) are more than 5 years older than our data from Facebook Ads and more than 6 years older than the WVS cultural map. During this time, significant cultural changes may have happened, given that the world is getting

more connected every day. Besides the time difference, the Foursquare sample is also expected to be highly unrepresentative because Foursquare is not widely used (e.g., younger people and men are over-represented) and the Foursquare check-ins are limited to people who also have a Twitter account and explicitly post their check-ins, which further skews and reduces the sample. In fact, we expect to see some minor changes between the cultural maps created by using different data sources, however, we decided to keep the comparison with Silva et al. (2014) to confirm that our methodology produces results in line with other studies.

### 2.4.3 Association with migration data

To quantify the association between our measure of Cultural Similarity and migration, we use the migration data provided by the United Nations. By using migration data, we can create, for each country, an *Immigrant ranking* by sorting, in descending order, the countries by the proportion of immigrants from each nationality living in the target country. For instance, the *Immigrant ranking* in Spain would have Great Britain as the origin country of the highest proportion of immigrants living in Spain (immigrants from Great Britain correspond to 0.65% of Spain's population) followed by Argentina, France, Brazil, Russia, Chile, Mexico, United States, Australia, Japan, South Korea, Turkey, Indonesia, Malaysia, and Singapore. Similarly, we can create a *Cultural Similarity ranking* by sorting the same list of countries according to the Asymmetric Cultural Similarity between countries in terms of the interest in foreign food and drink by Facebook users living in the target country. Notice that, for each country, the *Cultural Similarity ranking* corresponds to its column in Figure 2.3a sorted in ascending order. Following the same example, the *Cultural Similarity ranking* in Spain would have Chile as the most similar country to Spain in terms of interests in top popular Chilean foods and drinks (the asymmetric cultural distance between Chile and Spain is 0.07) followed by France, Argentina, Brazil, Mexico, Russia, Great Britain, Turkey, Australia, Japan, United States, Indonesia, South Korea, Singapore, and Malaysia.

To understand how strongly the proportion of immigrants living in a country is associated with the cultural similarity between the countries of origin and destination, we measure the correlation between these two rankings. For instance, if we use the Spearman correlation to measure the correlation between the *Immigrant ranking* and the *Cultural Similarity ranking* in Spain, the correlation is equal to 0.82, a high positive correlation. Moreover, we can also compare the rankings using a more intuitive measure by calculating the proportion of countries in the top 5 of the *Cultural Similarity ranking* that are also in the top 5 of the *Immigrant ranking*. In Spain, the top 5 countries in the *Cultural Similarity ranking* are Chile, France, Argentina, Brazil, and Mexico, whereas the top 5 countries in the *Immigrant ranking* are Great Britain, Argentina, France, Brazil, and Russia. Because three (Argentina, France, and Brazil) of the countries in the top 5 of the *Immigrant ranking* are also in the top 5 of the *Cultural Similarity ranking*, it means that 60% of the most similar countries to Spain are also the countries where most immigrants living in Spain come from.

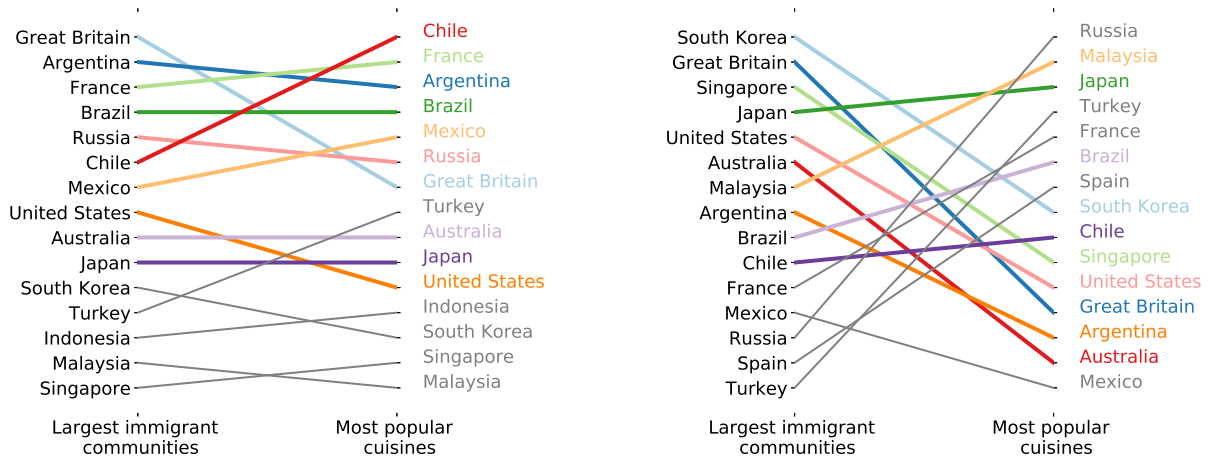
| Country | Spearman r | Proportion (top 5) |
|---|---|---|
| **Spain** | **0.82 (0.0)** | **0.6** |
| **Malaysia** | **0.75 (0.0)** | **0.6** |
| **Chile** | **0.74 (0.0)** | **0.6** |
| **Australia** | **0.65 (0.01)** | **0.8** |
| **France** | **0.65 (0.01)** | **0.8** |
| **Great Britain** | **0.62 (0.01)** | **0.8** |
| **Argentina** | **0.61 (0.02)** | **0.6** |
| **Mexico** | **0.55 (0.03)** | **0.4** |
| **Brazil** | **0.54 (0.04)** | **0.8** |
| **Singapore** | **0.5 (0.06)** | **0.4** |
| Turkey | 0.31 (0.25) | 0.4 |
| Russia | 0.11 (0.7) | 0.6 |
| Japan | 0.08 (0.78) | 0.2 |
| South Korea | 0.07 (0.81) | 0.4 |
| United States | 0.02 (0.94) | 0.4 |
| Indonesia | -0.17 (0.55) | 0.2 |

**Table 2.1:** Correlation between the *Immigrant ranking* sorted by the proportion of immigrants living in each country and the *Cultural Similarity ranking* for each country, sorted by the most similar countries in terms of cultural similarity.

Table 2.1 shows the Spearman correlation and the *p-value* associated with each correlation for each country. We also present the proportion of countries in the *Cultural Similarity ranking* top 5 that are also present in the *Immigrant ranking* top 5 for each one of the countries. For most of the countries, at least 60% of the *Cultural Similarity ranking* top 5 are also present in the *Immigrant ranking* top 5. Moreover, we observe that for most countries, there is a positive correlation between the number of immigrants from a country and the similarity between Facebook users' interests in popular food and drink from that country.

The countries shown in Table 2.1 are sorted according to the Spearman correlation between the *Immigrant ranking* and the *Cultural Similarity ranking*. Spain is the country that shows the highest correlation. Figure 2.5b illustrates the exact *Immigrant ranking* and the *Cultural Similarity ranking* for Spain in more detail. This pattern illustrates the high positive correlation we found for Spain, which showed that Spain is one of the most preferred destinations by immigrants from Great Britain, Argentina, France, and Brazil, and also shows interest in these countries' most popular foods and drinks.

However, for other countries, the correlation is lower or even negative and not significantly different from zero given the high p-values. For instance, in Indonesia the number of immigrants is not associated with higher preferences for foreign popular food and drink. Figure 2.5a illustrates the exact *Immigrant ranking* and the *Cultural Similarity ranking* for Indonesia in more detail. This pattern illustrates the negative correlation we found for Indonesia, which showed that Indonesia is one of the most preferred destinations by immigrants from South Korea, Great Britain, Singapore, and Japan but only Japan's foods and drinks are popular in Indonesia.




**(a)** Spain **(b)** Indonesia

**Figure 2.5:** Comparison between the *Immigrant ranking* sorted by the proportion of immigrants living in each country and the *Cultural Similarity ranking* for each country, sorted by the most similar countries in terms of cultural similarity.

Moreover, there is a positive association between the Spearman correlation of *Immigrant ranking* and the *Cultural Similarity ranking* (Table 2.1) and the total number of immigrants in the destination country. The Spearman correlation is 0.13 when we consider the total number of immigrants and 0.43 when we sum over the immigrants from the sixteen countries we consider in our work. As outlined above, it is expected that countries with fewer immigrants, such as Indonesia, Russia, Japan, and Turkey, would have more idiosyncratic tastes in food and drink due to limited exposure to foreign cuisines than countries with more immigrants, such as Great Britain and Australia. Finally, it is important to mention that the results are dependent on the subset of countries we are comparing. For instance, not all countries with large immigrant populations in the United States are covered by our data. For example, China is one of the most frequent top origin countries from immigrants in the United States, as well as Indonesia and South Korea (DESA, 2019), but because Facebook is not used in China, we do not have enough information to compare the *Immigrant ranking* and the *Cultural Similarity ranking* considering all possible countries.

## 2.5 Conclusion

In this work, by using Facebook Ads data, we proposed a scalable approach to obtain proxies for culture in order to measure cultural similarity between countries. Two measures of cultural similarity were proposed, Asymmetric Cultural Similarity and Symmetric Cultural Similarity. Our measures of cultural similarity were compared with the WVS data, and we highlighted some advantages of using social media data in this study. Unlike previous empirical studies, which were based on survey data and large batteries of questions, our methodology is easily scalable,

uses passively-collected information internationally available, and considers food and drink as a key dimension which explains a large fraction of the cultural similarities between countries.

Our measure of cultural similarity was also compared with migration data. Our results show that some countries (e.g., Indonesia) are more distant from most of the others in terms of their interest in foreign food and drink. Other countries (e.g., the United States) seem to be more culturally diverse or eclectic in their interest in foreign food and drink, especially if those types of food and drink are popular in countries where most of the immigrants came from. Overall, countries with fewer immigrants have more idiosyncratic tastes in food and drink, possibly due to limited exposure to foreign cuisines. Moreover, in a majority of countries, larger immigrant populations are associated with more similar food and drink preferences between their countries of origin and their destination countries.

Despite the multiple advantages of our proposed methodology using the Facebook Ads data to study cultural similarity, it is important to point out some of the limitations of this approach. First, our analyses are limited to correlational data. Overall, we observed that in a majority of countries, larger immigrant populations are associated with more similar food and drink preferences between their countries of origin and their destination countries. Our hypothesis is that immigrants help bring the culture of their home countries to new countries (Algan et al., 2012). More than just a unidirectional process of acculturation, we believe that there is a bidirectional relationship between culture and migration where cultural similarity influences the decision to migrate (Caragliu et al., 2013; White, 2013) and the process of migration then leads to increased acculturation and cultural similarity (Berry, 2006; Bierbrauer and Pedersen, 1996; Mesoudi, 2018). However, studying the causal pathways between cultural similarity and migration would require longitudinal data instead of a snapshot of a single year. To investigate the causal relationship between cultural similarity and migration longitudinally, we propose to automate the data collection to collect data from Facebook Ads annually. Second, the mechanisms behind the Facebook Advertising Platform are a black box and some information, such as interests, can be inferred by Facebook based on unknown algorithms. Previous research showed that demographic information, such as sex, location, and age, is accurately reported by Facebook users or estimated by Facebook (Grow et al., 2021). As Facebook's business model relies on targeted advertisements related to interests, we expect that Facebook's models identify users' interest in a fairly accurate way. However, it would be important for future research to independently validate whether interests, too, are accurately captured by the Facebook Advertising Platform. Third, our conclusions are limited to the sample of countries selected. The cultural similarities between countries, as well as the correlation between cultural similarity and migration, may differ depending on the subset of countries chosen, which is also constrained by our data source (e.g., Facebook is not used in China). However, our measure of asymmetric cultural similarity between any two countries in our sample would be unaffected if we included other countries because this measure is defined solely on the basis of top dishes in these two countries. When it comes to other metrics and other parts of the world, future research should investigate the extent to which

our conclusions can be generalized by replicating our approach for more countries, as well as by considering additional metrics. An exciting area of further development includes incorporating culture and our methods within the perspective of social network analysis. Within this context, shared interests in food and drinks can be seen as a defining component of relationships between countries. This approach would build and enrich the growing literature on identifying the role and boundaries of migration networks (Abel et al., 2021; Fagiolo and Mastrorillo, 2013; Massey and España, 1987; Tranos et al., 2015) and the relationship between cultural production and migration (Baily and Collyer, 2006). Finally, as a matter of fact, it is important to emphasize that while the cuisine of a country is an important cultural marker to study cultural similarity, the proposed methodology could be extended other types of attributes and interests, which might be relevant for studies with other goals. In this sense, future research can shed new light on the importance of additional cultural markers to characterize and measure cultural similarity between countries, as well as the network structure of relationships between countries.

The methodology that we proposed in this article is based on a scalable approach that makes use of social media data to study the cultural similarity between countries. The proposed measure of cultural similarity, even if it is just focusing on food and drink as cultural markers, seems to capture a substantial portion of the variability measured in data from the WVS. In addition to that, our measure of cultural similarity takes into consideration factors such as international data availability, time, cost, and asymmetry, which have hampered previous efforts to study cultural similarity. Finally, the high correlation between cultural similarity between countries and the number immigrants in the country suggests that cultural similarity related to food and drink preferences can be considered among relevant predictors of migration, in addition to more established quantities which include economic variables, trade (Egger et al., 2012) or indicators of cultural diffusion in networks, such as communication and the flow of information (Poot, 1996).

CHAPTER 3

# Evaluating the Impact of Cultural Similarity on Migration Prediction

## 3.1 Introduction

One of the strongest empirical regularities in spatial demography is that flows of migrants are positively associated with population size at origin and destination and inversely related to distance. This pattern was observed in the 19th century by Ravenstein (1889) and later formalized by Zipf (1946) into what are known as gravity models of migration. Traditionally, distance is measured geographically. However, other measures including those based on economic and cultural factors have also been found to be relevant for explaining migration flows (Anderson, 2011; Böhme et al., 2020; Caragliu et al., 2013; Esses, 2018; Lewer and Van den Berg, 2008).

The cultural distance between two countries could therefore be a valuable predictor of migration flows given the bi-directional relationship between culture and migration. For instance, cultural fit in terms of language, norms, and values are important factors that people consider before moving between countries (Caragliu et al., 2013; Pedersen et al., 2004). After moving, migrants then transmit cultural elements, such as food habits (Opare-Obisaw et al., 2000), from their origin country to their destination country and back (Mesoudi, 2018).

Measures of cultural distance are difficult to estimate and thus have not yet been widely adopted in gravity models for assessing and predicting migration. The few studies that have examined the impact of cultural dimensions or cultural distance on migration flows have typically relied on survey responses regarding norms, values, and beliefs, such as from the World Values Survey (Inglehart, 1997). Immigration data from Denmark, Germany, and the Netherlands illustrate that higher cultural distance derived from the World Values Survey is associated with less long-term mobility (White, 2013). Overall, cultural distance derived from the World Values Survey seems to hold an important role in predicting migration flows between European countries (Caragliu et al., 2013). However, survey approaches for migration suffer from significant limitations, such as the difficulty of reaching migrants, and the complexity and high costs associated with running a cross-national survey with a migration focus.

In this paper, we complement measures of cultural similarity based on cultural norms, values, and beliefs derived from surveys (i.e., the World Values Survey) for the study of migration flows. We expand the analysis of the impact of culture on migration flows by adding measures of cultural similarity based on cultural attributes regarding food and drink interests derived from social media data (i.e., Foursquare and Facebook Ads). This article cannot distinguish between all the mechanisms underlying the complex relationship between culture and migration, and the aim of this study is not to establish a causal link. We document how culture is an important aspect in the study of migration, and how measures of cultural similarities improve predictions of migration flows even after accounting for classic predictors of migration. We expand the literature by showing the impact of adding measures of cultural similarity derived from social media data based on food and drink interests (i.e., food and drink similarity).

Food and drink are two of the most basic needs of human beings. The manner in which people interact with food, from the procurement and selection of food to its preparation and consumption, reflects complex interrelationships and interactions among individuals, the society in which they live, and their culture (Alexander et al., 2019; Ferguson et al., 2020). Food studies became an important interdisciplinary field of study that focuses on the relationship of food and human experience, and the relationships between food, culture, and society (Almerico, 2014). Food production, distribution, and consumption are all shaped by cultural codes (Counihan et al., 1997) and represent a cultural act (Montanari, 2006). Food preparation [10] is one of the topics included in the Lists of Intangible Cultural Heritage provided by UNESCO [11] covering cultural practices and expressions of intangible heritage. Overall, food is used as an important marker of cultural identity (Kittler et al., 2016). Out of all categories of interests on Facebook (e.g., food and drink, news and entertainment, hobbies and activities, sports and outdoors), food and drink are the only interests that belong to a universal category given that food and drink are two of the most basic needs of human beings. Some interests are specific to certain demographic groups. For instance, not everyone is interested in sports or celebrities. However, food is popular across a wide demographic spectrum. In this context, given the importance of food to culture (Ashley et al., 2004; De Solier and Duruz, 2013; Recchi and Favell, 2019), we consider measures of cultural similarity, derived from social media data, based on food and drink interests.

We propose the use of measures of food and drink similarity developed by Vieira et al. (2022c) and evaluate their potential in predicting migration. These measures are timely, cost-effective, and scalable, and use aggregate data from Facebook users' food and drink interests that are freely and publicly available through the Facebook Advertising Platform (we will refer to these data as Facebook Ads data). We illustrate the applicability of the proposed approach by

[10] Although Facebook provides a range of interests broadly related to food and drink, most of those interests do not represent food as naturally found in nature (e.g., grapes, corn). Additionally, Vieira et al. (2022c) manually validated the data set by removing interests such as restaurants and brand names. The majority of the interests related to food and drink on Facebook represent dishes or any food or drink processed by humans (e.g., wine, quesadilla).

[11] https://ich.unesco.org/en/lists?term[]=vocabulary_thesaurus-10

showing how these new measures of food and drink similarity can be used to predict migration and explain migration flows. The measures of food and drink similarity derived from Facebook Ads have, despite their limitations, a predictive capacity of migration flows that is comparable to classic variables used in the literature to represent the cultural dimension, such as shared language and shared history. Additionally, the measures of food and drink similarity derived from Facebook Ads are able to capture changes quickly, especially when migration changes rapidly due to crises. In this context, we expect that these measures of food and drink similarity derived from Facebook Ads could represent, almost in real-time, the cultural changes during big and unexpected migration events (e.g., the migration of Ukrainians after Russia's invasion). Further, the Facebook Ads measures of food and drink similarity introduce a more nuanced view of symmetric and non-symmetric measures of similarity, opening new opportunities for further predicting and understanding the determinants of migrations.

## 3.2 Background

Cultural distance measures operational parameters that can be used as proxies for cultural dimensions. They allow researchers to estimate the extent to which countries differ culturally (Tung and Verbeke, 2010). The cultural dimensions used to measure culture can vary depending on the focus of the research (Mohr et al., 2020). For instance, the study of culture can focus on aspects of our daily life by considering cultural objects such as the clothes we wear, the music we listen to, and the food we eat (Kwantes and Glazer, 2017; Recchi and Favell, 2019). Food studies are an important interdisciplinary field that recognizes food as a central aspect for acculturation, cultural practices, and cultural identity (Ashley et al., 2004; De Solier and Duruz, 2013; Ferguson et al., 2020; Kittler et al., 2016; Montanari, 2006).

Operationally, culture has been traditionally measured in terms of norms, values, and beliefs via sampling surveys (Kwantes and Glazer, 2017) in which the survey responses are used to characterize cultural aspects of a country (e.g., Schwartz's value survey (Schwartz, 1994), the World Values Survey (Inglehart, 1997), and Hofstede's [12] cultural characteristics (Hofstede, 1983)) and to evaluate relative distance between countries (De Santis et al., 2016; Gupta et al., 2002; Mucciardi and De Santis, 2017; Muthukrishna et al., 2020).

Such studies based on surveys are highly valuable, but also have important limitations. In addition to measurement error (Groves and Lyberg, 2010), results may suffer from various biases (Suchman, 1962) like social desirability bias, question order bias, and acquiescence bias. Furthermore, surveys are costly and require a long time to run. For example, most government statistics are updated only once a year (often with a delay), and major surveys such as the European Value Study (EVS) and the World Values Survey (WVS) are often spaced even further apart (the EVS every 9 years and the WVS every 5 years). This lack of timely information makes it diffi-

[12] https://www.hofstede-insights.com/models/national-culture/

cult for decision-makers to respond dynamically to shifting circumstances. Overall, uncertainty and complexity are involved in demographic studies, in particular, migration predictions (Bijak, 2022). For instance, migration flows related to refugee movements [13] is one of the most volatile and therefore the most difficult to predict. These dynamic shifts in migration flows illustrate that we need more frequent data to track migration flows and improve predictions. To overcome part of these limitations, we propose an approach that relies on passively-collected data from social media, and that can complement existing sources.

Social media advertising platforms provide complementary tools that can be used to measure cultural preferences and allow comparisons across regions via passively collected data (You et al., 2017). As one of the first studies to address this question by using online data sources, Silva et al. (2014) identified cultural boundaries and similarities across populations by clustering them based on the analysis of food and drink habits. However, their analyses of culinary habits around the world were limited to Foursquare check-ins, which considered only 101 categories and, consequently, underestimated the variety of users' interests.

In a first study in this area using Facebook Ads data, Vieira et al. (2020) examined the similarity between selected countries and Brazil based on their population's interests in typical Brazilian dishes. However, the results were limited to dishes listed on Wikipedia. This limits the potential list of dishes and, importantly, some countries do not have a Wikipedia page dedicated to listing their typical dishes and therefore, the methodology was not scalable. Obradovich et al. (2020) also used data from Facebook Ads to examine cross-national cultural differences across nearly 60,000 interests. They validated their work by comparing the cultural distance calculated from their measurement with traditional survey-based measures. However, the authors included different types of features, from politics to national parks, and used a 'black-box' model to represent countries' cultures, which makes it hard to assess what exactly their index is measuring. More recently, Vieira et al. (2022c) presented a scalable, data-driven methodology to measure the cultural similarity between countries in terms of the most popular food and drink in each country from a list containing more than 200,000 interests on Facebook Ads (Speicher et al., 2018). Compared to Obradovich et al. (2020), the methodology proposed by Vieira et al. (2022c) compared countries in terms of fewer, but explicitly known attributes selected from a very large dataset. In other words, interests that are not relevant to any of the countries were disregarded to reduce feature sparsity. They presented two measures of cultural similarity, including the first asymmetric measure of cultural similarity derived from social media data.

The literature that we just summarized suggested methodologies based on social media data to measure cultural similarity between countries, and then correlated these measures with survey-based measures. To the best of our knowledge, ours is the first paper that evaluates how suitable the use of an asymmetric measure of similarity is to predict migration. In this work, we decided to evaluate the impact of adding these measures of cultural similarity to a gravity model

[13] *https://ourworldindata.org/explorers/migration?time=latest&facet=none&Metric=Net+migration+rate&Period=Total&Sub-metric=Total*

to predict migration. In order to test the Facebook measures against the most stringent baseline possible, we tested its predictive capacity in comparison with measures of cultural similarity derived from the World Values Survey (Inglehart, 1997) and Foursquare data (Silva et al., 2014).

Facebook Ads data have become an important tool in demographic research, especially for studying migration patterns (Alexander et al., 2019; Dubois et al., 2018; Leasure et al., 2023; Palotti et al., 2020; Spyratos et al., 2019; Zagheni et al., 2017). However, the study of international migration and the development of models to explain and to predict flows of people between countries is not new (Massey et al., 1993) and one of the most important approaches in traditional prediction is based on gravity-type models (Cohen et al., 2008; Lewer and Van den Berg, 2008; Ramos, 2016; Tinbergen, 1962). For example, Cohen et al. (2008) developed an algorithm to project future numbers of international migrants from any country or region to any other. The model considers the population and geographical area of origin and destination countries, and the geographic distance between origin and destination. Subsequently, researchers have added to this model by pointing out other variables, such as social variables (e.g., mortality rate) (Kim and Cohen, 2010), historical variables, such as shared history and shared language (Abel et al., 2019; Beine et al., 2016; Caragliu et al., 2013; Kim and Cohen, 2010; Lewer and Van den Berg, 2008), and online search keywords (Böhme et al., 2020). For instance, Böhme et al. (2020) showed how geo-referenced online search data can be used to measure migration intentions in origin countries and to predict bilateral migration flows. Moreover, distance measures beyond the geographic distance, like administrative, political, economic, or cultural distance (Ghemawat, 2001) are important variables that should be considered by models to predict migration.

Most of the studies that have so far analyzed cultural changes in relation to migration were restricted to one or a few countries or, when taking a broader international perspective, used cultural distance measures that are by construction symmetric (Rapoport et al., 2020). Since migration is neither homogeneous across countries nor symmetric, we apply an asymmetric measure of cultural similarity across many countries to more accurately represent processes of international cultural exchange.

## 3.3 Data

In this section, we describe the main data sources used to collect international data used in our prediction models. We present the data we used to measure the cultural and food and drink similarity between countries: the World Values Survey (Inglehart, 1997), Foursquare (Silva et al., 2014), and Facebook Ads (Vieira et al., 2022c). We also include a description of data sources for gravity model variables such as population, area, and geographic distance, as well as migration flows as the outcome variable. For comparability with previously suggested indices of similarity based on Foursquare data, our analysis focuses on a subset of sixteen of the most popular countries by number of Foursquare check-ins (Silva et al., 2014).

The countries selected for the analysis have been chosen based on a compromise across three different criteria. First, to match the list of countries selected by Silva et al. (2014) to make possible a comparison between the previous literature that uses Foursquare data with our results using Facebook data. Second, these countries cover a large portion of geographic areas and populations across the world. Finally, and importantly, our selection of countries reflects countries with high Facebook penetration rates: Argentina, Australia, Brazil, Chile, Great Britain, France, Indonesia, Japan, South Korea, Malaysia, Mexico, Russia, Singapore, Spain, Turkey, and the US. In other words, we favored a choice of countries with comparatively low and consistent biases over a broader selection that would be more heterogeneous in terms of biases and for which interpretation of results would be more complex.

### 3.3.1 Facebook Ads data

Vieira et al. (2022c) collected data regarding Facebook users' interests in food and drink, and proposed measures of cultural similarity between countries. The data collected from Facebook Ads refers to the number of Facebook monthly active users (i.e., active over the past 30 days) who match the demographic attributes targeted at the time of data collection. "The Facebook Marketing API allows marketers and researchers to obtain an estimate of the number of monthly active users for a proposed advertisement that matches given input criteria (Kosinski et al., 2015). For that, the platform provides a list of demographic attributes, such as age, gender, home location, and interests that can be customized by the advertiser as the input query. Thus, after specifying the target audience, and before the ad is launched, advertisers are provided with the size of the audience that matches the target specifications. Attributes like age, gender, and location are explicitly declared by the users in their profiles, whereas interests can be either declared by the user or inferred by Facebook based on user activities such as posting or interacting with contents (e.g., liking content, sharing content, or updating one's status). Facebook users generate traces of their preferences in multiple domains." (Vieira et al., 2022c). Their methodology consisted of selecting a subset of popular food and drink for each country, and then creating a vector representation according to their Facebook users' interests in those food and drinks. Finally, they measured the similarity between those country-level vectors. Methodological details are available in Vieira et al. (2022c). [14] Measures derived from the Facebook Ads data including the code to analyze the datasets and generate the figures are available in a public web repository. [15]

Two measures of similarity were proposed by Vieira et al. (2022c) – **Facebook Asymmetric Similarity** and **Facebook Symmetric Similarity** – depending on the subset of food and drink used to create the vector representations. The *Asymmetric Similarity* between two countries, $c_1$ and $c_2$, is measured in terms of the most popular food and drink in $c_1$ whereas the similarity between $c_2$ and $c_1$ is measured in terms of the most popular food and drink in $c_2$. In this case,

[14] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262947

[15] https://github.com/carolcoimbra/cultural-similarity-fb

since the similarity between $c_1$ and $c_2$ is different from the similarity between $c_2$ and $c_1$, this measure is not symmetric.

However, we could also measure the similarity between two countries by considering a fixed set of interests for both countries. In this case, we can refer to the measure as *Symmetric Similarity*, corresponding to the measure of similarity between two countries considering the union of the most popular food and drink in these countries. Since the subset of interests is fixed, the similarity between $c_1$ and $c_2$ is equal to the similarity between $c_2$ and $c_1$. In our models, we refer to the Facebook asymmetric measure of similarity as **Facebook asymmetric similarity** – food origin or food destination – depending on which subset of top food and drink, from the country of origin or destination, we considered in the measure of similarity. We refer to the symmetric measure of similarity as **Facebook symmetric similarity**.

For the Facebook measures of food and drink similarity, Vieira et al. (2022c) selected the top 50 types of food and drink in each country. [16] In this case, the asymmetric measure of similarity between two countries, $c_1$ and $c_2$, corresponds to the cosine similarity between the 50-dimensional vector representation of each country in terms of the 50 top food and drinks in the country $c_1$. The symmetric measure of similarity between two countries, on the other hand, is given by the cosine similarity between the vector representation of each country in terms of the 394 food and drinks. The set of 394 interests corresponds to the union of the top 50 interests in each one of the 16 countries.

The focus of this paper is to assess the extent to which the considered Facebook measures of cultural similarity – using only food and drink as cultural markers – are meaningful predictors of migration flows. In this sense, it is important to validate our results obtained with the Facebook Ads data and ensure comparability with prior research. We selected the two most relevant datasets to compare measures of cultural similarity: the World Values Survey and Foursquare. The World Values Survey is an established and traditional dataset based on large-scale representative survey data along several cultural dimensions regarding cultural norms, values, and beliefs. The Foursquare data (Silva et al., 2014) is based on users' food and drink habits collected from Foursquare check-ins. Although the measures of cultural similarity from the World Values Survey and Foursquare data focus on different aspects of culture, both measures are used as baselines for the Facebook measures of food and drink similarity. However, as mentioned before, both carry significant limitations. For the World Values Survey, the main disadvantages involve the costs and operational time needed to release new waves, which has typically been about 5 years. The Foursquare platform, in contrast, is not as widely used as Facebook and is heavily biased from a demographic point of view (e.g., young men are more likely to use

[16] We conducted additional analyses to show how stable the results are when we vary the number of interests we consider in the Facebook measures of similarity. The top 50 foods and drinks generate the best results considering all the calculated metrics, such as the adjusted R-squared, and significant coefficients.

Foursquare than women or older people [17] [18] and most users are from the US [19], although only 2% of the American population uses Foursquare [20]), and limited to people who explicitly share their locations when visiting a place (i.e. check-in). Moreover, access to the Foursquare API is not free. [21]

### 3.3.2 World Value Survey (WVS) data

The World Values Survey dataset considers several cultural dimensions such as religion, politics, economics, and lifestyle. Inglehart (1997) identified two major dimensions derived from the World Values Survey and proposed a cultural map [22] of the world where the location of each country is given by the scores on these two dimensions. Based on the location of each country in the World Values Survey cultural map from 2020, each country was represented by a vector of two dimensions corresponding to the two dimensions of the World Values Survey cultural map: Traditional values (which emphasize the importance of religion, parent-child ties, deference to authority, and traditional family values) versus Secular-rational values (represented by societies which place less emphasis on religion, traditional family values and authority); and Survival values (emphasis on economic and physical security) versus Self-expression values (high priority on environmental protection, equality, and rising demands for participation in decision-making in economic and political life). Then, we measured the cosine similarity between each pair of countries to obtain the **World Values Survey similarity**.

The measure of cultural similarity derived from the World Values Survey is a representation of a more traditional measure of culture based on norms, values, and beliefs. The selection of the World Values Survey as a data source for measuring cultural similarities across countries is based on two main criteria: coverage and wave updates. The World Values Survey covers an extensive geographical and thematic scope, is conducted globally every 5 years, [23] and is considered one of the most authoritative and widely used cross-national surveys in social sciences. In addition to that, there is a high association between the cultural dimensions from the World Values Survey cultural map and other cultural dimensions derived from surveys (Kaasa and Minkov, 2022; Taras et al., 2009).

[17] ***https://brandongaille.com/26-great-foursquare-demographics/***

[18] ***https://www.statista.com/statistics/814726/share-of-us-internet-users-who-use-foursquare-by-age/***

[19] ***https://99firms.com/blog/foursquare-statistics/#gref***

[20] ***https://financesonline.com/foursquare-statistics/***

[21] ***https://foursquare.com/products/pricing/***

[22] ***https://www.worldvaluessurvey.org/WorldValuesSurveyEventsShow.jsp?ID=428***

[23] The most recent 7th wave of the World Values Survey (2017-2022) covers 80 countries.

### 3.3.3 Foursquare data

Silva et al. (2014) identified cultural boundaries and similarities across populations by clustering them based on the analysis of food and drink habits via Foursquare check-ins. They also propose a cultural map where the location of each country is given by the two first principal components after applying the Principal Component Analysis (PCA) algorithm over a high dimensional preference vector based on Foursquare check-ins in different sub-categories of bars and restaurants. Based on the location of each country in the cultural map proposed by Silva et al. (2014), we measured the cosine similarity between them to obtain the **Foursquare similarity**.

We compared different measures of similarity derived from Euclidean and cosine distance. Both measures are highly correlated with each other (WVS data (0.64) and Foursquare data (0.89); see figures A.2 and A.3 in the Appendix). We did not observe substantive changes in our model when using cultural similarities derived from Euclidean or cosine similarity. For consistency purposes, all the measures of cultural similarity derived from the World Values Survey, Foursquare, and Facebook Ads data are based on cosine similarity.

### 3.3.4 United Nations data

Population size of each country in 2019 [24] comes from the United Nations. The data include estimates of the total **population** for all countries and are made available via the 'World Population Prospects', the 2019 Revision. We also collected the estimates of the number of international migrants, **migrant stocks**, from each country of origin in each country of destination from 2019 from the United Nations website. [25]

### 3.3.5 CEPII GeoDist data

GeoDist (Mayer and Zignago, 2011a) makes available the exhaustive set of gravity variables developed by Mayer et al. (2005) to analyze market access difficulties in global and regional trade flows. The dataset incorporates country-specific geographical variables for the world's countries, including the **area** of each country in square kilometers ($km^2$). Moreover, the dataset includes variables that apply to pairs of countries, from which we used the geographic distance and shared history. The **geographic distance** is based on bilateral distances between the biggest cities of those two countries, weighted by the share of the city in the overall country's population between each pair of countries (Mayer and Zignago, 2011b). Finally, **shared history** is the name given to an indicator of whether the two countries ever had a colonial link. Based on the information from the dataset, the colonial link represents whether the two countries have had a common colonizer

[24] https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp

[25] https://www.un.org/development/desa/pd/content/international-migrant-stock

after 1945, have ever had a colonial link, have had a colonial relationship after 1945, are currently in a colonial relationship or were/are the same country.

### 3.3.6 CEPII Language data

The CEPII Language dataset (Melitz and Toubal, 2012) provides separate measures of common native language, common spoken language, common official language, and linguistic proximity between different native languages. In our model, we use an indicator of whether the two countries shared a **common official language**, and the **linguistic proximity** between two countries' languages. The first indicator is a binary variable that codes whether the two countries share at least one official language. The second indicator calculates linguistic proximity on the basis of the Ethnologue classification of language trees between trees, branches, and sub-branches (Fearon, 2003; Laitin, 2000). This variable can take four values: 0 for languages belonging to separate family trees; 0.25 for languages belonging to different branches of the same family tree (e.g., English and French); 0.50 for languages belonging to the same branch (e.g., English and German); and 0.75 for languages belonging to the same sub-branch (e.g., German and Dutch).

### 3.3.7 World Bank data

The gross domestic product per capita, **GDP per capita** (constant 2010 US$) of each country in 2019 was collected from the World Bank. [26]

### 3.3.8 Migration flow data

Finally, the dependent variable of our models is the migration flow by origin and destination country. Due to the lack of migration flow data at the global level, we use the estimated values from the Demographic accounting, pseudo-Bayesian approach proposed by Abel and Cohen (2019). The estimations are available only for five-year bilateral migration flows. For our analysis, we use the latest estimates of migration flows for the period 2015-2019.

## 3.4 Gravity models to predict migration

In order to investigate whether the measures of cultural similarity can improve the prediction of migration flows, we test different models inspired by traditional gravity models (Cohen et al., 2008). Although they have their limitations (Beyer et al., 2022), gravity models are among the most traditional models used to predict migration flows. The gravity model is a log-linear model [27] as shown by Equation 3.1. The classic model uses only the population of origin and

[26] https://databank.worldbank.org/home

[27] The logarithm scale used in this study is the logarithm base 10.

destination ($P_o$, $P_d$), the area of the country of origin and destination ($A_o$, $A_d$), and the geographic distance between origin and destination ($D_{o,d}$) as independent variables to predict migration flows ($M_{o,d}$) between the country of origin and destination:

$$log_{10}(M_{o,d}) = \beta_0 + \beta_1 log_{10}(P_o) + \beta_2 log_{10}(A_o) + \beta_3 log_{10}(P_d) + \beta_4 log_{10}(A_d) + \beta_5 log_{10}(D_{o,d}) + \epsilon_{o,d} \quad (3.1)$$

We added additional independent variables that might promote or deter migration based on prior literature that has identified them as relevant predictors (Anderson, 2011). In this way, we could test the extent to which measures of culture affect the predictive capacity of the model beyond the more stringent baseline. We examined a series of gravity models. The first model (Model 1) refers to the classic version of a gravity model (Cohen et al., 2008) represented by Equation 3.1 plus the GDP per capita of both countries of origin and destination, and migrant stocks between countries of origin and destination. All the independent variables included in Model 1 are referred to as basic variables as they are important predictors commonly added to gravity models (Böhme et al., 2020; Cohen et al., 2008; Tinbergen, 1962). Given the importance of common language and common colonial history (Abel et al., 2019; Beine et al., 2016; Caragliu et al., 2013; Kim and Cohen, 2010; Lewer and Van den Berg, 2008), the second model (Model 2) builds on the first model and adds variables related to shared language and shared history. Shared language and history (e.g., colonial history) are often included in gravity models as cultural variables. However, other types of cultural differences may also be relevant to predict and explain migration. For instance, differences in cultural norms, values, and beliefs between countries affect migration flows (Caragliu et al., 2013). To take into account the impact of cultural norms, values, and beliefs in a country on migration (Caragliu et al., 2013; Esses, 2018), we specify the third model (Model 3) by adding the World Values Survey cultural similarity measure to the second model.

Finally, we assess the impact of adding measures of food and drink similarity derived from social media data to each one of the three models presented. For instance, Model 2 + FB asymmetric adds the first asymmetric measure of similarity using data from Facebook Ads to Model 2. The Facebook asymmetric measure of similarity consists of two variables, Facebook Asymmetric Similarity (food origin) and Facebook Asymmetric Similarity (food destination), which represent measures of similarity by food and drink that are popular in the country of origin and food and drink that are popular in the country of destination, respectively. Model 2 + FB symmetric and Model 2 + Foursquare add, respectively, the symmetric measure of similarity derived from Facebook Ads data and the measure of similarity derived from Foursquare data to Model 2. Each measure of cultural similarity derived from social media data was added separately to the models to assess their individual impact in predicting migration flows. Due to the relatively high correlation between the Facebook measures (0.38) and between the Foursquare and Facebook measures of similarity (0.32 and 0.37), as shown in Figure 3.1, we only tested those measures separately, to reduce potential issues of collinearity.

| Models | RMSE | R-squared | MAE | WAIC* |
|---|---|---|---|---|
| Model 1: Area, Population, Distance, GDP per capita, Migrant Stocks | 0.61 | 0.78 | 0.46 | 440.17 |
| Model 1 + FB asymmetric | 0.59 | 0.79 | 0.45 | 430.88 |
| Model 1 + FB symmetric | 0.61 | 0.78 | 0.46 | 438.47 |
| Model 1 + Foursquare Similarity | 0.61 | 0.78 | 0.46 | 438.96 |
| Model 2: Model 1 + Shared language and history | 0.58 | 0.80 | 0.45 | 419.60 |
| Model 2 + FB asymmetric | 0.59 | 0.80 | 0.45 | 419.62 |
| Model 2 + FB symmetric | 0.58 | 0.80 | 0.45 | 419.82 |
| Model 2 + Foursquare Similarity | 0.58 | 0.80 | 0.45 | 418.44 |
| Model 3: Model 2 + WVS Similarity | 0.55 | 0.82 | 0.43 | 390.04 |
| Model 3 + FB asymmetric | 0.55 | 0.82 | 0.43 | 392.12 |
| Model 3 + FB symmetric | 0.55 | 0.82 | 0.43 | 391.63 |
| Model 3 + Foursquare Similarity | 0.55 | 0.82 | 0.43 | 390.09 |

**Table 3.1:** Overall prediction errors for each model evaluated using cross-validation.
∗ Metric applied just to the final model using the full input dataset (240 pairs of countries).

In all these models, the dependent variable is the logarithm of the estimated migrant flow in 2019. In order to calculate the logarithm of the dependent variable, we added an offset (equal to 1) to all the observations. The resulting coefficients and statistics for each one of these models are presented in Table A.1. In this table, the columns represent the models and the rows describe each of the variables in the prediction model.

To properly evaluate the models and predictions, while avoiding over-fitting, we decided to evaluate each model against a test sample of the data that was not seen by the model during the fitting phase that relied on the training data. Model fitting and evaluation were done only on the training data. The evaluation of predictive accuracy or errors was done on the testing data. For this evaluation, we used the Leave One Out Cross-Validation (LOOCV).

The Leave One Out Cross-Validation approach requires one model to be evaluated for each point in the training dataset. Given a dataset with N data points (in our case, pairs of countries), the cross-validation works in the following way: (i) train the model on N-1 data points, (ii) test the model against that one data point which was left out in the previous step, (iii) calculate prediction error, (iv) repeat the above 3 steps until the model is trained and tested on all data points, and (v) generate the overall prediction error (e.g., taking the average of prediction errors across all models).

Table 3.1 shows the average prediction errors for three different measures for each model. The rows represent the models and the first three columns represent the three measures of prediction errors. The Root Mean Squared Error (RMSE) is a proxy for the average difference between the predictions made by the model and the actual observations. The lower the root mean squared error, the more closely a model can predict the actual observations. The Mean Absolute Error (MAE) is the average absolute difference between the predictions made by the model and the actual observations. The lower the mean absolute error, the more closely a model predicts the actual observations. Finally, the R-squared is a measure of the correlation between the predictions made by the model and the actual observations. The higher the R-squared, the more variance in the data is explained by the model. Each of the three metrics provided in the

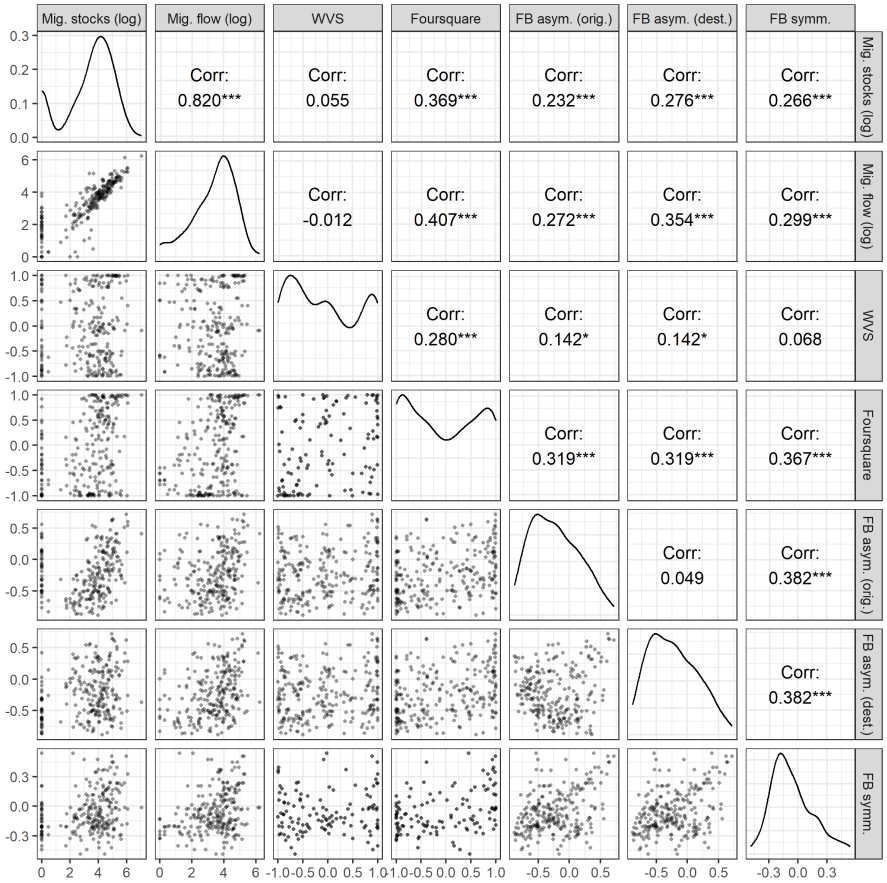



**Figure 3.1:** Distribution and correlations between the measures of similarity and migration flows (in logarithm scale) between countries. Each dot represents a pair of countries within the 16 countries we analyzed.
$***p < 0.001$; $**p < 0.01$; $*p < 0.05$

output gives us an idea of how well the model performed on previously unseen data. However, the R-squared measure itself is not reliable since more variables always increase the metric, even if new variables are only marginally predictive. To address this, we included other measures that penalize the number of variables in their calculation. The last column in Table 3.1 shows the Watanabe–Akaike or Widely Applicable Information Criterion (WAIC) (Gelman et al., 1995) for each one of the models tested. We use the full input dataset (without cross-validation), which consists of 240 pairs of countries. The lower the Watanabe–Akaike, the more closely a model can predict the actual observations. Table A.1 shows also the adjusted R-squared and the resulting coefficients and statistics for each one of these models. In this table, the columns represent the models and the rows describe each of the variables in the prediction model. In the next section, the resulting coefficients and statistics from Table 3.1 and Table A.1 are described in more detail.

## 3.5 Results

Table A.1 shows in detail all the coefficients for each one of the variables included in the gravity models that we tested using the full input dataset corresponding to 240 pairs of countries. Table 3.1 shows the results averaged across cross-validations, except for the WAIC, which was calculated from the model using the full input dataset. To evaluate the impact of adding measures of cultural similarity to the migration model, we first assess the correlation between the variables considered. Figure 3.1 shows the correlation between each one of the measures of similarity and migration flows (in logarithm scale) between each pair of countries within the 16 countries we analyzed. We observed that the symmetric and asymmetric measures of food and drink similarity derived from Facebook Ads data showed a positive correlation (0.38). Similarly, the measures of food and drink similarity based on Foursquare and on Facebook Ads data were also positively correlated (0.37 between Foursquare and the Facebook symmetric measure, and 0.32 between Foursquare and the Facebook asymmetric measure). Foursquare and Facebook measures of food and drink similarity are highly correlated with each other and capture similar patterns regarding food and drink interests across countries. Next, we compared the cultural similarities derived from social media versus the World Values Survey. Although cultural similarities based on the World Values Survey and the Facebook symmetric measure do not capture the same cultural attributes and were not substantially associated (0.07), the World Values Survey was positively correlated with both Facebook asymmetric (0.14) and Foursquare measure of food and drink similarity (0.28).

The measures of food and drink similarity derived from social media data do not exhibit a strong correlation with metrics obtained from survey data. Whereas the metric derived from the World Values Survey encompasses culture in terms of norms, values, and beliefs, the metrics derived from Foursquare and Facebook data primarily emphasize the interest in food and drink as cultural markers. The difference in the data nature suggests that the World Values Survey cultural similarity captures different aspects of culture that are not reflected by interests in food and drink from social media data. The measures derived from Foursquare and Facebook data, on the other hand, are highly correlated to each other and capture a similar pattern on food and drink interests across countries.

We observed that cultural similarity based on the World Values Survey was not positively correlated with migration flows (-0.01). This result means that countries that are close to each other in the World Value Survey cultural map have slightly less migration flows between them. Table A.1 shows a significant negative effect of the World Values Survey cultural similarity on migration flows, meaning that high World Values Survey cultural similarity was associated with less migration flows. Despite the unexpected negative effect of the World Values Survey cultural similarity on migration flows, the adjusted R-squared improved (0.83) when the World Values Survey cultural similarity was added to the gravity model. This result confirms the importance of cultural norms, values, and beliefs in a country when fitting migration models (Caragliu

et al., 2013; Esses, 2018). In contrast, the measures derived from social media data all showed a positive correlation with migration flows (Foursquare 0.41; Facebook Ads asymmetric 0.27 and 0.35, Facebook Ads symmetric 0.3), which means that a high food and drink similarity was associated with larger migration flows.

Even with this stringent baseline of adding geographic and economic variables, we found that including measures of cultural similarity derived from Facebook data focusing on food and drink improved predictions beyond what could be achieved with all these other predictors. The coefficients from Model 1, including the Facebook measures of food and drink similarity, are statistically significant and the predictive capacity of the model increased. This suggests that the Facebook measures of food and drink similarity are an important predictor of migration, capture different patterns, and, additionally, enable us to identify directional processes. The same does not hold for the model that includes all the basic variables and shared language and history (Model 2). In other words, we did not observe improvements when adding the Facebook measures of food and drink similarity, which means that there is likely an overlap in the explanatory power of the measures of food and drink similarity derived from Facebook data and shared language and history. Figure A.1 in Appendix shows the significant positive correlation between the measures of food and drink similarity derived from Facebook data and shared language and history. The estimated coefficients in Table A.1 for the measures of cultural similarity based on interests in food and drink derived from Facebook data in Model 2 are not significant. Finally, even though the measures of cultural similarity derived from the World Values Survey and social media data capture different aspects of culture, we did not observe significant coefficients when we added them the model that includes all the basic variables and the World Values Survey cultural similarity (Model 3). However, we observed a slight improvement in the adjusted R-squared compared to Model 3 when the measures derived from Foursquare data and the asymmetric measure of food and drink similarity derived from Facebook data were added to the model.

Despite the low improvement in migration flow prediction for the time point considered, the measures of food and drink similarity derived from Facebook data have a predictive capacity comparable to classic variables used in the literature, such as shared language and shared history. In addition to that, the measures of food and drink similarity derived from Facebook data contribute to predictive models of migration by adding not just a dynamic, but also an asymmetric component. The measures of similarity from Facebook are measuring country-level indices of cultural interests, which can shift precisely through migration. While systems of belief within a single culture (e.g., the majority culture in a country) should not change quickly, the ratio of majority culture to minority culture(s) can shift, leading to changes in country-level interests that would be observable in digital trace data. Particularly given the limitations of other measures of cultural similarities, Facebook Ads data offers an effective way to capture such changes in a way that complements other measures.

Figure 3.2 shows a comparison between the expected migration flows estimated by Abel and Cohen (2019) and the migration flows predicted by each model. The orange line represents

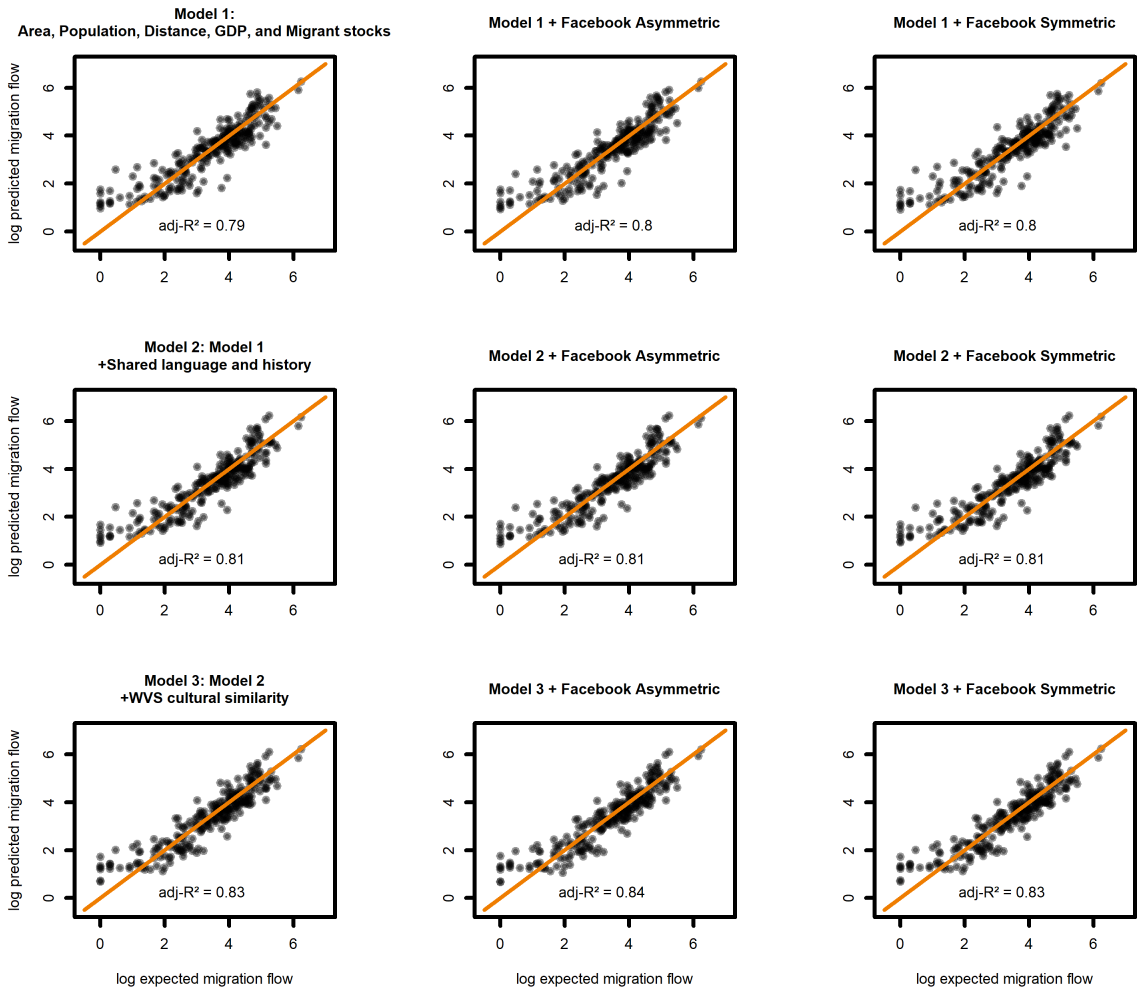



**Figure 3.2:** Comparison between the expected migration flows (x-axes) and the migration flows predicted (y-axes) by each one of the models using the full input dataset (240 pairs of countries). Both axes are on a logarithmic scale. Each dot represents a pair of countries within the 16 countries we analyzed.

the expected distribution, where the predicted migration flow is equal to the expected one. The distribution of the dots, which corresponds to pairs of countries, changes from one model to the other, and the predictions become closer to the expected values for migration flows. Overall, we observed that the baseline and more traditional models overestimate migration flows for pairs of countries between which there is little migration, and underestimate migration flows for pairs of countries between which there are larger migration flows. This pattern becomes slightly less evident with the inclusion of other variables, including the measures of food and drink similarity derived from Facebook.

## 3.6 Discussion

We showed the impact of adding measures of cultural similarity, both derived from surveys and social media data, to gravity models in order to predict migration. Our results indicate that the measure of cultural similarity derived from the World Values Survey contributes to improving migration prediction and the measure of food and drink similarity derived from Foursquare data is highly correlated with migration flows. However, in terms of scalability and reproducibility, they may have some disadvantages. As mentioned before, surveys are costly and require substantial operational time. For example, the World Values Survey runs every 5 years. Foursquare data have a different set of limitations, including the fact that the Foursquare platform is not as widely used as Facebook and is heavily biased from a demographic point of view. Moreover, the Foursquare dataset considered is over 5 years older than the data from Facebook Ads, and over 6 years older than the World Values Survey. During this period of time, significant cultural changes may have happened, given that the world is continuously changing in terms of connectivity across regions. With more than 2.7 billion worldwide users, [28] Facebook captures a larger and more diverse population than other social media. Considering Facebook's data availability, our methodology could be easily scaled to consider more countries. Moreover, the data from Facebook Ads are freely available and can be continuously updated and collected, which makes the approach timely, cost-effective, reproducible, and scalable. This will increase the relevance of these types of analysis in traditionally data-poor contexts, like in low- and middle-income countries.

Given the advantages of using Facebook Ads data, we provide a stringent test of the Facebook measures' incremental effects and our results show that cultural similarity, as measured by food and drink interests, explains migration flows in a way that is comparable to standard predictors such as shared language and shared history. Besides the advantages of using Facebook data to measure food and drink similarity, the use of an asymmetric measure of similarity adds more advantages to the approach presented. In fact, most of the gravity models consider symmetric variables to predict migration, which is itself an asymmetric phenomenon. Since the migration flow between countries is asymmetric (e.g., there are more Chileans in Spain than Spaniards in Chile), we should expect that the similarity in terms of food and drink interests will be asymmetric as well (e.g., there are more Chileans interested in Spanish food than Spaniards interested in Chilean food). We see this result reflected in the coefficients of the model using the Facebook asymmetric similarity, which shows a stronger effect of how popular the destination country's dishes are in the country of origin than vice versa. For example, Chileans' interest in Spanish dishes would be a stronger predictor of how many Chileans move to Spain than Spaniards' interest in Chilean dishes. We find evidence of asymmetric patterns, such that the cultural markers in the country of destination are more closely associated with migration flows

[28] *https://www.facebook.com/iq/insights-to-go/2740m-facebook-monthly-active-users-were-2740m-as-of-september-30/*

than cultural markers in the country of origin. We propose that future research in this area should take asymmetry into account when predicting migration.

To the best of our knowledge, we present the first study that considers a scalable, rapidly available, and asymmetric measure of similarity derived from social media data to predict migration. Our findings contribute to the literature by (i) showing the importance of cultural similarity derived from food and drink interests in social media data for predicting migration and (ii) allowing rapid predictions of current migration flows ahead of official statistics. For instance, Leasure et al. (2023) leveraged data from Facebook Ads to monitor in real-time subnational population sizes and internal displacement in Ukraine, on a daily basis, and disaggregated by age and sex. Similarly to Leasure et al. (2023)'s work, our methodology could capture rapid changes in populations' interests across countries, for instance, due to unexpected migration, and could help in predicting migration flows.

The primary objective of this study was to assess the value of examining cultural similarity in studying migration. Specifically, we aimed to test measures of cultural similarity based on food and drink interests in social media to predict international migration flows. Measures of cultural distance are difficult to estimate and thus have not yet been widely adopted in gravity models for assessing and predicting migration. However, culture plays an important role in the processes of migration. In addition, culture may be especially important to study because cultural attributes from our daily life, as well as migration flows, are dynamic and sensitive to change, in comparison to the static variables that are typically used in gravity models to predict migration, such as the area and geographic distance between two countries (Cohen et al., 2008).

As mentioned before, the relationship between migration and culture is likely bidirectional since cultural fit is an important factor that people consider before moving between countries, and migrants transmit cultural elements from their origin country to their home country and back during migration. In this paper, we focused on showing how measures of cultural similarity derived from Facebook users' food and drink interests can be used to explain migration flows between countries. For instance, imagine that the number of Facebook users living in the US interested in some traditional dishes from Brazil increases. One possible reason would be that the number of Brazilian immigrants in the US increased and people are getting exposed to Brazilian interests. In this example, if these Brazilian immigrants establish a big Brazilian community in the US, the number of Brazilian immigrants would potentially increase even more. In this case, the number of Facebook users interested in Brazilian food and drink works as a proxy for the Brazilian community established in the US. One of our main results shows exactly the importance of the cultural similarity between countries in terms of their Facebook users' interests in food and drink to predict migration flows between these countries. [29] While our study has broader ramifications,

[29]We conducted additional analyses to investigate the role of the immigrant community in the host country in shaping the significant coefficients observed for cultural similarities. We added migration stocks from 2019 to all the models and observed that overall, the coefficients regarding cultural similarities decreased by 30%, but were still significant. This result indicates that even though food and drink from the origin country may have been introduced to the destination country by immigrants, the interest in

the scope of this article is more limited, as we showed the positive association between cultural similarity and migration flows, without attempting to establish a causal direction in this complex bi-directional relationship. Future studies could collect additional data, and develop methods, in order to address this issue and move towards more causal estimates of the directions between culture and migration flows.

Caution should be exercised when interpreting our results due to their limitations, which we would like to acknowledge. First, the present analysis is constrained by data availability: only 16 countries were included in our analysis. The 16 countries selected for the analysis have been identified to match the list of countries chosen by Silva et al. (2014), in order to compare the results from different types of social media data. Besides enabling a comparison with other studies, these 16 countries cover a large and diverse portion of the world's regions and reflect countries where the Facebook penetration rate is high, thus reducing the potential size of the biases in the data. The data collection could be extended to more countries. That said, while the Facebook audiences' interests for the most current period could be collected, migration data remain a crucial bottleneck. The last time period for which global estimations of migration flow data are available from our main source (Abel and Cohen, 2019) is 2015-2019. In other words, we do not have migration flow data, or not even migration stock data, after 2019. Moreover, the COVID-19 pandemic affected migration and we do not have updated data that we could use as a dependent variable in our models. Once new migration data are available, new data from Facebook can be collected in real-time, and the predictions can be updated.

The measures of similarity that we used relies only on data regarding Facebook users' interests in food and drink. Although the cuisine of a country is an important cultural marker to study cultural similarity, the proposed methodology could be used with other types of attributes and interests, which might be relevant for studies with other goals or angles. We expect that a broader operationalization of measures of culture would lead to models with even higher predictive accuracy. In this sense, what we showed is likely a lower bound in terms of predictive capacity.

Besides that, social media data, including the Facebook Ads data, and the interest categories provided by Facebook, may be neither exhaustive nor representative. Facebook data include a number of biases since the users who use Facebook are not necessarily representative of the underlying population in their respective countries. There is a growing literature that has expanded our knowledge on how to identify and correct biases in social media data (see Appendix).

In addition to representativity, the user classification into the categories provided on Facebook Ads could also be a source of bias. Grow et al. (2022) evaluated the bias regarding location, age, and gender on Facebook. The authors compared the information provided by participants

those food and drink cannot only be explained by the size of the immigrant population. In other words, the interest in food and drink from the origin country is spread across the population in the destination country.

of an anonymous online survey with Facebook Ads' classification of the same individuals. The results showed that about 86%–93% of respondents' answers matched Facebook's classification. Although location, age, and gender appear to be classified mostly in a correct way on Facebook Ads, the accuracy of classification regarding Facebook users' interests has not been tested systematically. We hypothesize that Facebook covers only a subset of users' interests, but future research is needed to assess the extent to which the representation of interests is accurate.

We would like to emphasize the importance of addressing issues such as biases in digital trace data, and we point the readers to the resources mentioned above for a series of approaches developed to tackle this problem. With our article, we are entering partially uncharted territories in terms of assessing biases related to our methods, as our approaches are novel and not yet part of the conventional toolbox. We hope that our study will further stimulate methodological research on identifying and correcting biases when studying cultural dimensions using social media data. While the reader should be aware that the data used in this article are not necessarily representative of the entire underlying populations, it should also be noted that the choice of focusing on 16 countries with high Facebook penetration rates limits the extent of the biases, favoring comparisons across countries where Facebook is used in relatively similar ways by comparable demographic segments of the population. It is noteworthy that, despite the biases, the predictive model performs very well. Once the biases are fully modeled, we expect that the predictive capacity can only increase. We hope that this article lays the foundation for further analyses that can help us better understand these data and their potential, especially in countries and contexts that have historically been data-poor.

## 3.7 Conclusion

In this paper, we discuss how measures of cultural similarity derived from surveys and social media data can be important variables to predict migration flows. We compared a measure of cultural similarity derived from the World Values Survey with measures of food and drink similarity derived from Foursquare and Facebook Ads data. By using the measures derived from the Facebook Ads data, we introduce a more nuanced view of symmetric and asymmetric measures of similarity and show how these measures of similarity can be used to explain migration flows between countries. Our results show that the Facebook measures of food and drink similarity hold an important role in predicting migration, comparable to standard predictors, such as shared language and shared history. Finally, while some variables such as shared language, history, and geographic distance are static and symmetric, cultural attributes from our daily life are sensitive to changes in the environment and can be represented as an asymmetric measure of similarity between countries, adding value to models of migration, from both the substantive and predictive perspectives.

CHAPTER 4

# Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

## 4.1 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." [30] Despite these large numbers, persons in need of international protection [31] – refugees, asylum-seekers, displaced persons, other persons in need of international protection, and stateless persons – constitute a distinct migrant subgroup that is often not fully captured by traditional data (Robinson, 1998).

Studies on forced migration, such as those observing patterns of movement and estimating forced migration flows, are important for enabling international organizations, governments, and humanitarian agencies to allocate resources more effectively and to provide timely protection. A growing body of literature focuses on repurposing digital trace data to quantify sudden displacement, observe patterns of population movement, and nowcast forced migration flows (Avramescu and Wiśniowski, 2021; González-Leonardo et al., 2024; Leasure et al., 2023). Digital traces represent an alternative data source for migration studies, particularly in contexts where timely and granular information is critical.

Timely information is also crucial for forced migrants, as forced displacement increases the need for quick and reliable access to information throughout the migration journey. In the context of standard labor migration, the preparation for migration and the gathering of information typically take time (De Haas, 2021), as studies linking online information-seeking behavior to migration flows have shown (Böhme et al., 2020; Wladyka, 2017). In contrast, in the context

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

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

of forced displacement, this preparation happens much more rapidly. Since forced migrants often flee imminent danger, the need for information extends beyond the pre-migration stage and into the peri- and post-migration phases. Recent studies using digital trace data to examine forced migration patterns suggest that the search for migration-related information intensifies after crises (Anastasiadou et al., 2024; Sanliturk and Billari, 2024). Survey-based studies further show that forced migrants frequently use smartphones and the internet to search for information during their journey and after their arrival at their destination (Merisalo and Jauhiainen, 2020). As an example, Wikipedia has been identified as an important information source for Middle Eastern asylum-seekers in Germany (Zimmer and Scheibe, 2020).

Given the importance of online sources of information for forced migrants, we investigate the relationship between online information-seeking behavior and forced migration flows. As a case study, we focus on the Ukrainian refugee crisis that began on February 24, 2022, following Russia's invasion of Ukraine. We examine the use of Wikipedia as an online source of information by Ukrainian refugees. In addition to being the largest and most widely used free online encyclopedia, Wikipedia is also appealing from a data collection perspective. Unlike Google Trends data, which provide only relative search popularity indices, Wikipedia provides publicly accessible data on the absolute number of daily page views dating back to 2015, making it a valuable resource for analyzing users' interests over time. Previous research has shown that Wikipedia readership reflects real-world events and trending topics (Miz et al., 2020), and it has been used to monitor, forecast, and assess information-seeking behavior related to health crises and natural disasters (Jemielniak et al., 2021; Ribeiro et al., 2021; Tizzoni et al., 2020). However, to the best of our knowledge, this is the first study to assess the relationship between information-seeking behavior on Wikipedia and refugee flows. We address the following research questions:
**RQ1:** How did the Ukrainian refugee crisis affect information-seeking behavior on Wikipedia?
**RQ2:** What was the temporal relationship between information-seeking behavior on Wikipedia and Ukrainian refugee flows?

According to the United Nations High Commissioner for Refugees (UNHCR), [32] more than 5.9 million refugees left Ukraine for destinations across Europe up to April 2024. For statistical purposes, the UNHCR uses the term refugees to broadly refer to all individuals who have left Ukraine due to the war. We employ this usage throughout this paper. The majority of Ukrainian refugees initially sought refuge in neighboring countries. In particular, Poland and Germany emerged as primary destinations of Ukrainian refugees. We focus our analysis on Ukrainian refugees in Poland and Germany, examining patterns of information-seeking behavior in Polish and German cities. Poland is particularly relevant, not only as a major destination during the early stages of the Ukrainian refugee crisis (Duszczyk and Kaczmarczyk, 2022), but also because official data on border crossings are available from Polish authorities. Building on these data, we analyze the temporal relationship between Wikipedia-based information-seeking behavior and refugee flows into Poland.

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

Our methodology leverages data from Wikipedia Pageviews to assess how the number of views of Wikipedia articles about cities, used as proxies for potential destinations, varied over time across different language editions, which serve as proxies for countries or regions of origin, in response to migration events. We collected the daily number of views in Ukrainian (the official language of Ukraine), Russian, and English of Wikipedia articles about European capitals. In addition, to account for views by domestic populations in Poland and Germany, we included daily views in the Polish- and German-language editions of Wikipedia. Finally, to assess the null effect on information-seeking behavior related to cities less affected by refugee flows, we collected daily views in Ukrainian, Russian, and English of Wikipedia articles about five of the most populous capitals worldwide. To account for fluctuations in the overall popularity of Wikipedia across different language editions, we computed the proportion of views for each article relative to the total number of views of Wikipedia in that language during the same time period.

We found a positive correlation between the proportion of views of Ukrainian-language Wikipedia articles about capitals of countries in the European Union (EU) and the stocks of Ukrainian refugees in those countries by year since 2022. Wikipedia articles about the capitals of Poland and Germany, for instance, were consistently among the five most viewed articles in the Ukrainian-language Wikipedia in the years analyzed (2022-2024). Within Poland, we also observed a strong correlation between views of Ukrainian-language Wikipedia articles about the most populous Polish cities and the numbers of Ukrainian refugees registered in those locations. The same pattern was observed for the German context, where we noticed a stronger correlation in 2022 between the numbers of Ukrainian refugees under temporary protection in the most populous German cities and views of Wikipedia articles about those locations in Ukrainian than in other Wikipedia languages.

We further explored the temporal dynamics of information-seeking as measured by views of Wikipedia articles about Polish cities in relation to the timing of Ukrainian refugee border crossings into Poland, according to data from the UNHCR. We found a consistent positive association between article views and the daily number of refugees crossing the border. Additionally, we applied Granger causality analysis to examine the lag structure and direction of the relationship between refugee flows and Wikipedia readership. The results of the analysis showed that border crossings by Ukrainian refugees into Poland Granger-caused increases in views of Ukrainian-language Wikipedia articles about Polish cities, with an average optimal lag of around eight days. This indicates that spikes in information-seeking behavior typically followed refugee arrivals by just over a week, rather than preceded them, underscoring the reactive nature of information-seeking behavior during forced migration.

Our results open up new avenues for understanding the relationship between information-seeking behavior and forced migration flows, highlighting the role of Wikipedia as an online source of information during crises. This study makes several key contributions. First, we demonstrate how real-world crises leading to refugee flows influence online information-seeking

behavior, as reflected in Wikipedia views. Second, we introduce Wikipedia data as a novel and timely source for analyzing information-seeking in the context of forced migration. By leveraging the increase in views of Wikipedia articles during crises, we show that forced migration follows a distinct temporal pattern. During forced migration, individuals often leave the origin location first and seek information later, whereas during traditional forms of migration, such as labor migration, individuals tend to engage in more preparatory information searches. This insight sheds light on how people seek information in times of crisis and uncertainty, including during forced migration, and how the behavior of such people differs from the pre-departure planning typical of regular labor migrants. Importantly, the migration process did not end at the moment of crossing the border. While official applications for protection often lagged border crossings by at least a couple of weeks, Wikipedia activity increased almost immediately after border crossings. This finding positions Wikipedia as a near real-time indicator of emerging migration patterns during crises, and highlights Wikipedia's potential role as an early-warning system in migration monitoring.

## 4.2 Related work

### 4.2.1 Refugees and online sources of information

Information and communications technology, as well as online sources of information and social media platforms, have become increasingly important tools for refugees seeking information on which to base their migration decisions (Dekker et al., 2018; Felton, 2015; Merisalo and Jauhiainen, 2020). Studies about Ukrainian refugees report that 92% of Ukrainian refugees in Poland have a mobile phone and 86% have reliable internet access (Social Progress Imperative, 2022). Furthermore, the International Organization for Migration (IOM) reports that Ukrainian refugees in Germany (International Organization for Migration, 2022) consider social media and the internet as their top sources of information. Online sources of information help migrants and refugees with their decision-making processes (Dekker et al., 2018; Felton, 2015; Merisalo and Jauhiainen, 2020; Zimmer and Scheibe, 2020). Additionally, the availability of digital trace data has paved the way for a growing literature that predicts and analyzes forced migration using innovative data sources (Anastasiadou et al., 2024; González-Leonardo et al., 2024; Leasure et al., 2023; Sanliturk and Billari, 2024).

Online search engines provide a useful tool to measure interest in migration-related topics and predict migration patterns (Avramescu and Wiśniowski, 2021; Böhme et al., 2020; Lin et al., 2019). However, there are limitations associated with the use of online search engine data. For example, there are a variety of search engines available, but it is often difficult to compare search data across different platforms. While Google is the most popular search engine worldwide, Bing and Yandex also compete in various regions. Additionally, each search engine reports online search interest in its own way, using different parameters and algorithms. These algorithms may

introduce biases. For instance, Google Trends provides only a normalized index for a given place and time and applies an unobservable threshold that prevents results from being produced when interest is sufficiently low.

Complementing online search engines, Wikipedia Pageviews data regarding the daily absolute number of views of Wikipedia articles are easily accessible through the API or via download from the website. Although Wikipedia use is often associated with deeper topical reading (Kämpf et al., 2015), previous work has shown a high correlation between frequently searched keywords and views of Wikipedia articles. This suggests that Wikipedia Pageviews can be a valuable source for determining popular global web search trends (Yoshida et al., 2015). Wikipedia has also been identified as a key source of information for asylum-seekers from the Middle East in Germany (Zimmer and Scheibe, 2020).

### 4.2.2 Wikipedia readership during crises

Wikipedia is the most popular free online encyclopedia, enabling readers to engage with a variety of information content. Wikipedia readership is known to be influenced by real-world developments (Miz et al., 2020). Numerous studies have used Wikipedia data to monitor trends in health information during outbreaks such as those of COVID-19 and the Zika virus (Ribeiro et al., 2021; Tizzoni et al., 2020).

In addition, Wikipedia has been used to monitor, forecast, and assess information-seeking behavior with regard to diseases and natural disasters (Jemielniak et al., 2021; McIver and Brownstein, 2014). For instance, McIver and Brownstein (2014) showed that Wikipedia-derived models have been effective in estimating influenza cases, outperforming Google Flu Trends in certain contexts. Jemielniak et al. (2021) found that Wikipedia readership patterns are often correlated with external events, such as tropical cyclones, further supporting the utility of Wikipedia data in tracking responses to crises. These findings suggest that Wikipedia data can serve as a valuable complementary data source, especially when traditional surveillance systems are not available in real time in crisis contexts.

In this study, we investigate the relationship between Wikipedia article views and forced migration flows resulting from the Russian invasion of Ukraine. Our aim is to assess how the Ukrainian refugee crisis affected information-seeking behavior on Wikipedia and to analyze the temporal relationship between information-seeking behavior on Wikipedia and Ukrainian refugee flows. To the best of our knowledge, this is the first study to use Wikipedia Pageviews data to assess information-seeking behavior specifically within the context of forced migration.

## 4.3 Data

We examine the relationship between changes in readership of Wikipedia articles about European cities and the situations of Ukrainian refugees across Europe. Our analysis integrates

data collected from Wikipedia alongside supplementary data from various sources, enabling a comprehensive understanding of Ukrainian refugee flows across the continent, with a particular focus on Poland and Germany.

We compiled a list of European capitals to assess the changes in views of Wikipedia articles about European capitals. Then, we narrowed our focus to the Polish and German contexts. We specifically focused on the 19 most populous cities in Poland, [33] which are also among the main destinations of Ukrainian refugees in Poland. For the German context, we applied our methodology to the 40 most populous cities in Germany. [34] [35] To contextualize the magnitude and temporal dynamics of Wikipedia readership changes in these Polish cities, we also collected comparative data for five of the most populous capital cities globally that have been less directly impacted by Ukrainian refugee movements: Beijing, Tokyo, Kinshasa, Jakarta, and Lima. All code and data necessary to reproduce our results are available in a public web repository. [36]

### 4.3.1 Official statistics

#### European data

We collected data from Eurostat in order to compare these data with Wikipedia data. This allowed us to assess the association between the number of Ukrainian refugees across 31 European countries[37] and the proportion of views of Ukrainian-language Wikipedia articles about European capitals after the Russian invasion of Ukraine. We used the Eurostat data on the number of Ukrainian beneficiaries of temporary protection at the end of each month by country. [38] Some countries, such as Germany, started reporting the monthly stocks after August 2022. In addition, to ensure consistency across the other datasets used in our analysis, we selected the stocks as of December of each year (2022–2024) to construct annual stock measures.

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

[34] Cities with at least 200,000 inhabitants.

[35] Berlin, Hamburg, Munich, Cologne, Frankfurt, Stuttgart, Düsseldorf, Leipzig, Dortmund, Essen, Bremen, Dresden, Hanover, Nuremberg, Duisburg, Bochum, Wuppertal, Bielefeld, Bonn, Münster, Mannheim, Karlsruhe, Augsburg, Wiesbaden, Mönchengladbach, Gelsenkirchen, Aachen, Braunschweig, Chemnitz, Kiel, Halle (Saale), Magdeburg, Freiburg im Breisgau, Krefeld, Mainz, Lübeck, Erfurt, Oberhausen, Rostock, Kassel

[36] ***https://github.com/carolcoimbra/wikimig***

[37] Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, and Sweden as members of the European Union, along with Iceland, Liechtenstein, Norway, and Switzerland.

[38] ***https://ec.europa.eu/eurostat/databrowser/view/migr_asytpsm__custom_17904294/default/table***

### Polish data

To study the Polish context, we collected data provided by the Polish government on the daily number of Ukrainian refugees crossing the border from Ukraine to Poland from February 24, 2022, to March 7, 2023, from the UNHCR refugee data platform. [39] After their arrival in Poland, Ukrainian refugees were offered transportation from the border to one of the 27 reception centers, where they received assistance related to transfers, accommodation, meals, and medical care. [40]

Additionally, we gathered data from Poland's Data Portal (DANE) [41] on the stocks of Ukrainian refugees who registered for temporary protection in Polish cities between April 2022 and August 2024. Ukrainian refugees registered for temporary protection in Poland, and were assigned a PESEL number (*Powszechny Elektroniczny System Ewidencji Ludności*, in English: Universal Electronic System for Registration of the Population), which serves as an official identification number in Poland. While this dataset consists of a record of active applications of Ukrainians seeking protection, we used the latest release of each year (2022–2024) [42] to construct the annual stock measures of Ukrainian refugees in Polish cities.

### German data

For the German context, we collected data from the German Federal Statistical Office (GENESIS) [43] on the stocks of Ukrainians seeking protection by administrative district [44] in Germany by reference date (December 31 of each year). [45] In terms of government support and coordination, as in Poland, border authorities in Germany directed refugees to the nearest reception center, where they were provided with accommodation, food, and other essential support services. At the reception center, Ukrainian refugees were given a place to sleep, food, and other support services until longer-term housing was found. [46]

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

[40] https://euaa.europa.eu/sites/default/files/2022-06/Booklet_Poland_EN.pdf

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

[42] The last report for 2022 is from 12/26/2022, the last report for 2023 is from 12/12/2023, and the last report for 2024 is from 12/10/2024.

[43] https://www-genesis.destatis.de/

[44] For the 40 most populous cities in Germany, we collected the number of Ukrainians under temporary protection at the independent city level (*kreisfreie Stadt*). For Hanover and Aachen, however, the data are available at the city-regional level (*Städteregion*), rather than at the independent city level.

[45] https://www-genesis.destatis.de/datenbank/online/statistic/12531/table/12531-0041

[46] https://euaa.europa.eu/sites/default/files/2022-06/Booklet_Germany_EN.pdf

### 4.3.2 Wikipedia Pageviews

We gathered data on Wikipedia article views using the library `WikiToolkit` [47] implemented in Python. The data contained daily counts of views (i.e., each time a page is loaded) of Wikipedia articles across different languages since July 2015. We collected data on user views, which included views by editors, anonymous editors, and readers. Views by search engine "web crawlers" or automated programs were not included. Additionally, the data used in this study included both direct pageviews [48] and those arriving via redirects. [49] By accounting for redirects, we captured views redirected through searches on Wikipedia using alternative spellings, abbreviations, misspellings, or variations in capitalization or spacing (Hill and Shaw, 2014), thus ensuring more comprehensive coverage of page views. Finally, for normalization purposes, we collected the total number of views across different languages from Wikimedia Statistics. [50]

Our analyses examined changes in the readership of Wikipedia articles about European capitals, Polish cities, and German cities across multiple languages. English-language Wikipedia served as a proxy for international attention and as a baseline language, given that English is the most accessed language on Wikipedia. Ukrainian and Russian are the most accessed languages in Ukraine. Readership of articles in Ukrainian, as the official language of Ukraine, was used as a strong proxy for Ukrainians accessing Wikipedia. For articles about Polish and German cities, views in Polish and German, respectively, were included as proxies for domestic attention.

#### Proportion of views

To account for changes in the popularity of Wikipedia across different languages, such as the increased use of Ukrainian over Russian during the war (Kulyk, 2024), we calculated the proportion of daily views for each specific Wikipedia article. Considering the number of views $WV$ on a page $p$ of a Wikipedia article in language $l$ at a time $t$ (e.g., day, week, month, or year), the proportion of views $PWV$ is given by:

$$PWV_{p,l,t} = \frac{WV_{p,l,t}}{WV_{l,t}} \tag{4.1}$$

In the following sections, we present the methodology used to address each of our research questions, along with the key results that support our findings.

[47] https://github.com/pgilders/WikiToolkit/tree/main

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

[49] https://pageviews.wmcloud.org/pageviews/faq/#redirects

[50] https://stats.wikimedia.org/

## 4.4 RQ1: How did the Ukrainian refugee crisis affect information-seeking behavior on Wikipedia?

To address the first research question, we adopted a two-step approach. First, we examined the association between the proportion of views of Ukrainian-language Wikipedia articles and the stocks of Ukrainian refugees under temporary protection in EU countries and in the most populous cities in Poland and Germany. Second, we narrowed our focus to the Polish and German contexts. We assessed the percentage change in views of Ukrainian-language Wikipedia articles about Polish and German cities after the invasion. For comparison, we also analyzed the percentage change in views of Wikipedia articles about five of the most populous capital cities in the world that were less affected by the refugee flows, as well as of articles in language editions other than Ukrainian.

### Association between Wikipedia views and stocks of Ukrainian refugees

First, we examined the association between the readership of Ukrainian-language Wikipedia articles and the numbers of Ukrainian refugees who applied for temporary protection across Europe and Polish and German cities. We compared the yearly proportion of views for each Wikipedia article in five languages: English, Ukrainian, Russian, Polish, and German.

Most datasets on applications for temporary protection are reported on a yearly basis. To ensure consistency when comparing these data with Wikipedia views of Ukrainian refugee stocks across locations, we aggregated the Wikipedia page view data into yearly views. This aggregation was performed separately for each Wikipedia article and language, as defined in Equation 4.1. The resulting time series represents the proportion of yearly views of Wikipedia articles across different languages.

We hypothesized that the most viewed Ukrainian-language articles would show a stronger correlation with refugee presence, serving as a proxy for the information-seeking behavior of Ukrainian refugees. Building on this hypothesis, we further extended our analysis to explore the relationship between views of Wikipedia articles about EU capitals and the broader distribution of Ukrainian refugees across EU countries.

For the European context analysis, we created a ranking of EU countries sorted by the highest number of temporary protection applications from Ukrainian refugees, based on Eurostat data for 2022, 2023, and 2024. Similarly, we ranked EU capitals according to the proportion of views of their Wikipedia articles for the same years. We then calculated the Spearman's rank correlation between the ranking of European countries by number of Ukrainian refugees and the ranking of Wikipedia articles about EU capitals by number of views.

Table 4.1 summarizes the Spearman's rank correlation and shows that the proportion of views of Ukrainian-language Wikipedia articles about EU capitals was positively correlated with the number of Ukrainian refugees in EU countries for all three years. The correlation was lower

| | Europe | | | Poland | | | Germany | | |
|---|---|---|---|---|---|---|---|---|---|
| Wikipedia language | 2022 | 2023 | 2024 | 2022 | 2023 | 2024 | 2022 | 2023 | 2024 |
| Ukrainian | 0.69*** | 0.63*** | 0.63*** | 0.87*** | 0.87*** | 0.85*** | 0.77*** | 0.71*** | 0.71*** |
| Russian | 0.60*** | 0.53** | 0.51** | 0.82*** | 0.81*** | 0.76*** | 0.72*** | 0.68** | 0.72*** |
| English | 0.61*** | 0.56** | 0.59*** | 0.87*** | 0.84*** | 0.83*** | 0.73*** | 0.66** | 0.70*** |
| Polish | 0.57*** | 0.53** | 0.51** | 0.88*** | 0.88*** | 0.88*** | 0.73*** | 0.68** | 0.71*** |
| German | 0.45* | 0.39* | 0.43* | 0.85*** | 0.82*** | 0.83*** | 0.76*** | 0.75*** | 0.75*** |

**Table 4.1:** Spearman's rank correlation between the yearly proportion of views of Wikipedia articles about EU capitals, Polish cities, and German cities and the stocks of Ukrainian refugees with temporarily recognized protection status by year.
$***p < 0.001$; $**p < 0.01$; $*p < 0.05$

when comparing the number of Ukrainian refugees in EU countries with the ranking of views of Wikipedia articles about EU capitals in Russian, English, Polish, and German. For a visual representation of these rankings, see Figure B.1 in the Appendix.

Poland and Germany emerged as key destination countries during the Ukrainian refugee crisis. On Ukrainian-language Wikipedia, the article on Poland's capital city of Warsaw consistently ranked among the top articles in terms of the proportion of yearly views, while the article on Berlin appeared regularly within the top five articles. Building on these patterns, we narrowed our analysis to focus specifically on the Polish and German contexts.

For the analysis on Poland, we investigated the relationship between the number of Ukrainian refugees assigned a PESEL number in Polish cities and the proportion of yearly views of Wikipedia articles about those cities. We created a ranking of Polish cities based on the yearly stocks of Ukrainian refugees assigned PESEL numbers, using data from Poland's Data Portal (DANE) from 2022 to 2024. Similarly, we ranked Polish cities according to the proportion of yearly views of their Wikipedia articles. Following the approach used in the European context analysis, we calculated the Spearman's rank correlation between the ranking of Polish cities by the number of Ukrainian refugees assigned PESEL numbers in those cities by year and the ranking of Polish cities by the proportion of yearly views of Wikipedia articles about those cities.

Table 4.1 also includes a summary of the Spearman's rank correlations for the Polish context. We observed that views of Ukrainian-language Wikipedia articles about Polish cities were strongly aligned with the distribution of Ukrainian refugees across those cities, as measured by PESEL registrations. The correlation between Ukrainian-language views and PESEL numbers declined slightly after 2023, which may reflect a shift in refugees' information needs over time as initial settlement challenges gave way to longer-term integration processes. By contrast, Polish-language Wikipedia views reflected more stable, long-term interest from the host population, and therefore produced rankings that remained consistent across years. Finally, the results for the Polish context also showed that Ukrainian-language correlations remained stronger than those for English-, Russian-, or German-language Wikipedia, underscoring the centrality of the Ukrainian language in the refugees' information-seeking behavior during displacement.

In the Appendix, Figure B.2 shows a visual representation of these rankings. Despite the strong correlation between the proportion of the number of views of Ukrainian-language Wikipedia articles about Polish cities and the number of Ukrainian refugees registered in those cities, Rzeszów stands out. Although Rzeszów had fewer registered refugees than cities like Warsaw and Wrocław, the Wikipedia article about Rzeszów, and especially the Ukrainian-language article, had one of the highest proportions of views over the years 2022-2024. Rzeszów has emerged as a pivotal location in the Ukrainian crisis due to its proximity, of roughly 100 km, to the Ukrainian border. In the month and a half following Russia's invasion of Ukraine in February 2022, Rzeszów hosted an estimated 100,000 refugees and served as a transit point for approximately 1.5 million more (Pietka and Sielska, 2025). Wikipedia captured the information-seeking behavior focused on this city, which played a key role during the migration process.

For the analysis in Germany, we compared the rankings of German cities based on the yearly stocks of Ukrainian refugees under temporary protection from 2022 to 2024 with the ranking of German cities according to the proportion of yearly views of their Wikipedia articles across languages. Table 4.1 summarizes the Spearman's rank correlations for the German context. Results show that views of Ukrainian-language Wikipedia articles about German cities were positively correlated with the official stocks of Ukrainian refugees in those cities, particularly in 2022, indicating that during the initial phase of displacement, refugees turned to Wikipedia in their native language to access information about German destinations as well. Similar to the pattern observed in the Polish context, the correlation between Ukrainian-language views and refugee stocks decreased after 2022, which may reflect changing information needs as the crisis evolved. By contrast, views of German-language Wikipedia articles remained more stable over time, reflecting host population interest, while views of English- and Russian-language articles showed weaker alignment with refugee distributions. In the Appendix, Figure B.3 illustrates these rankings.

Next, we calculated the percentage change in Wikipedia article views about Polish and German cities to estimate the impact of the Russian invasion of Ukraine on readership across different languages.

### Relative change in Wikipedia views

We quantified the percentage increase in views of Wikipedia articles about Polish cities across different language editions, using language as a proxy for the geographic origin of attention. In particular, we focused on Ukrainian-language Wikipedia to approximate the information-seeking behavior of Ukrainian users. To measure changes in the volume of views of articles after the Russian invasion of Ukraine, we computed the relative change in Wikipedia views across language editions.

To reduce noise and provide a more stable representation of trends in Wikipedia readership over time, we aggregated the daily view data into weekly time series. This aggregation was performed separately for each Wikipedia article and language, as defined in Equation 4.1. The

resulting time series represents the proportion of weekly views of Wikipedia articles across different languages. Then, we calculated the relative change in the proportion of weekly views of Wikipedia articles across language editions. This metric captured temporal fluctuations in attention, while accounting for seasonal patterns, by comparing each week to the corresponding week in the previous year.

The relative change ($RC_{p,l,t}$) in the proportion of weekly views for a given page $p$ of a Wikipedia article in language $l$ at a time $t$, is defined as follows:

$$RC_{p,l,t} = \frac{(PWV_{p,l,t} - PWV_{p,l,t-52})}{PWV_{p,l,t-52}} \times 100 \tag{4.2}$$

where $PWV_{p,l,t}$ denotes the proportion of weekly views for a page $p$ in language $l$ during the week $t$, and $PWV_{p,l,t-52}$ is the proportion for the corresponding week one year earlier (i.e., 52 weeks prior). This year-over-year comparison helps isolate the impact of the refugee crisis by controlling for regular seasonal fluctuations in Wikipedia readership.

The inset in Figure 4.1 illustrates the relative change in the proportion of weekly views, compared to the same week in the previous year, of the Wikipedia article about Katowice across four languages (English, Polish, Russian, and Ukrainian) from August 24, 2020, to August 24, 2023. While no noticeable increase in views of the English and Polish articles was observed after the Russian invasion of Ukraine, there was a marked relative change in views of both the Ukrainian and Russian articles. Most strikingly, a relative increase of over 500% in views of the Ukrainian-language article for Katowice occurred immediately after the onset of the war on February 24, 2022. Moreover, during this period, the relative increase in views of Ukrainian-language Wikipedia articles for all other analyzed Polish cities exceeded 200%. These findings underscore the potential impact of the Ukrainian refugee crisis on information-seeking behavior, as reflected in the increase in views of Ukrainian-language Wikipedia articles. For a complete overview of the time series showing relative changes for each one of the Polish cities, see Figure B.4 in the Appendix.

We contrasted the increases in views not only across languages, highlighting that the most significant increases were for views of Ukrainian-language articles, but also across cities that were affected differently by refugee flows in the early stages of the war. Specifically, we compared views of Wikipedia articles about Polish cities that received a large influx of Ukrainian refugees with views of articles about cities less directly impacted by Ukrainian refugee flows. Figure 4.1 illustrates this contrast by showing the maximum relative change and the date of this peak in the month after the Russian invasion. The figure includes Wikipedia articles about the 19 most populous Polish cities and five of the most populous global capitals (Beijing, Jakarta, Kinshasa, Lima, and Tokyo) across the English, Polish, Russian, and Ukrainian editions. For clarity, the article names are shown only for the global capitals articles in Ukrainian (in black) and for the Polish city articles for which the relative change in views exceeded 200% (in gray).

For the Wikipedia articles about five of the world's most populous capitals, we observed either marginal increases or overall decreases in views, suggesting that these cities were less

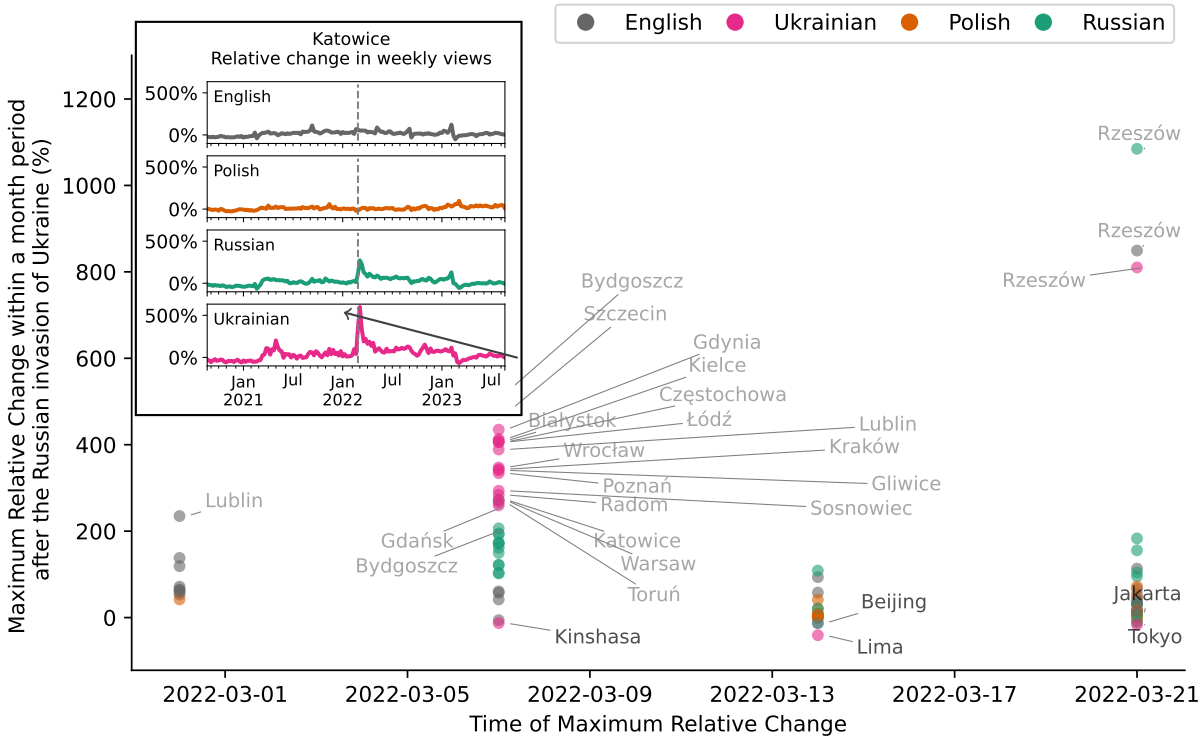



**Figure 4.1:** Maximum relative change in the proportion of weekly views over the month following the Russian invasion of Ukraine, compared to the same period in the previous year. Results are shown for Wikipedia articles about the 19 most populous Polish cities and five of the most populous cities in the world (Beijing, Jakarta, Kinshasa, Lima, and Tokyo) across four languages (English, Polish, Russian, and Ukrainian). As an example, we also show the relative change in the proportion of weekly views compared to the previous year of the Wikipedia article about **Katowice** across four languages (English, Polish, Russian, and Ukrainian) from August 24, 2020, to August 24, 2023.

affected by refugee-related information-seeking during this period. In contrast, we found relative increases of at least 200% in views of all the Ukrainian-language articles about the most populous Polish cities in the week following the Russian invasion of Ukraine. This result points to a strong increase in information-seeking among Ukrainians about Polish cities immediately after the Russian invasion of Ukraine. The peak increase in views of Ukrainian-language Wikipedia articles about Polish cities was consistently much higher and more concentrated in the week following the invasion than the increase in views of Wikipedia articles about other cities or in other languages in the month after the invasion.

We also observed clear spikes in views of Ukrainian-language Wikipedia articles about German cities shortly after the Russian invasion. The largest increases in views of the articles about cities such as Düsseldorf and Munich occurred within two weeks of the outbreak of the war, while views of the articles about Oberhausen, Münster, and Bielefeld peaked slightly later, at around three weeks after the invasion. These temporal dynamics suggest that different German

destinations became the focus of the information-seeking behavior of refugees at varying stages of the initial displacement wave. Figure B.5 in the Appendix illustrates these peaks by showing the maximum relative changes in views and the timing of these surges in the month immediately following the invasion.

To complement the findings regarding the timing of the changes in Wikipedia views, we employed an autoregressive model (AR(1)) and structural break tests. The aim was to identify break points in the time series of the proportion of daily views of Ukrainian-language Wikipedia articles about Polish cities. We used the `strucchange` package in R, which estimates the optimal number of breaks and their confidence intervals (Bai, 1997; Bai and Perron, 2003; Zeileis et al., 2003). The structural break analysis identified changes in the time series of views of Ukrainian-language Wikipedia articles about Polish cities around the time of the Russian invasion of Ukraine. Out of the 19 cities, 15 exhibited break points within one week of the invasion, indicating abrupt increases in information-seeking activity (Table B.1). One city, Gliwice, showed a break point one week prior to the invasion (February 16, 2022), while Rzeszów showed a break point within three weeks after the invasion. Łódź and Lublin were the outliers, as for these cities a structural break in the time series was observed about three months after February 24, 2022. Similarly, the analysis of the time series of the proportion of daily views of Ukrainian-language Wikipedia articles about German cities showed that a structural break occurred within one week after the start of the invasion for 28 out of the 40 cities (Table B.3). We observed a structural break within one month after the start of the invasion for all 40 cities except Krefeld, for which the break date was identified as one day before February 24, and Aachen, for which we did not observe a structural break until May 2022. All observed breaks were followed by sharp rises in daily views of the respective Ukrainian-language articles. In contrast, we did not observe such a pattern for the five most populous world capitals (Table B.2). Out of the five most populous capitals, only Beijing showed a structural break in the time series in proximity to the start of the Russian invasion; however, this was followed by a decrease in the proportion of daily views of the Ukrainian-language Wikipedia article on Beijing.

As hypothesized, the increase in views of Ukrainian-language Wikipedia articles about European cities showed a strong association with the presence of refugees in those locations, particularly in Poland. Building on the consistently large increase in views of articles about Polish cities after the Russian invasion of Ukraine, we then investigated our second research question regarding the temporal relationship between Wikipedia article views and refugee flows.

## 4.5 RQ2: What was the temporal relationship between information-seeking behavior on Wikipedia and Ukrainian refugee flows?

To answer the second research question, we investigated the relationship between the number of Ukrainian refugees in Poland and the increase in the number of views of Ukrainian-language Wikipedia articles about Polish cities. For the following analysis, we used the proportion of daily views of Wikipedia articles about Polish cities, along with official statistics provided by UNHCR on the number of refugees crossing the border from Ukraine into Poland from February 24, 2022 to March 7, 2023.

Figure 4.2 presents the time series of UNHCR data on the daily number of Ukrainian refugees crossing into Poland, alongside data on Wikipedia views. Figure B.6 in the Appendix shows similar time series for each of the Polish cities analyzed. For the Wikipedia data, we reported the proportion of daily views of articles about Polish cities across four different languages: English, Polish, Russian, and Ukrainian. We observed that the proportion of daily views in Russian, and especially in Ukrainian, increased dramatically in 2022, following the start of the war.

We began this analysis by calculating the correlation between the time series. Figure 4.3 shows that the numbers of views of Ukrainian- and Russian-language Wikipedia articles about the 19 most populous cities in Poland were consistently positively correlated with the numbers of Ukrainian refugees who crossed the border into Poland. In the figure, the whiskers represent the 5th and 95th percentiles, while the box spans from the first quartile (25%) to the third quartile (75%) of the data, with a horizontal line indicating the median.

Next, to estimate the temporal relationship between information-seeking behavior as reflected in views of Wikipedia articles about Polish cities and Ukrainian refugee flows to Poland, we conducted Granger causality tests (Granger and Newbold, 2014). Our aim was to assess whether Ukrainian refugee flows to Poland could predict subsequent increases in views of Wikipedia articles about Polish cities (or vice versa). Granger causality is a statistical test that evaluates temporal precedence, indicating whether past values of one time series improve the prediction of another. It is important to note that Granger causality assesses whether including past values of one variable enhances the predictive accuracy of another.

In this context, we tested whether the time series of the proportion of daily views of Wikipedia articles about Polish cities could help forecast the time series of the number of Ukrainian refugees crossing the border into Poland. The test involved regressing the dependent variable (e.g., daily number of Ukrainian refugees crossing the border) on both its own lagged values and the lagged values of the independent variable (e.g., daily views of Wikipedia articles about Polish cities). The null hypothesis stated that the independent variable would not improve the forecast of the dependent variable. If the associated p-value was below a chosen significance level (e.g., 0.05), we would reject the null hypothesis and infer evidence of Granger causality.

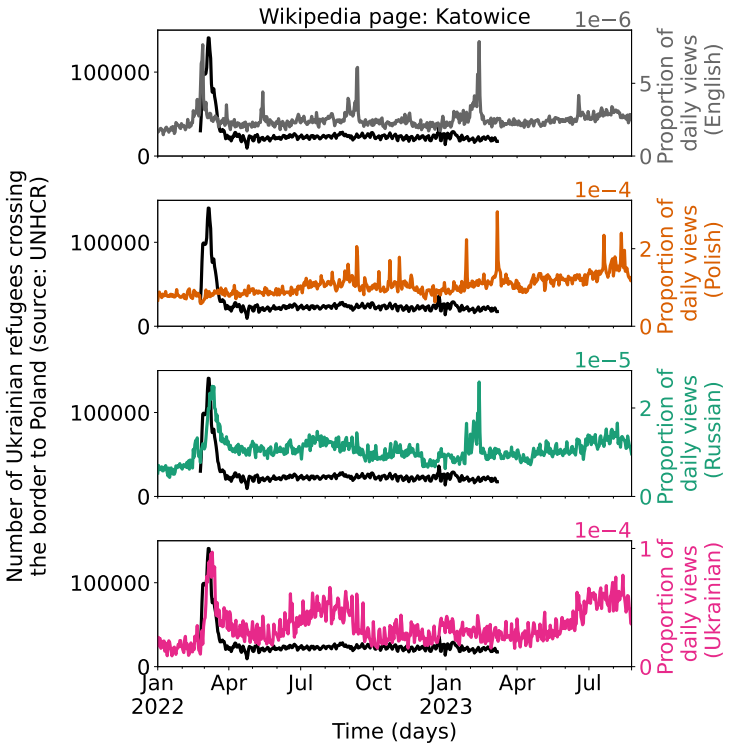



**Figure 4.2:** Time series representing, in black, the daily number of Ukrainian refugees crossing the border from Ukraine to Poland (from February 24, 2022 to March 7, 2023) and, in colors, the proportion of the daily number of views of Wikipedia articles about **Katowice** across four languages (English, Polish, Russian, and Ukrainian).

Thus, if the Granger causality test revealed a temporal relationship between daily views of Wikipedia articles about Polish cities and refugee flows, this would suggest that information-seeking on Wikipedia could serve as an indicator of migration patterns or decisions to cross into Poland.

Before conducting the Granger causality tests, we employed a Vector Autoregressive (VAR) modeling framework (Lütkepohl, 2013), following an approach similar to that in Bailard et al. (2024), using the implementation available in the Python library `statsmodels`.[51] This framework models each variable as a function of its own lags and the lags of other variables in the analysis. We used this approach to identify statistically significant lag lengths for inclusion in the Granger causality tests. The Akaike information criterion (AIC) (Bozdogan, 1987) was employed to determine the optimal lag structure, which was found to be eight days for most Wikipedia articles about Polish cities. The only exceptions were for the articles about Radom and Wrocław, for which the optimal lag was, respectively, 15 and 17 days when modeling border crossings alongside the proportion of daily views in Ukrainian. Next, we verified that all time series satisfied the stationarity requirement using the Augmented Dickey–Fuller (ADF) test (Dickey

[51] ***https://www.statsmodels.org/stable/index.html***

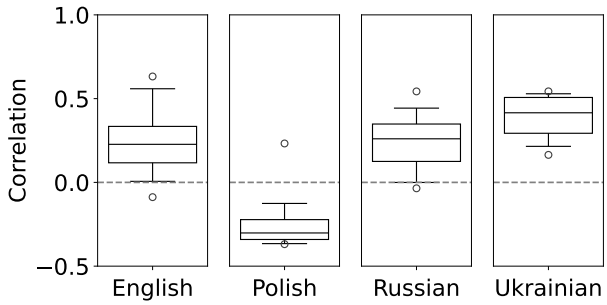



**Figure 4.3:** Correlation between the numbers of views of Wikipedia articles about the 19 most populous cities in Poland, across different languages, and the numbers of Ukrainian refugees crossing the border into Poland. The whiskers of each box plot extend from the 5th to the 95th percentile, and the horizontal line in the middle indicates the median.

and Fuller, 1979), which confirmed stationarity at the identified lag length. We also tested for autocorrelation in the residuals with Lagrange Multiplier (LM) tests, confirming the absence of residual autocorrelation (Johansen, 1995). Additionally, the stability of the VAR models was assessed and confirmed using standard eigenvalue stability diagnostics.

After validating these assumptions and identifying the optimal lag for each pair of the proportion of daily Wikipedia article views and the border crossing data, we applied the Granger causality test. Figure 4.4 presents the F-statistics from the Granger causality tests conducted for each relationship analyzed. Each box plot summarizes the distribution of F-statistics for one type of relationship across the Wikipedia articles about Polish cities. The whiskers extend from the 5th to the 95th percentile, and the box spans from the first quartile to the third quartile, with a vertical middle line indicating the median. Each dot represents the result for a specific Wikipedia article about a Polish city. Dots in blue and red indicate, respectively, statistically significant (p-value $<$ 0.05) and non-significant relationships (p-value $\geq$ 0.05).

Overall, we observed that the F-statistics were higher and significant (p-value $<$ 0.05) in the direction where the number of Ukrainian refugees crossing the border to Poland Granger-caused the daily number of views of Wikipedia articles about Polish cities in Ukrainian and Russian. As Ukrainian and Russian are official or widely spoken languages in Ukraine, this finding is suggestive of information-seeking by Ukrainian refugees. This temporal ordering implies that spikes in information-seeking behavior tended to follow, rather than precede, refugee arrivals. For a detailed summary of the relationships tested, including the optimal lag, F-statistics, and p-values, see Table B.4 in the Appendix. This table includes only relationships with statistically significant p-values ($p < 0.05$).

Our findings align with the understanding that forced migration is often sudden and driven by urgent safety concerns, leaving little opportunity for extensive preparatory research before displacement. Unlike traditional migration processes, such as labor migration, in which individuals typically engage in preparatory information searches (Böhme et al., 2020; Wladyka,

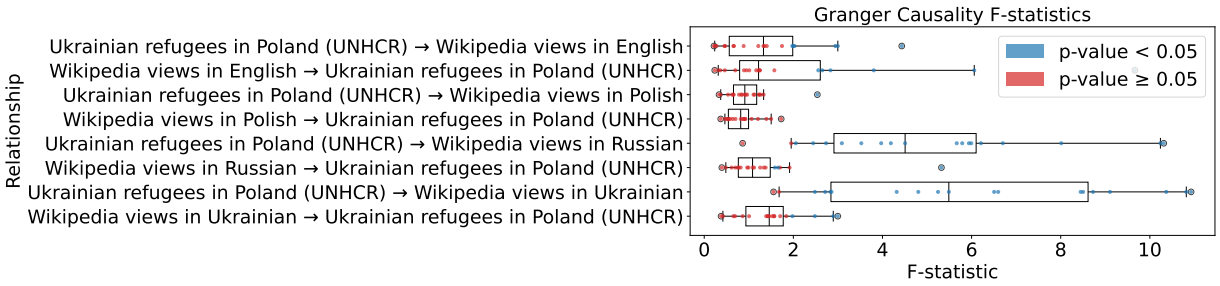



**Figure 4.4:** Distribution of F-statistics from Granger causality tests between time series of Wikipedia views of articles about Polish cities and Ukrainian refugees crossing the border to Poland. The whiskers of each box plot extend from the 5th to the 95th percentile, and the vertical line indicates the median. Each colored dot represents the F-statistic for a Wikipedia article about one of the 19 most populous cities in Poland, with blue indicating statistically significant relationships ($p < 0.05$) and red indicating non-significant relationships ($p \geq 0.05$).

2017) well in advance to inform their destination choices and logistical planning, forced migrants often make immediate decisions to flee and only subsequently seek out detailed information about their new surroundings, leading to an intensification of migration-related information searches after crises (Anastasiadou et al., 2024; Sanliturk and Billari, 2024).

## 4.6 Discussion

In this study, we provided empirical evidence that the Ukrainian refugee crisis significantly increased information-seeking behavior on Wikipedia. By using data from Wikipedia Pageviews, we identified large increases in views of Ukrainian-language Wikipedia articles about European cities after the Russian invasion of Ukraine. There were sharp increases in views of Wikipedia articles about cities that hosted large numbers of Ukrainian refugees, such as Warsaw and other major Polish urban centers, and particularly of articles in Ukrainian. This suggests a strong alignment between forced displacement and digital information-seeking behavior. Moreover, by comparing the Wikipedia views across different languages, we showed that while English- and Polish-language views remained relatively stable, Ukrainian-language views increased immediately after the onset of the war, highlighting Wikipedia's role as a key information resource for displaced populations.

Our findings indicating that Wikipedia readership, especially in the Ukrainian language, increased not before but shortly after the Russian invasion of Ukraine, shed light on the temporal relationship between information-seeking behavior and refugee flows. This timing suggests that, unlike other types of migrants, such as labor migrants, who typically search for information during the pre-migration phase, forced migrants tend to seek information reactively, both during and after displacement. This pattern reflects the urgent and unplanned nature of forced migration.

By providing evidence of these temporal dynamics between online information-seeking and forced migration, our study also contributes to the computational forced migration literature and paves the way for further research. Moreover, our findings highlight the potential of Wikipedia to serve as a real-time indicator of emerging information needs during humanitarian crises and as a complementary tool for detecting shifts in forced migration flows and settlement patterns as they unfold.

This study contributes to the growing literature on the use of digital trace data in migration research by introducing Wikipedia as a timely and scalable source for observing information-seeking behavior during forced migration. However, several limitations affect the generalizability of our findings. First, it is important to emphasize that the Wikipedia Pageviews data are available only from July 2015 onward. In our case, the Ukrainian refugee crisis is a recent event that fits within the period of data availability, but this would not be the case for some major refugee crises, such as those in Syria or Venezuela.

In addition, uneven internet penetration rates across populations may affect the use of Wikipedia as a potential source of information during a crisis. In the context of the Ukrainian refugee crisis and the flow of Ukrainian refugees to Poland, the internet penetration rate was not a major concern. Ukraine's digital infrastructure is comparatively robust, with 80% of individuals accessing the internet daily by early 2024. Poland and Germany have similarly high rates of digital connectivity, but this is not the case for many countries in the Global South. [52]

The most significant limitation is the use of language as a proxy for the users' origin. Information on the location of Wikipedia readers is not publicly available, and for our study, we used views of articles in the Ukrainian language as a proxy for views by Ukrainian refugees. While this methodology is less problematic in the case of Ukrainian, which is largely concentrated in Ukraine, it becomes a more serious limitation when it is applied to languages spoken widely across multiple countries, such as Arabic and Spanish. As a result, our findings may not extend to other major refugee crises, such as those in Syria or Venezuela, where digital access and language use patterns differ considerably. Consequently, the findings may not be generalizable to all forced displacement contexts, especially those with different digital infrastructures, official languages used in multiple countries, or crisis conditions predating 2015. Nevertheless, for the Ukrainian refugee crisis, we have provided empirical evidence of the impact of the Russian invasion of Ukraine on information-seeking behavior on Wikipedia, and established a temporal association between this behavior and Ukrainian refugee flows.

Finally, a further limitation concerns the extent to which our findings are specific to Wikipedia, or instead reflect broader patterns of online information-seeking mediated by search engine rankings. For instance, if much of Wikipedia's traffic originates from Google, the relationship we documented may depend on Wikipedia's current visibility in search results. Given the growing prominence of AI-driven search tools, the use of search engines or readership of online sources of

[52] *https://worldpopulationreview.com/country-rankings/internet-penetration-by-country*

information, such as Wikipedia, might change. These caveats underscore that our results should be interpreted within the present information-seeking ecosystem, in which Wikipedia continues to hold a central role.

Another limitation comes from the official data obtained from Eurostat and the Polish and German statistical offices about the number of temporary protection applications, which were used as ground-truth data in our analysis. These official data suffer from gaps or delays in reporting, particularly in tracking the number of Ukrainian refugees arriving in various cities in a structured and consistent manner. Additionally, once Ukrainian refugees have obtained a PESEL number in a Polish city, for instance, they are free to move within Poland, and we lack reliable data on their internal migration patterns.

To the best of our knowledge, this is the first study using Wikipedia Pageviews data to assess the relationship between information-seeking behavior and forced migration flows. We presented a timely, cost-effective, reproducible, and scalable methodology that leveraged Wikipedia as an innovative data source for studying migration. This approach has important implications.

First, from a policy perspective, the changes in Wikipedia readership in response to the Ukrainian refugee crisis were rapid, and having access to such timely information could provide significant benefits to decision-makers, especially in countries experiencing political instability. Governments can use such real-time data to respond more effectively to migration surges as they unfold, enabling better resource allocation and planning.

Second, the migration literature has previously shown evidence of a time gap between online information-seeking activity and migration flows, with the information search typically preceding the move by several months (Böhme et al., 2020; Wladyka, 2017). In contrast, our study demonstrated how these temporal dynamics changed in the context of forced migration. We found evidence of a lag of about eight days between the onset of the Russian invasion of Ukraine, which led to Ukrainian refugees crossing the border into Poland, and the subsequent peak in views of Wikipedia articles about Polish cities.

Lastly, from a computational social science perspective, our work repurposed Wikipedia data to study a critical real-world phenomenon, demonstrating how digital trace data could be leveraged to provide insights into information-seeking behavior in the context of forced migration.

Beyond the methodological contributions, our work also has implications for civil society and humanitarian organizations. Looking at Wikipedia activity, as a near real-time complement to official statistics, can help organizations detect emerging migration flows during crises, when timely information is scarce. Because refugees often seek information in their native language shortly after displacement, monitoring Wikipedia page views could serve as a low-cost and scalable tool for anticipating where support services, such as housing, legal aid, or language assistance, are most urgently needed. While digital trace data are not a replacement for official records, integrating them into existing monitoring frameworks could enhance the responsiveness of humanitarian interventions and improve coordination across local and international actors.

Finally, our study has underscored the important role of Wikipedia, in particular the role of non-English language editions such as those in Ukrainian and Russian, as an information resource for refugees. The recognition of Wikipedia as a valuable source of information for refugees can help motivate contributors and communities to strengthen public information goods that directly support vulnerable populations during crises.

We collected only publicly available data from the UNHCR website, Poland's official data portals, and the Wikimedia Pageviews platform, following established ethical guidelines (Rivers and Lewis, 2014). Our study used aggregated data on the number of refugees in European countries and in select Polish cities, as well as aggregated views of Wikipedia articles. In the Wikipedia data, the geographic location of readers is not available for privacy reasons. Finally, we did not attempt to identify individual users or link any personal information to specific users.

In future work, the causal dimension of the relationship demonstrated in our study could be further investigated, and data collection could be expanded to cover more cities affected to varying degrees by Ukrainian refugee flows. Additionally, similar to changes in readership, Wikipedia edits are influenced by real-world events (Ruprechter et al., 2021). In this context, we aim to assess the impact of refugee crises on Wikipedia edits. Finally, the dedication, the size of the editors' community, and the quality of the information provided by Wikipedia editors may act as pull factors for certain destinations in refugees' decision-making processes. We would like to explore to what extent online information sources, such as Wikipedia, compensate for the absence or weakness of migration networks for refugees during their movements, as highlighted in previous research (Dekker et al., 2018; Merisalo and Jauhiainen, 2020).

## 4.7 Conclusion

In this study, we assessed the impact of forced migration on Wikipedia readership and provided evidence of a temporal relationship between information-seeking behavior on Wikipedia and refugee movements. Focusing on the Ukrainian refugee crisis, we showed that the number of views of Ukrainian-language Wikipedia articles about Polish cities increased by over 200%, with these views serving as a proxy for information-seeking by Ukrainian users. In addition, we found a strong correlation between the number of views of Wikipedia articles about European cities and the number of Ukrainian refugees recorded across Europe. These findings demonstrated the potential of Wikipedia readership trends to reflect real-world migration patterns.

Regarding the temporal dynamics, we identified a lag of approximately eight days between the onset of mass migration to Poland and the peak in Wikipedia views, which confirmed that spikes in information-seeking behavior tended to follow, rather than precede, refugee arrivals. These results highlight the reactive nature of information needs during forced migration. Despite this reactive pattern, Wikipedia activity increased almost immediately after border crossings, whereas official protection applications often lagged border crossings by weeks. This contrast underscores Wikipedia's potential to serve as a near real-time indicator of migration during crises,

or an early-warning system for monitoring it. Finally, our approach complements traditional data sources by offering faster and more dynamic insights into migration flows, thus opening up new avenues for research and policy development aimed at improving responses to large-scale displacement.

CHAPTER 5

# Mapping Global Gender Balance in STEM: Evidence from Facebook

## 5.1 Introduction

Rapid technological advances disrupt industries and labor markets, leading to an increasing demand for STEM – Science, Technology, Engineering, and Mathematics – talent (Tovey, 2017; UNESCO, 2021). From an industry perspective, greater participation of women in technology enhances workforce diversity, which is crucial for businesses to remain relevant, foster innovation, and stay competitive (Forbes, 2018; Gompers and Kovvali, 2018).

Over the past decades, several studies and initiatives have addressed gender inequality in domains such as health, education, economy, and politics. A range of factors, from family encouragement to cultural perceptions (e.g., stereotypes), have been identified as potential reasons why women avoid STEM subjects (Ertl et al., 2017). Despite ongoing efforts to reduce the gender gap in STEM, many initiatives appear ineffective, and progress toward parity remains slow (Forum, 2020). A substantial gender gap persists due to the lower proportion of women graduating and pursuing careers in STEM fields (García-Holgado et al., 2020; UNESCO, 2017; World Economic Forum, 2016).

Efforts to narrow the gender gap in STEM must take into account cultural contexts and regional specificities (WEF, 2021), as well as variations across disciplines. For example, aggregated gender gap statistics often mask disparities within STEM fields. In the U.S., women earn more than half of undergraduate degrees in Biology, Chemistry, and Mathematics, but only 20% of degrees in fields such as Computer Science, Engineering, and Physics (Cheryan et al., 2016; Munoz-Boudet and Revenga, 2017).

Most existing studies assessing gender gaps rely on surveys (Garcia-Holgado and Garcia-Penalvo, 2022; Tandrayen-Ragoobur and Gokulsing, 2021). These surveys require significant time and financial resources to administer, and global surveys are particularly costly and often limited to developed countries (Kashyap et al., 2020). Moreover, research tends to focus on gender disparities in education and labor markets, with less emphasis on broader perspectives

such as people's preferences. Conducting such analyses at a global scale requires consistent methodologies across countries, yet reliable statistics remain scarce for many countries in the Global South.

To address these limitations, we propose using social media data to assess gender balance in individuals' interests in STEM and non-STEM majors. In this paper, we present a large-scale analysis of the global STEM gender gap using data from the Facebook Advertising Platform (Facebook Ads). Facebook Ads data has been widely applied in diverse contexts, including assessing population health (Araujo et al., 2017), inferring political views (Guimarães et al., 2021), measuring cultural similarities (Vieira et al., 2022c), predicting migration patterns (Alexander et al., 2019; Palotti et al., 2020), and analyzing gender gaps (Garcia et al., 2018; Mejova et al., 2018). For example, Garcia et al. (2018) demonstrate strong correlations between Facebook data and gender inequality indexes at scale across multiple countries.

In this study, we analyze the STEM gender gap globally by considering a broad set of STEM and non-STEM interests associated with college majors. As part of our methodology, we curated a list of Facebook interests related to both STEM and non-STEM majors. We then collected the estimated number of Facebook users expressing interest in each of these majors across different countries. To evaluate gender gaps, we derived two measures: the Overall Gender Balance (OGB) and the Gender Balance (GB). The OGB represents the proportion of male Facebook users in a given country, while the GB is defined as the median proportion of male Facebook users in a given country interested in each of our selected majors, either overall or separately for STEM and non-STEM. Using this methodology, we examined variations in gender gaps across 142 STEM and non-STEM interests in 198 countries.

Our findings reveal clear contrasts between STEM and non-STEM majors. Women constitute the majority of Facebook users interested in non-STEM majors, while men dominate interests in STEM. Within STEM, however, important differences emerge: Life Sciences and Mathematics are more popular among women, whereas Engineering and Technology are more popular among men. In the non-STEM category, majors such as Economics, Business, History, Government, and Journalism are more popular among men than women.

We compared our STEM gender balance estimates with those reported in the 2021 Global Gender Gap Report (WEF, 2021). The results demonstrate the feasibility of using Facebook data to measure global gender gap, particularly in relation to college majors. Our analysis provides gender balance measures for 48 additional countries not covered in the official report. Among the 152 countries included in both datasets, we find that higher levels of gender parity correlate with a greater proportion of women interested in college majors, particularly non-STEM majors.

In addition, we conduct a case study of Brazil to further examine gender balance in STEM. Brazil ranks third worldwide in number of Facebook users, and women represent 54% of this audience. [53] Brazil also presents an interesting scenario where women account for 49% of the

[53] *https://www.statista.com/statistics/866227/facebook-user-share-brazil-gender/*

country's researcher population (Intelligence, 2017), but remain underrepresented in certain STEM fields such as Computer Science. In this case study, we investigate how gender balance varies across Brazilian states, between STEM and non-STEM majors, and among demographic groups (e.g., education levels and age). We apply our proposed methodology to assess gender balance in STEM fields using Facebook Ads data and characterize the Brazil in terms of gender balance in STEM. Finally, we compare our Facebook-based results with data from Brazilian surveys (INEP, 2018; UNESCO, 2017).

Our results suggest that despite women being the majority of Facebook users in Brazil, STEM interests remain concentrated among men. The disparity is even more pronounced when accounting for demographic attributes such as age and education, where women's interest in STEM declines further. These findings offer insights into how programs and initiatives could better target women at different educational stages and age groups to foster greater participation in STEM.

## 5.2 Related work

According to UNESCO, young women accounted for only 25% of students in engineering, manufacturing, construction, or information and communication technology programs in more than two-thirds of countries in 2020 (UNESCO, 2020). Female participation in STEM is not only low, but the attrition rate during and after college is also high. Several studies have examined what motivates women to pursue STEM careers (Botella et al., 2019; Clewell and Campbell, 2002; Steinke et al., 2009), while others have focused on initiatives to improve women's recruitment and retention in STEM. Examples include training teachers to encourage STEM vocations among young women and implementing gender-inclusive policies (Garcia-Holgado and Garcia-Penalvo, 2022; Moss-Racusin et al., 2021).

The gender gap in STEM has been studied primarily through surveys (Garcia-Holgado and Garcia-Penalvo, 2022; Tandrayen-Ragoobur and Gokulsing, 2021). For instance, Tandrayen-Ragoobur and Gokulsing (2021) surveyed undergraduate students and women in STEM occupations, identifying family environment, teacher–student relationships, sense of community, and personal future expectations as key factors. Garcia-Holgado and Garcia-Penalvo (2022) proposed a survey-based model to enhance women's attraction, access, and retention in STEM within higher education institutions in Latin America.

Research has also examined strategies for supporting women's transition into STEM careers. Findings highlight that a successful transition from schooling to university depends heavily on the support provided (Botella et al., 2019; Christie et al., 2017). In addition, stereotypical portrayals of STEM professionals in the media have been identified as potential contributors to women's underrepresentation in STEM fields (Clewell and Campbell, 2002; Steinke et al., 2009; Steinke and Tavarez, 2018).

Despite the usefulness of surveys, they are resource-intensive in terms of time and cost. As an alternative, some studies have employed digital trace data to study gender gaps. Magno and Weber (2014), for example, showed that online inequality is strongly correlated with offline inequality. Analyzing Twitter and Google+ data, the authors concluded that women in less developed countries often have higher relative online social status (e.g., more followers). Other studies contrasted Facebook gender gaps with indices of inequality in education, health, and economic opportunity across multiple countries (Garcia et al., 2018; Kashyap et al., 2020). These findings suggest that social media can reduce barriers to information access for women, though limitations persist in regions such as South Asia and sub-Saharan Africa, where women face restricted internet access and limited digital skills. Gender inequality was also found to be greater in states with higher educational inequality and lower social development, such as India (Kashyap et al., 2020). LinkedIn's Advertising Platform has similarly been used to assess the gender skills gap in the U.S. (Haranko et al., 2018).

More recently, the Facebook Ads API has been used to infer diverse demographic characteristics from audience estimates. Applications include studies of lifestyle diseases (Araujo et al., 2017), human mobility (Spyratos et al., 2019), rural–urban inequalities (Rama et al., 2020), cultural influences on migration (Vieira et al., 2020), media bias (Ribeiro et al., 2018), disparities in Facebook adoption (Gil-Clavel and Zagheni, 2019), and gender inequality (Al Tamime and Weber, 2022; Kashyap et al., 2020). Gender gaps on Facebook have been shown to serve as proxies for broader gender inequalities (Weber et al., 2018). For example, Garcia et al. (2018) found that countries with smaller Facebook gender gaps correlated with higher levels of economic gender equality. Similarly, Kashyap et al. (2020) found strong correlations between gender gaps in internet use, low-level digital skills, and data from both Facebook and Google Ads.

Al Tamime and Weber (2022) explored the potential of Facebook and Instagram Ads data to model gender gap declines in STEM across different age groups in U.S. cities, using APIs filtered by age, gender, and STEM interests. While the study demonstrated the usefulness of social media advertising data, it was limited to a single country and broad STEM categories. To address these limitations and account for well-established gender differences in preferences (Falk and Hermle, 2018), including those observed on Facebook (Cuevas et al., 2021), our study extends this line of work by collecting data on 142 STEM and non-STEM interests associated with college majors across 198 countries, while also providing an in-depth analysis of the Brazilian case.

## 5.3 Data

In this section, we outline the data sources used in our analysis. We first describe how we collected information on interests related to college majors through the Facebook Ads API. We then present two complementary offline datasets used to contrast the gender balance estimates derived from Facebook data with established statistics, thereby assessing the validity and limitations of digital trace data in this context.

### 5.3.1 Facebook Ads data

We use Facebook users' interests in college majors as a proxy for the interest in those fields of study. Based on this, we measure the overall gender balance across STEM and non-STEM majors in 198 countries. In addition to providing a global characterization of gender gaps in college majors, we also focus on the Brazilian context, evaluating regional variations in gender balance across different parts of the country.

The first step of our data collection involved identifying a list of college majors. We used a list published by Handshake, [54] an industry-leading early-career network and career management platform that connects universities, employers, and students. The list provided by the platform consists of 177 college majors grouped into 15 areas of knowledge. Students populate this list when importing data into the platform, while the broader major groups are defined in cooperation with partner universities.

We then annotated each major as STEM or non-STEM. STEM refers to fields in Science, Technology, Engineering, and Mathematics. However, the exact definition of STEM varies across educational, political, and social contexts (Aguilera et al., 2021; Manly et al., 2018). In this study, we follow the classification of the National Center for Education Statistics (NCES), [55] which specifies which majors are categorized as STEM. After annotating the 177 majors, we queried the Facebook Ads API to obtain estimated audience sizes for each.

The Facebook Marketing API [56] allows advertisers to obtain an estimated number of monthly active users for a proposed advertisement that matches given input criteria (Kosinski et al., 2015). For that, the platform provides a list of demographic attributes, such as age, gender, home location, and interests, that the advertiser can customize as the input query. Thus, after specifying the target public and before the ad is launched, advertisers are provided with the audience corresponding to the number of Facebook users that match the target specifications. The users' interests, for example, can be informed by the user or inferred by Facebook based on user activities while posting or interacting with content (e.g., liking, sharing, or updating status). Other attributes, such as age, gender, and location, are explicitly declared by the users in their profiles. For the purposes of this study, we use only publicly available data from the Facebook Marketing API, following established ethical guidelines (Rivers and Lewis, 2014). Our analysis relies on aggregated data, ensuring user anonymity and compliance with Facebook's Marketing API terms of service. [57]

We collected the estimated number of Monthly Active Users (MAU) by gender for all Facebook interests associated with college majors worldwide. Of the 177 majors from Handshake,

[54] https://support.joinhandshake.com/hc/en-us/articles/360019970434-List-of-Major-Groups

[55] https://www.ice.gov/doclib/sevis/pdf/stemList2022.pdf

[56] https://developers.facebook.com/docs/marketing-apis

[57] https://developers.facebook.com/policy/#marketingapi

| | College Majors |
|---|---|
| **STEM** | Aerospace Engineering, Agriculture, Agronomy, Animal Science, Astronomy, Automation Engineering, Automotive Engineering, Aviation, Biochemistry, Biological Engineering, Biology, Biomedical Engineering, Biotechnology, Botany, Cartography, Cell Biology, Chemistry, Computer Engineering, Computer Programming, Computer Science, Computer Systems Networking, Construction Engineering, Construction Management, Cyber Security, Data Science, Earth Sciences, Ecology, Electrical Engineering, Energy Engineering, Environmental Engineering, Environmental Management, Epidemiology, Food Science, Forensics, Forestry, Genetics, Geography, Geology, Immunology, Industrial Engineering, Information Systems Management, Kinesiology, Landscape Architecture, Management Science, Marine Biology, Materials Science, Mathematics, Mathematics Education, Mechanical Engineering, Microbiology, Molecular Biology, Natural Resource Management, Network Engineering, Neuroscience, Nuclear Engineering, Nursery, Oceanography, Physics, Plant Biology, Plant Sciences, Software Design, Soil Science, Statistics, User Experience, Veterinary Sciences, Zoology |
| **Non-STEM** | Accounting, Actuarial, Advertising, Agriculture Business, Agriculture Education, American Sign Language, Anthropology, Applied Arts, Architecture, Art History, Business Administration, Business Analytics, Classical Studies, Consulting, Consumer Science, Counseling, Criminal Justice, Criminology, Culinary Arts, Dentistry, Design, Early Childhood Education, Economics, Education Administration, Elementary Education, Emergency Management, Entrepreneurship, Ethics, Ethnic Studies, Exercise Science, Finance, Financial Management, Foreign Languages, Gender Studies, Government, Graphic Design, History, Homeland Security, Hospital Administration, Human Resources, Human Services, Industrial Design, Interior Design, International Business, International Studies, Journalism, Linguistics, Management, Marketing, Media Studies, Medicine, Music Education, Nursing, Nutrition, Occupational Therapy, Operations Management, Pharmacy, Philosophy, Physical Education, Political Science, Product Design, Psychology, Public Administration, Public Health, Public Policy, Public Relations, Religious Studies, Secondary Education, Social Work, Sociology, Special Education, Speech Pathology, Sport Business, Theatre Arts, Theology, Urban Planning |

**Table 5.1:** College majors grouped into STEM and non-STEM categories.

we matched 193 corresponding Facebook interests. We removed ambiguous or overly broad interests (e.g., Music, Photography) and excluded those with an estimated audience below 1,000, in line with Facebook Ads' privacy threshold. After filtering, we retained 142 interests, consisting of 66 STEM and 76 non-STEM majors, classified according to NCES. Table 5.1 provides the full list of majors included in our analysis.

Our global analysis covers 198 countries, excluding those where Facebook is restricted [58] or where audiences fell below the minimum privacy threshold. Since Facebook applies lower limits to its reporting, some majors with small audiences (below 1,000 users in a given country) could not be included. As a result, the number of majors analyzed varies across countries. In the Appendix, Figure C.2 shows the number of STEM and non-STEM majors available for the top 50 countries with the largest number of interests above the minimum threshold of 1,000 users. The U.S. and India have the broadest coverage and the largest proportions of STEM majors. Figure C.1, also in the Appendix, maps the proportion of STEM majors across countries. In most countries, non-STEM majors outnumber STEM majors, with Turkmenistan and Yemen showing the highest concentration of non-STEM interests.

For the Brazilian case study, we collected audience data disaggregated by demographic attributes such as home location, age, and education level (see Table C.1 in the Appendix). Among the previously identified majors, 73 met the minimum threshold of 1,000 users across all specified demographic subgroups. For the analysis, we consider these 73 majors. Table 5.2 lists

[58] https://www.facebook.com/business/help/1155157871341714?id=176276233019487

| Major Group | College Majors |
|---|---|
| Agriculture/Food/Horticulture (3) | Agriculture, Agronomy, Horticulture |
| Arts/Design (8) | Architecture, Design, Graphic Design, Industrial Design, Interior Design, Landscape Architecture, Music Education, Product Design |
| Business/Entrepreneurship/Human Resources (10) | Accounting, Management, Economics, Entrepreneurship, Finance, Insurance, Marketing, Real Estate, Retail, Sales |
| Civics/Government (7) | Criminology, Law, Political Science, Government, Public Administration, Public Policy, Urban Planning |
| Communications (3) | Advertising, Journalism, Public Relations |
| **Computer Science/Information Systems/Technology** (2) | Computer Programming, Computer Science |
| Education (4) | Early Childhood Education, Physical Education, Higher Education, Secondary Education |
| **Engineering** (3) | Computer Engineering, Electrical Engineering, Mechanical Engineering |
| Health Professions (5) | Dentistry, Kinesiology, Medicine, Nursing, Pharmacy |
| Humanities/Languages (6) | Gender Studies, History, Linguistics, Philosophy, Ethics, Theology |
| **Life Sciences** (6) | Biochemistry, Biology, Botany, Ecology, Microbiology, Zoology |
| **Math/Physical Sciences** (4) | Chemistry, Mathematics, Physics, Statistics |
| **Natural Resources/Sustainability/Environmental Science** (5) | Astronomy, Aviation, Forestry, Geology, Oceanography |
| Social Sciences (7) | Anthropology, Cognition, Neuroscience, Geography, Psychology, Social Work, Sociology |

**Table 5.2:** College major groups used for the Brazilian case study of gender balance. Major groups in bold represent areas of knowledge containing STEM college majors.

all the 73 college majors grouped in 14 groups corresponding to areas of knowledge according to Handshake. Five of the 14 areas of knowledge contain STEM college majors. In total, the Brazilian case study includes 20 STEM majors related to Technology, Engineering, Math/Physical Sciences, Life Sciences, and Environmental Sciences.

### 5.3.2 Offline data

We use data from the Global Gender Gap Report 2021 published by the World Economic Forum (WEF, 2021). The report provides the Global Gender Gap Index (GGGI), a composite measure of gender-based disparities across 156 countries. The index is structured around four key sub-indices: Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment. Each sub-index captures different dimensions of gender inequality, ranging from labor force participation and wage equality to access to education, life expectancy, and representation in political institutions. The GGGI is expressed on a scale from 0 to 1, where a value of 0 denotes maximum inequality and 1 represents full gender parity.

Additionally, we incorporate benchmark statistics from the UNESCO Survey on Women in STEM (UNESCO, 2017). This survey provides country-level measures of women's representation in research and higher education, with particular attention to disparities in STEM fields. For the purposes of our analysis, we focus specifically on the statistics provided for Brazil, which allow us to evaluate the extent to which the results derived from Facebook Ads data align with established survey evidence in this national case study.

## 5.4 Gender balance metric

To assess the global gender distribution among Facebook users, we compute the Overall Gender Balance (OGB), which indicates the proportion of male users in a given population $p$:

$$OGB_p = \frac{MAU_p(male)}{MAU_p(male) + MAU_p(female)} \tag{5.1}$$

To evaluate gender balance in college majors using Facebook users' interests, we adopt the Gender Balance (GB) metric introduced in prior studies (Haranko et al., 2018). This metric quantifies the ratio of male to female Facebook users interested in a specific major. Defining the metric requires specifying the target population, which can be obtained by combining various demographic attributes available in Facebook Ads. For example, the target population may consist of female Facebook users in Brazil interested in Computer Science, or female users aged 20–35 enrolled in graduate school, living in Brazil, and interested in Computer Science. Given a population $p$, we compute the proportion of users with gender $g$ interested in a college major $m$ as follows:

$$A_p(g, m) = \frac{MAU_p(g, m)}{MAU_p(g)} \tag{5.2}$$

Normalization is essential because of the prevalent imbalance in gender distributions on Facebook, where female users outnumber male users in most countries (see Figure 5.1a). To account for this, we use the normalized audience shares to calculate the Gender Balance (GB) of a college major $m$ within a population $p$, as shown in Equation 5.3:

$$GB_p(m) = \frac{A_p(male, m)}{A_p(male, m) + A_p(female, m)} \tag{5.3}$$

The $GB$ scores range from 0 to 1, with 0.5 indicating gender parity. Values above 0.5 indicate a male majority, while values below 0.5 indicate a female majority.

## 5.5 Facebook gender balance across countries

In this section, we present the Overall Gender Balance (OGB) and Gender Balance (GB) derived from Facebook Ads users' interests in college majors across countries. We begin by examining the overall gender balance, considering the gender composition of Facebook users in each country.

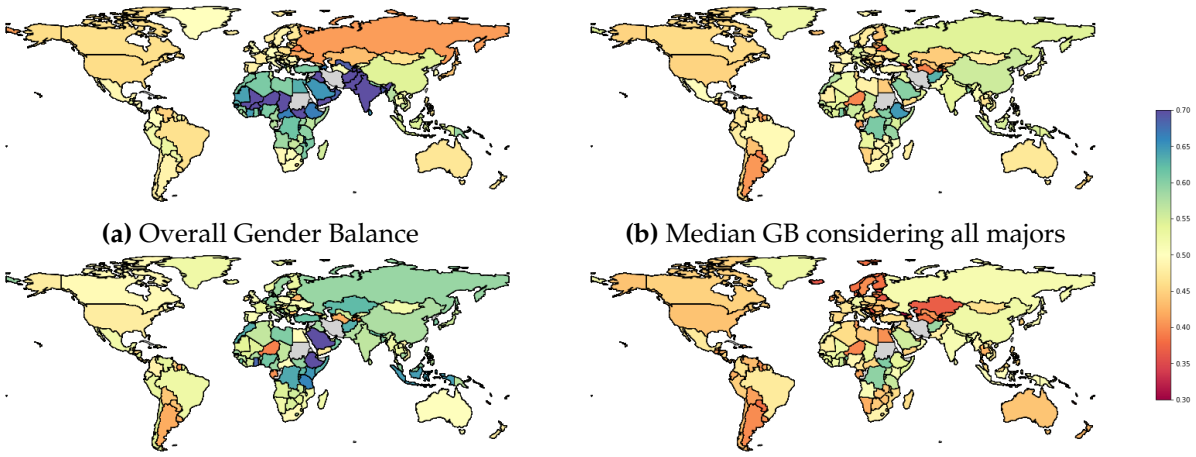
Figure 5.1a displays OGB values across countries on a color scale. Red hues denote a lower proportion of men (low OGB values), while blue hues indicate higher proportions of men (high OGB). Gray countries lack available Facebook data. OGB ranges from 0.39 in Belarus to 0.85 in Yemen (median = 0.51). In most countries, female audiences surpass male audiences, consistent with earlier findings that women are more active on Facebook (Gil-Clavel and Zagheni, 2019).

Figure 5.1b presents the median GB values for all majors across countries. GB values range from 0.37 in Georgia to 0.65 in Ethiopia (median = 0.48). Notably, 64% of countries have GB below 0.5, indicating a higher proportion of women than men interested in college majors, with exceptions concentrated in selected African and Asian countries. OGB and GB show a moderate positive correlation (Pearson's $r = 0.45$, Figure 5.3).

Figures 5.1c and 5.1d show the median GB values for STEM and non-STEM majors, respectively. GB values are generally higher for STEM, suggesting that in most countries men are more likely than women to express interest in STEM majors. In fact, 74% of countries exhibit a male majority in STEM, with GB values ranging from 0.37 in New Caledonia to 0.71 in Saudi Arabia (median = 0.57). Conversely, 72% of countries show a female majority in non-STEM, with GB values ranging from 0.31 in Georgia to 0.6 in South Sudan (median = 0.45). The 75th percentile of GB in non-STEM is 0.49, further emphasizing women's stronger interest in these fields.

In countries where men outnumber women overall (high OGB, Figure 5.1a), men also tend to show greater interest in STEM majors (high GB STEM), particularly in North Africa and Asia. Only 48 countries exhibit a female majority in STEM (GB $< 0.5$). Examples include Niger (OGB = 0.8, GB STEM = 0.39), Tajikistan (OGB = 0.78, GB STEM = 0.41), Togo (OGB = 0.7, GB STEM = 0.45), and Yemen (OGB = 0.7, GB STEM = 0.48). Interestingly, the opposite holds for non-STEM: in countries where men dominate the Facebook user base, women tend to show higher interest in non-STEM majors. This pattern is observed in Tajikistan (OGB = 0.78, GB non-STEM = 0.38), Azerbaijan (OGB = 0.67, GB non-STEM = 0.39), Egypt (OGB = 0.63, GB non-STEM = 0.4), Niger (OGB = 0.8, GB non-STEM = 0.4), Uzbekistan (OGB = 0.69, GB non-STEM = 0.41), and Gambia (OGB = 0.66, GB non-STEM = 0.42). Only 44 countries have median GB values for non-STEM above 0.5 (male majority). Overall, OGB and GB are moderately correlated for both STEM ($r = 0.3$) and non-STEM ($r = 0.45$). Despite global differences in Facebook penetration and the overall female-biased user base (OGB $< 0.5$), a consistent pattern emerges: men are more interested in STEM majors (GB STEM $> 0.5$), while women are more interested in non-STEM majors (GB non-STEM $< 0.5$).

To provide more granular insights, we further examine GB values by major for the five countries with the highest number of majors represented (Figure C.2, in the Appendix). Figure 5.2



**(a)** Overall Gender Balance

**(b)** Median GB considering all majors

**(c)** Median GB of all STEM majors

**(d)** Median GB of all non-STEM majors

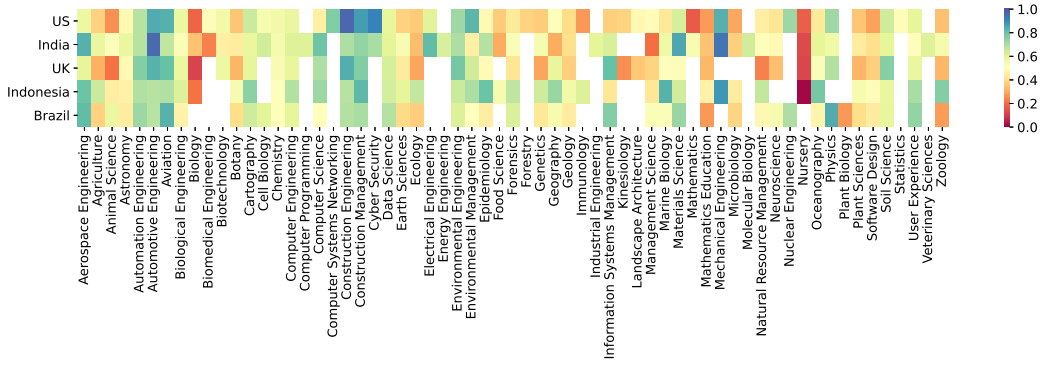
**Figure 5.1:** Overall Gender Balance (OGB) and Gender Balance (GB) across countries. Coloring ranges from red for the highest proportion of women to blue for the highest proportion of men. Gray indicates countries with unavailable information.

shows the GB for each major in these countries, with Figures 5.2a and 5.2b separating STEM and non-STEM majors, respectively. The color scale follows Figure 5.1, where red indicates male-dominance, blue indicates female-dominance, and white marks missing data due to Facebook's 1,000-user threshold. Across the two figures, about 60% of STEM majors exhibit GB $> 0.5$ (male-dominated interest), while about 60% of non-STEM majors exhibit GB $< 0.5$ (female-dominated interest).

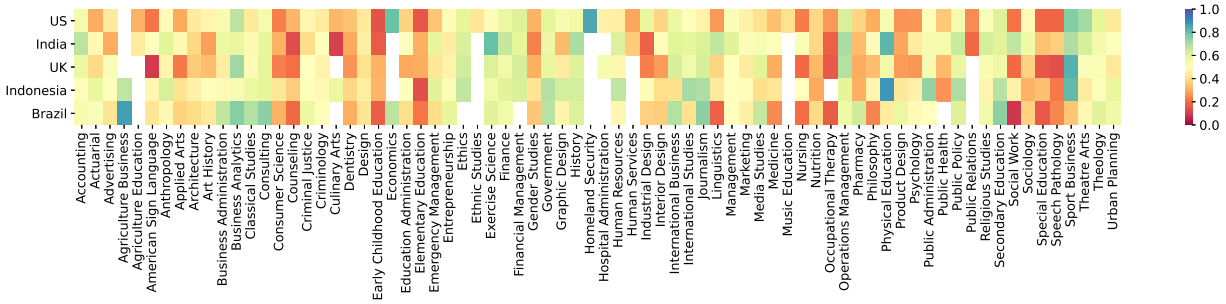
However, not all majors conform to these overall patterns. In STEM, women outnumber men in certain fields, particularly Life Sciences and Mathematics. Two broad patterns emerge: (i) Engineering and Technology majors are predominantly male-dominated, while (ii) Science and Mathematics majors often exhibit female-dominance. This suggests that while gender gaps are widely recognized in STEM, they vary substantially across disciplines. Policies aiming to reduce gender disparities should therefore consider the heterogeneity within STEM. For non-STEM majors, exceptions also exist, with men outnumbering women in fields such as Economics and Business, History, Government, and Journalism.

### 5.5.1 Contrasting online and offline gender gaps

The contrast between online and offline gender gaps can offer valuable insights into the interconnectedness of online and offline measures of gender gap and shed light on the effectiveness of using social media data to measure gender gaps. Offline indicators can provide a benchmark for assessing the extent to which social media data can capture the gender gap, while also serving to highlight any methodological limitations. As an offline indicator of gender gap, we used




**(a)** STEM majors




**(b)** non-STEM majors

**Figure 5.2:** Gender Balance (GB) for each major in the top 5 selected countries. Colors range from red (low GB) to blue (high GB). White indicates unavailable data.

the Global Gender Gap Index (GGGI) data from the 2021 report provided by World Economic Forum (WEF, 2021). The GGGI values range from 0 (maximum disparity) to 1 (full parity).

Figure 5.3 shows Pearson correlations between gender balance measures derived from Facebook data, the GGGI, and its four sub-indices for 152 countries covered in both datasets. We find a strong negative correlation between Facebook OGB and the GGGI ($r = -0.69$). For example, Iceland, Finland, Norway, New Zealand, and Sweden, which score high on the GGGI, are close to gender parity and have a higher proportion of female than male Facebook users (low OGB). Conversely, Yemen and Afghanistan, with low GGGI scores, exhibit male-dominated Facebook audiences. As a counterexample, Bangladesh, the United Arab Emirates, Burundi, Rwanda, Albania, and Mozambique approach gender parity offline (high GGGI), but their Facebook user base remains male-skewed (high OGB).

Considering Facebook Gender Balance, we observe a moderate negative correlation with the GGGI ($r = -0.35$). This suggests that in countries with greater gender inequality offline (low GGGI), men are more likely than women to express interests in college majors on Facebook.

The correlation is stronger between the GGGI and non-STEM GB ($r = -0.40$). This result indicates that higher offline gender parity is associated with greater female interest in non-STEM majors on Facebook. For instance, the Nordic countries (Iceland, Finland, Norway, New Zealand, and Sweden) combine high GGGI scores with low GB values across non-STEM majors, meaning

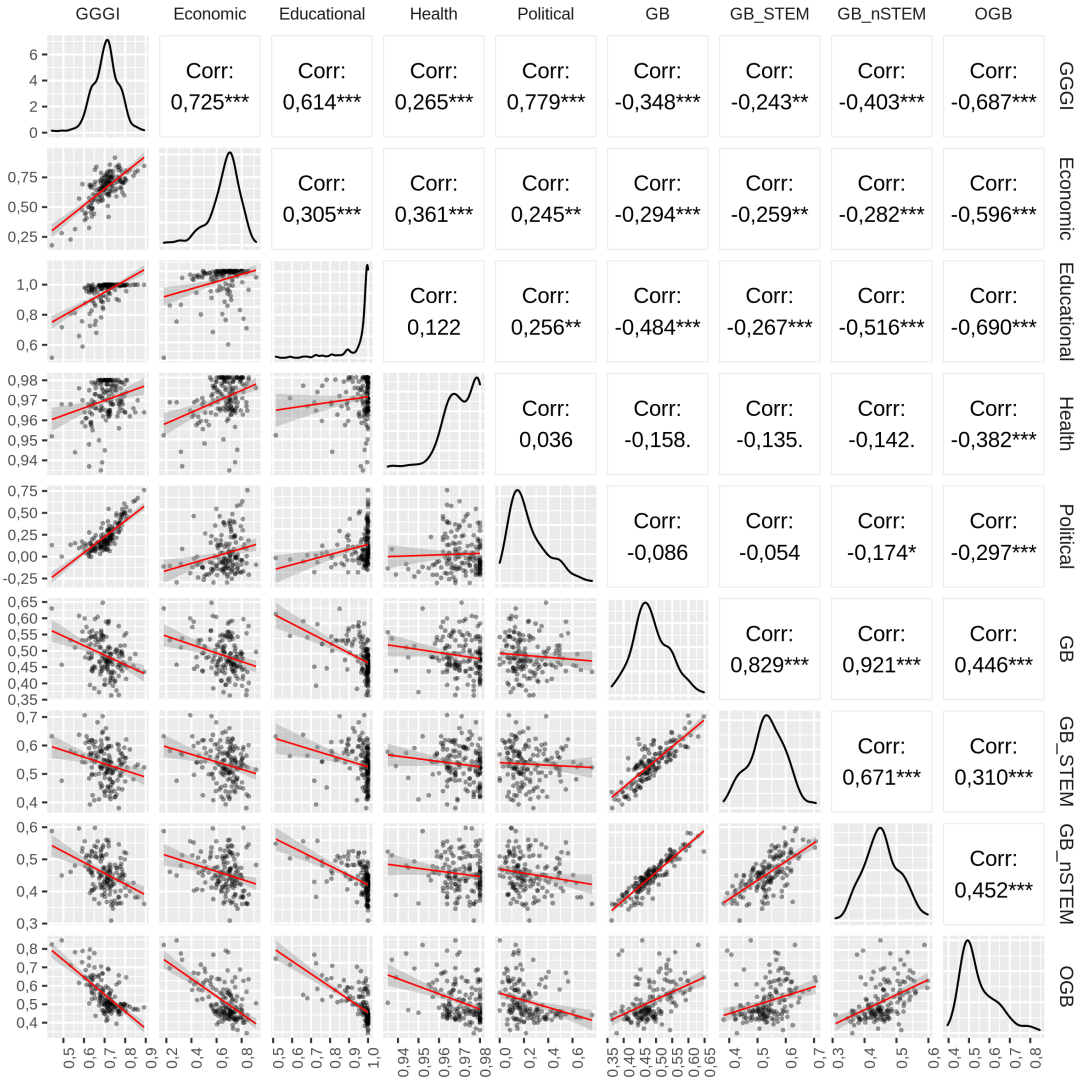



**Figure 5.3:** Correlation matrix of Gender Balance (GB) measures derived from Facebook data and the Global Gender Gap Index (GGGI), including its four components: Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment.
$***p < 0.001$; $**p < 0.01$; $*p < 0.05$

that most Facebook users interested in these fields are women. In contrast, Afghanistan, with a low GGGI, shows a male majority across interests in college majors (high GB). Yemen represents a counterexample where despite a low GGGI, women outnumber men in Facebook interests related to college majors (low GB).

Finally, we observed a low negative correlation between the GGGI and STEM GB values ($r = -0.24$). One explanation is that our measure uses the median GB across all STEM majors, which flattens variation. Engineering and Technology tend to be highly male-dominated (high GB), while Life Sciences and Mathematics often attract more women (low GB) (see Figure 5.2).

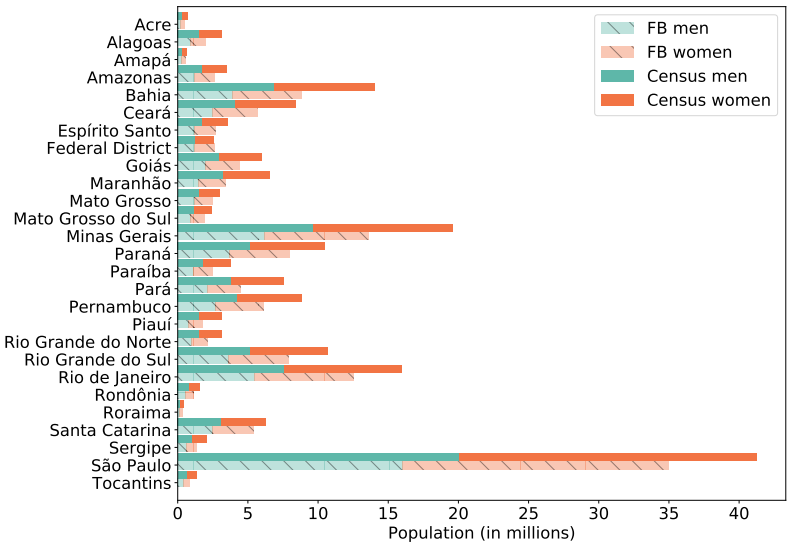



**Figure 5.4:** Population by gender across Brazilian regions. Darker shades represent official census data (IBGE, 2010), while lighter shades indicate Facebook audiences (September 2020).

## 5.6 Facebook gender balance in Brazil

Before analyzing the Gender Balance across Brazilian states, we first examine the overall population and Facebook user distribution by gender. Figure 5.4 presents the official population alongside the number of Facebook users across states, disaggregated by gender. In all Brazilian regions, the female audience on Facebook surpasses the male audience. According to the Brazilian Institute of Geography and Statistics (IBGE), [59] the Brazilian population consists of 49% men and 51% women. In contrast, Facebook Ads data reveal a stronger imbalance, with more than 54% of users being women. This finding aligns with Gil-Clavel and Zagheni (2019), who report that women are generally more likely than men to engage with the Facebook platform.

Focusing on the interests of Facebook users in Brazil across the 73 college majors, Figure C.3a (in the Appendix) presents the number of women and men interested in each major. Figure C.3b (in the Appendix) presents the normalized audience (Equation 5.2). Despite the female-biased population on Facebook, certain majors attract proportionally more male users. For example, Engineering, Technology, and Aviation are predominantly male-oriented, while majors such as Design, Retail, Ethics, and Psychology attract a higher proportion of women.

Figure 5.5 shows the distribution of Gender Balance across the 73 majors for each Brazilian state. The bottom and top of each box correspond to the first and third quartiles, the band inside the box is the median, and dots represent outliers. Across all states, the median Gender Balance is below parity ($GB < 0.5$), indicating that at least half of the majors have more female than

[59] *https://educa.ibge.gov.br/jovens/conheca-o-brasil/populacao/18320-quantidade-de-homens-e-mulheres.html*

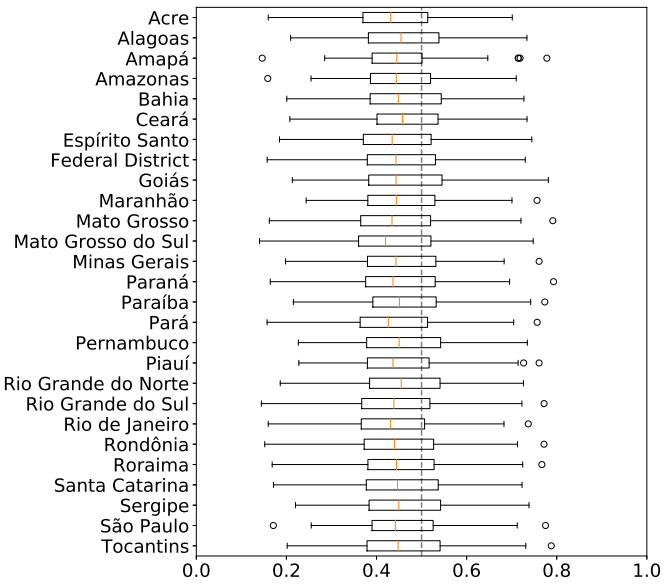



**Figure 5.5:** Distribution of Gender Balance (GB) in Facebook users' interests across college majors for each Brazilian region. A GB value of 0.5 indicates gender parity, while values below (above) 0.5 reflect a female (male) majority.

male Facebook users interested in them. Note that the vertical reference line aligns with the third quartile in most regions, suggesting that approximately 30% of majors are male-dominated ($GB > 0.5$). We also observe more outliers on the male-dominated side ($GB > 0.5$), highlighting majors where men strongly outnumber women. Most of these cases correspond to Technology and Engineering fields, pointing to the lower interest of women in these majors.

We computed the gender distribution across Brazilian states and compared it across all college majors. Figure 5.6 shows the Gender Balance (GB) distribution for each major in each state. The majority of majors with a higher proportion of men ($GB > 0.5$) are STEM fields (shown in gray). Most of these are part of the Environmental Science, Engineering, or Computer Science subgroups (see Table 5.2). In contrast, other STEM fields, such as Life Sciences and Math/Physical Sciences, have lower GB values, indicating a higher proportion of women interested in them.

Even within STEM, two distinct patterns emerge. Majors related to Environmental Science, Engineering, and Computer Science are male-dominated, while Life Sciences and Math/Physical Sciences attract more women. Therefore, STEM majors should not be treated as a homogeneous group and can be subdivided into two subgroups: (i) Life Sciences and Math/Physical Sciences, where women are well represented, and (ii) Environmental Science, Engineering, and Computer Science, where women are underrepresented.

For non-STEM majors, a higher proportion of men ($GB > 0.5$) is observed in fields such as Agriculture/Food/Horticulture, Civics/Government, Geography, Public Relations, Insurance, Product Design, and Music Education. This shows that gender imbalance is not restricted to STEM areas. Nevertheless, most non-STEM majors in Brazil have a higher proportion of women interested in them.

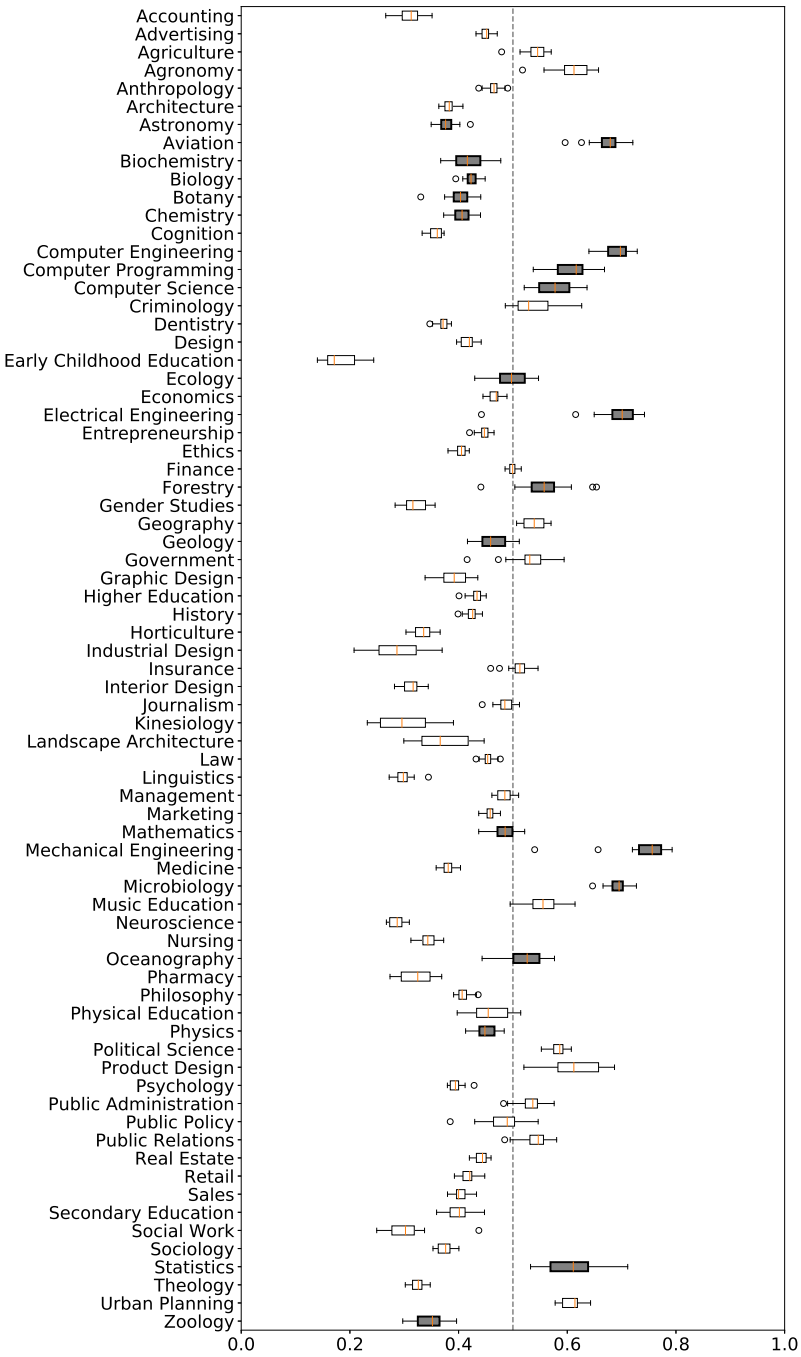
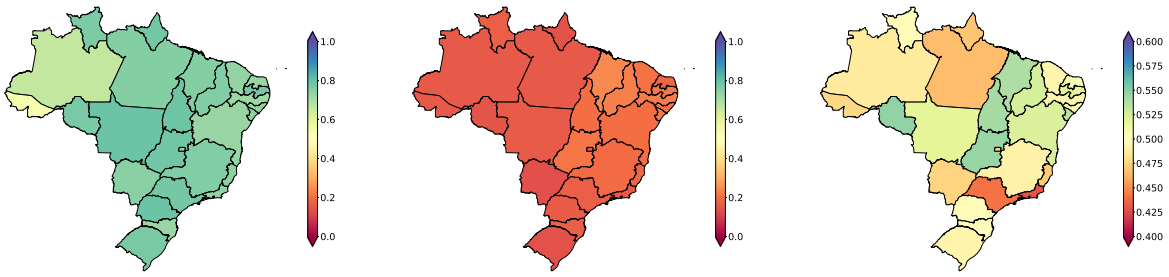



**Figure 5.6:** Gender Balance (GB) distribution of Facebook users' interests in college majors across Brazilian regions. Boxplots for STEM majors are shown in gray, while non-STEM majors are shown in white. A GB value of 0.5 indicates gender parity; values below (above) 0.5 reflect a female (male) majority.

We analyzed the spatial distribution of Gender Balance (GB) values for selected college majors across Brazilian states, as shown in Figure 5.7. Figures 5.7a and 5.7b show GB values for two majors: Mechanical Engineering and Early Childhood Education, respectively. As observed in Figure 5.6, Mechanical Engineering exhibits the highest GB values, indicating a male majority,




**(a)** Mechanical Engineering **(b)** Early Childhood Education **(c)** Ecology

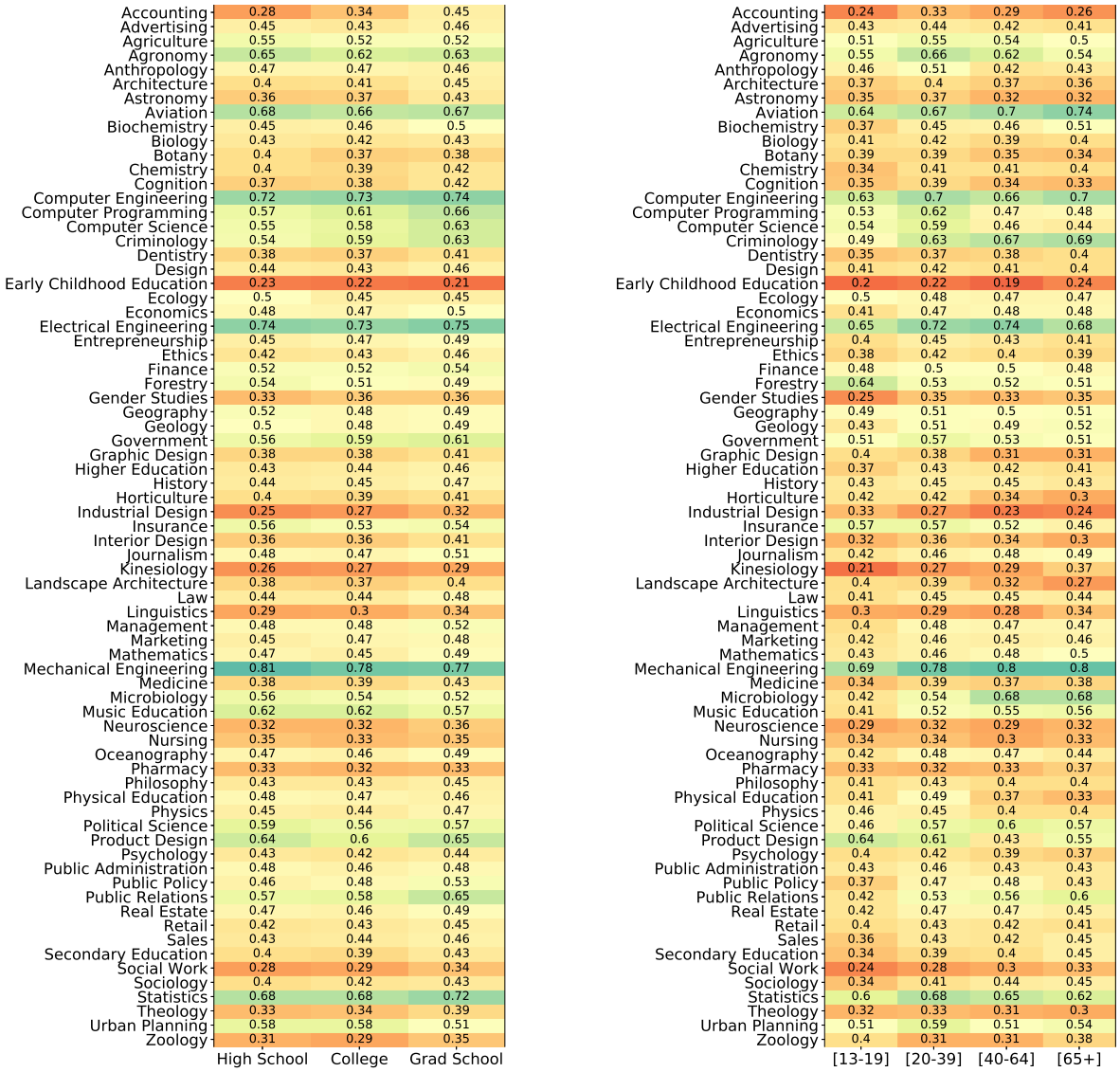
**Figure 5.7:** Gender Balance (GB) of Facebook users' interests in college majors across Brazilian regions. Hot (red) colors indicate lower GB values, while cold (blue) colors indicate higher GB values. A GB of 0.5 represents gender parity while values below (above) 0.5 reflect a female (male) majority.

consistently across all states. This suggests that regional factors do not significantly influence interest in this major. Conversely, Early Childhood Education has the lowest GB values, indicating a female majority, with no notable regional variation in interest.

Most boxplots in Figure 5.6 lie entirely above or below the perfect Gender Balance line ($GB = 0.5$). This indicates that gender parity is absent in most majors, with interest skewed predominantly toward either men or women. However, some majors show variation in GB across states. Ecology, for example, has a balanced national GB (median = 0.5), but state-level values vary considerably. Figure 5.7c illustrates this variation, with GB values ranging from 0.4 to 0.6. In the Southeast, South, and North regions, more women are interested in Ecology, whereas in the Midwest and Northeast, men show greater interest.

To further explore gender balance, we examine Facebook users in Brazil across demographic subgroups, focusing on education levels and age groups. For education, we consider three subpopulations based on Facebook Ads categories: *High School*, *College*, and *Grad School*, which include users either currently enrolled or reporting their highest education level. For age, we examine four groups: Adolescent (13–19 years), Early Adulthood (20–39 years), Adulthood (40–64 years), and Maturity (65 years or older), linking age with college major interests. Additionally, we compare the data provided by Facebook Ads with the results from the UNESCO Survey (UNESCO, 2017).

Figure 5.8a shows the distribution of GB values per college major and education level. The figure uses a color scale where red hues indicate a female majority, while blue hues indicate a male majority. For visual comparison, the GB values for each major are plotted alongside their corresponding education levels. Overall, we observe that as education level increases, GB values also tend to increase, indicating a higher proportion of men relative to women. For example, majors such as Statistics, Computer Engineering, Electrical Engineering, and Mechanical Engineering show increasing male dominance at higher education levels. In many STEM majors, the GB increases notably from College to Graduate School (e.g., Computer Programming, Computer




**(a)** Facebook Education Status

**(b)** Facebook Age Group

**Figure 5.8:** Gender Balance (GB) of Brazilian Facebook users' interests in college majors, broken down by education level and age group. A GB of 0.5 indicates gender parity while values below (above) 0.5 represent a female (male) majority.

Science). Even in fields dominated by women, such as Education (with the exception of Early Childhood Education), the proportion of women decreases as education level rises. This pattern aligns with findings from UNESCO's STEM and Gender Advancement (SAGA) project (Huyer, 2015), which shows that the gender gap in STEM widens significantly in the transition from Bachelor's to post-graduate levels (Master's and Doctorate) and into research careers.

Figure 5.8b shows the GB value for each major by age group. We observe a similar trend to that seen with education level: the proportion of women interested in each major tends to

decrease with age. Over their lifetime, women appear to lose interest in certain STEM majors, such as Aviation and Mechanical Engineering. For most majors, particularly STEM fields, GB values increase during the transition from Adolescent (13–19) to Early Adulthood (20–39), mirroring the pattern observed from High School to College. This suggests that female interest in STEM majors declines during early adulthood, consistent with both educational and career-stage trends.

## 5.7 Conclusion

This study demonstrates the potential of Facebook Ads data as a cost-effective, scalable, and reproducible tool to assess gender balance in STEM fields at both global and national levels. At the global scale, our analysis of 198 countries confirms a bias toward women in the Facebook audience, consistent with prior research, except in specific countries in Africa and Asia. However, while women are generally overrepresented in Facebook's audience and show greater interest in non-STEM majors, men predominate in STEM-related interests in the majority of countries. Noteworthy variations emerge across STEM majors: Life Sciences and Math attract more women, whereas Engineering and Technology appeal to a higher proportion of men. Similarly, certain non-STEM majors, such as Economics and Business, History, Government, and Journalism, are more popular among male users.

The case study of Brazil complements the global picture by revealing regional and demographic nuances that large-scale analyses might overlook. Despite Brazil's relatively high share of women in the population and research workforce, men are still disproportionately represented in several STEM majors, particularly in Environmental Science, Engineering, and Technology. Our analysis provides a general picture of how gender inequality is present in STEM majors as well as their distribution across different regions in Brazil. Our findings also indicate that women's interest in STEM declines with age and higher educational levels, pinpointing critical stages at which gender disparities intensify. Regional comparisons further highlight that certain majors, such as Ecology, exhibit substantial variation across states, whereas others, like Early Childhood Education or Mechanical Engineering, display consistent gendered patterns nationwide.

In general, our results align with those obtained from official statistics. Using Facebook Ads, however, allows us to overcome certain limitations, enabling large-scale data collection with fewer resources and in a shorter time. Our methodology provides valuable insights that complement traditional survey and administrative data and can be extended to other countries or applied at different granularities (e.g., cities within the same state). Moreover, the focus on Brazil contributes to the literature of gender imbalance studies providing statistics about geographic regions that have not been previously studied. Finally, by identifying when and where women disengage from STEM pathways, our methodology can inform policies and initiatives aimed at fostering greater gender equity in science, technology, engineering, and mathematics.

Together, these studies highlight both the promise and the limitations of using digital trace data to study gender inequality in STEM. Despite the demonstrated feasibility of using Facebook

Ads data for demographic studies, several limitations exist. First, some interests are explicitly declared by users, while others are inferred by Facebook based on activities such as posts, likes, and shares. The undocumented nature of these inferences poses challenges, although we assume that users' interests in college majors on Facebook provide a reasonable proxy for studying gender gaps across disciplines. Facebook Ads functions as a black box, and the mechanisms used to infer demographic attributes from the offline world are not publicly disclosed. This limitation prevents a full evaluation of the reliability of the collected data. Facebook usage also varies across cultures and regions, which may influence results. Furthermore, the Facebook population is known to be biased with respect to gender, age, and other socio-demographic characteristics (Araujo et al., 2017; Gil-Clavel and Zagheni, 2019). Nevertheless, previous studies have validated Facebook Ads data against offline sources and found promising results for analyses of digital gender inequality, migration, and cultural tastes (Garcia et al., 2018; Spyratos et al., 2019; Vieira et al., 2020).

Second, we do not differentiate the relative importance of each user's interest in a major. In other words, we assume that all declared interests carry equal weight, even if some users are interested in only one major while others engage with multiple majors. Third, demographic attributes are restricted to those provided by the Facebook Ads Platform, and gender is treated as a binary variable. Fourth, our analysis primarily focuses on country-level gender balance without age-group breakdowns to maximize data coverage, since when audience sizes fall below 1,000—the Facebook Ads API minimum—the true value can range from 0 to 1,000. For the Brazilian case study, we extend the analysis to include age and education level for demographic groups with audiences above 1,000. This approach assumes that a substantial portion of Facebook users accurately report their interests and personal information, such as age and educational level, which can be partially verified against external sources (Araujo et al., 2017; Grow et al., 2022).

Finally, there is no universal agreement on which majors constitute STEM. Classifications depend on stakeholders, including educators, students, and industry. Our methodology is designed to be reproducible and flexible, allowing for multiple STEM definitions, and we believe our results remain robust under reasonable alternative classifications.

Other social media platforms, such as LinkedIn, could also be used to measure gender balance. However, LinkedIn is biased toward individuals in the labor market (Najafikhah and Shamizanjani, 2018), which falls outside the focus of this study. In contrast, Facebook remains the largest social media platform, with the highest number of active users, particularly among young people (Duggan et al., 2013). Leveraging Facebook data allows us to focus on a broader audience, including users who are not necessarily enrolled in or working in STEM fields. Our primary goal is to understand the gender gap at the level of users' interests, regardless of professional or educational status. Future research could validate these findings against offline indicators, extend the methodology to other countries and platforms, and refine analyses by incorporating age, education, and regional dimensions.

CHAPTER 6

# Characterizing the Population of Missing Children in Guatemala: Evidence from Twitter

## 6.1 Introduction

Quantitative researchers often encounter "missing" values in their data, for instance, derived from measurement error or non-response. "Missingness" in real life, however, is an altogether different phenomenon. Missing people are individuals whose status as dead or alive is unknown, reported as missing by relatives and friends. The Wall Street Journal reported in 2012 that: "It is estimated that some 8 million children go missing around the world each year". [60] However, missing people have received surprisingly little attention in the literature. This may be partly due to the fact that governments do not always collect and release up-to-date data on missing people (Citroni, 2014), and disappearances are relatively rare in high-income countries. [61]

Missing people are a particular concern in enclaves of the Global South in which populations are exposed to high levels of economic uncertainty, poverty, violence, and conflict (García and Aburto, 2019; McIlwaine and Moser, 2001; Meneghel and Hirakata, 2011; Soares Filho, 2011). We focus on the case of Guatemala, which emerged from a bloody 36-year civil war in 1996 (CEH, 1999). Over the last decades, the country has experienced growing levels of violence related to gang activity and drug trafficking (Cruz et al., 2020) that, alongside high levels of poverty, have contributed to large-scale population displacement and migration (Jonas and Rodríguez, 2015).

Aggregated records from the National Civilian Police are one of the few data sources on missing people in Guatemala. According to these data, almost 40,000 individuals went missing in the country between 2003 and 2020, half of whom were under 18 years of age. Between 2018 and 2020, the period we focus on in this study, 4,000 individuals went missing in Guatemala, 65%

[60] https://www.wsj.com/articles/SB10001424052702304707604577424451609727644

[61] http://amberalert.eu/statistics/

of whom were children (0-17 years old). The police data, however, provide neither a full nor an updated picture of the population of missing people. The reasons why a person disappears may also be associated with age, gender, race, socioeconomic level, and contextual factors. However, little is still known about the characteristics of missing people.

Previous work has focused on quantifying public engagement with reports of missing people on Twitter (Crump, 2011; Ferguson and Soave, 2021; Solymosi et al., 2021). In this work, we adopt a methodology for collecting Twitter data in a systematic way, including image processing techniques to extract text from images, to track the population of missing children in Guatemala in real-time. We collected individual-level data on more than 700 missing children from the official Twitter account of *Alerta Alba-Keneth*, a governmental warning system responsible for disseminating information about missing children in Guatemala (Rodas Andrade, 2021).

We combine the Twitter data with official data sources to provide the first systematic description of the population of missing children in Guatemala during the 2018–2020 period. This work addresses the following research question: What is the composition of the population of missing children in Guatemala by age, sex, and geography, and how has this varied over time? We expect to provide a demographic overview of the missing children, avoiding the race- and gender-related media bias (e.g., African American missing children and female missing children being significantly underrepresented in television news coverage) as reported by other authors (Min and Feaster, 2010).

This is the first project that leverages digital data and demographic methods to study the population of missing children in Guatemala or, to the best of our knowledge, in any other setting. We focus on Guatemala, but the methodology we propose in this work can be replicated to other Twitter accounts sharing detailed information about missing people. Moreover, this work allows us to better understand the advantages of using social media data combined with official government data as a complementary data source to study missing children.

## 6.2 Related work

Existing studies have documented how individuals affected by the disappearance of an acquaintance use social networks to crowdsource support. Despite the fact that there is no guarantee that reporting missing people online will lead to their discovery, social networks provide much-needed emotional support to friends and relatives of the missing people (Hattingh and Matthee, 2016). Police departments around the world increasingly use social media to engage the surrounding communities using social media platforms (Dai et al., 2017).

Exploratory studies have examined the use of social media tools by police departments, with a particular focus on Twitter (Crump, 2011; Ferguson and Soave, 2021; Solymosi et al., 2021). Most of these studies have centered on understanding how to increase the public's engagement with the information on missing people shared online. Ferguson and Soave (2021) analyzed 373 missing person tweets posted over two years (2017–2019) from 15 Canadian police services on

Twitter to estimate which features are likely to increase public engagement (retweets, likes, and comments) with these tweets. Similarly, Solymosi et al. (2021) analyzed 1,008 tweets made by Greater Manchester Police between the period of 2011 and 2018 in order to investigate what features of the tweet, the Twitter account, and the missing person are associated with levels of retweeting. In both studies, the authors found several features to be significantly associated with higher engagement, such as the use of images and hashtags. These strategies increased community outreach and participation, as well as the likelihood of efficiently and successfully solving the missing person cases. A standardized structure for sharing details of missing people on Twitter may enhance the usefulness of social media in this respect (Ferguson and Soave, 2021).

In this work, we collect information about missing children from the official Twitter account of *Alerta Alba-Keneth*, a governmental warning system focused on missing children in Guatemala. The main goal of this work is to characterize the population of missing children, instead of predicting the best features responsible for high engagement with tweets. Moreover, we do not expect variations in engagement due to different features used to tweet (e.g., the use of different hashtags), since the *Alerta Alba-Keneth* Twitter account is highly structured and all the tweets and images shared have standardized components.

Finally, by focusing on the characterization of the missing children in Guatemala, we expect to provide a demographic overview of the missing child, minimizing the race- and gender-related media bias (e.g., African American missing children and female missing children being significantly underrepresented in television news coverage) as reported by other authors (Min and Feaster, 2010). To the best of our knowledge, this is the first work studying the demographic composition of missing children using Twitter data.

## 6.3 Data

### 6.3.1 Guatemalan National Police data

Missing people in Guatemala can be directly reported to three government offices: the National Civilian Police (Policía Nacional Civil), the Prosecutor General's Office (Ministerio Público), or the Attorney General's Office (Procuraduría General de la Nación). Each of these offices produces its own statistics, which means that reports of disappearances are often duplicated across sources. In our experience, accessing these data is very difficult given the unwillingness of the relevant authorities to share them.

We obtained data on missing people in Guatemala via multiple Freedom of Information requests to the National Civilian Police. We made similar requests to a number of other government agencies that collect reports of missing people, but this information was refused due to

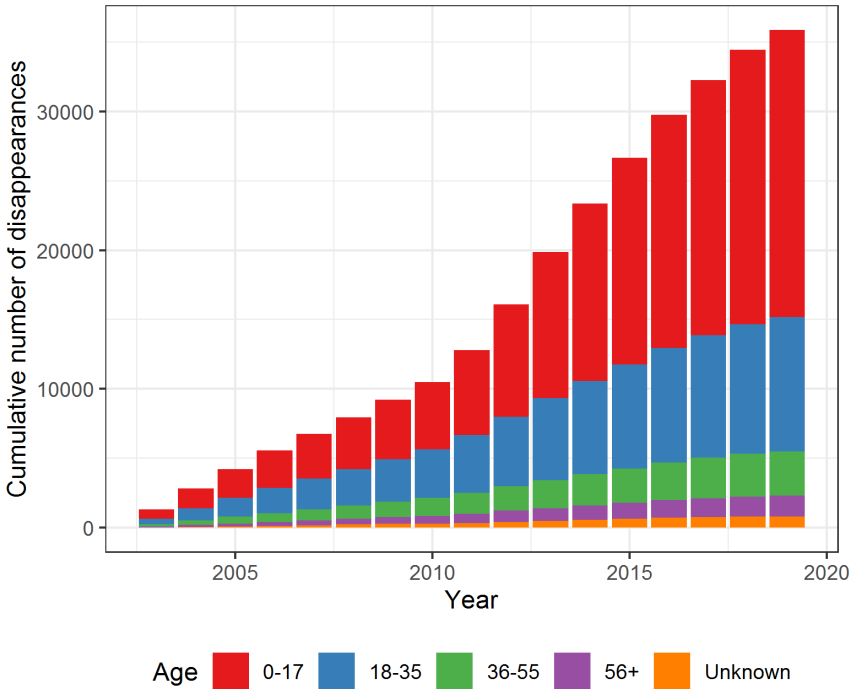



**Figure 6.1:** Age distribution of the cumulative number of disappearances (2003–2019) according to the Guatemalan National Police data.

data protection. The data from the National Civilian Police cover the 2003–2021 period, [62] but our analyses are limited to the 2018–2020 period.

The data include the number of missing people reported to the police by broad age groups (0–17; 18–35; 36–55; 56+) and month of occurrence. Figure 6.1 shows that the population under 18 years of age represents more than half of the known cases of missing people in the country. However, the police department did not share individual-level data or aggregated information on the composition of the population by age, sex, or ethnicity. Other details about the circumstances of the events (e.g., where the disappearance took place) were also unavailable. Finally, it is important to note that the National Police data only include disappearances reported to this institution directly.

### 6.3.2 Twitter data

Twitter is a popular microblogging platform that allows users to express themselves and record their thoughts in up to 280 characters. Twitter is also a social network since it is possible to follow users and stay updated on their tweets (which may contain text, photos, GIFs, videos, and links). Among its more than 200 million daily active users [63] are news outlets, academic institutions, and

[62] No data were provided for May–December 2019.

[63] *https://s22.q4cdn.com/826641620/files/doc_financials/2021/q2/Q2'21_InvestorFactSheet.pdf*

government agencies that use the platform to share official information with a wide audience. We are particularly interested in the latter's use of Twitter to share information on missing people. Twitter data have been used extensively by researchers in the social sciences (McCormick et al., 2017; Tsoi et al., 2018).

We leverage the use of Twitter by a Guatemalan government agency to gather detailed data on missing children, i.e., those under 18 years old according to Guatemalan law. We focus on the *Alerta Alba-Keneth*, [64] an intergovernmental agency tasked with the search, location, and immediate protection of missing or abducted children and adolescents. *Alerta Alba-Keneth* was established in 2010 to coordinate the efforts of multiple government agencies for locating missing children, including the National Civilian Police, the Prosecutor General's Office, and the Attorney General's Office. It collects reports on missing children submitted to any of these agencies, as well as those made directly to *Alerta Alba-Keneth* through its website and hotline, making it the most comprehensive source of data on missing children in the country (Rodas Andrade, 2021).

The *Alerta Alba-Keneth* Twitter account, set up in 2015, has high visibility with almost 20,000 followers. The *@alba_keneth* profile tweets images with information about newly missing children on a daily basis, following the structure of the example shown in Figure 6.2.

We used the Twitter API Academic Research product track [65] to download all the tweets and their embedded images from the *Alerta Alba-Keneth* account since 2015. Regular tweets about missing children in a consistent format began in 2018. Each tweet consists of a short text and an image. The text contains information about the name, age, date, and place of disappearance. The image includes not only a portrait photo of the missing child but also a "profile" with additional details in text form, such as sex, appearance or individual characteristics (e.g., eye color, hair color, and height), and contact information.

We collected a total of 13,696 tweets containing images. Since some images were tweeted more than once, we compared the *hash* [66] of each image and removed all duplicates. This resulted in 11,130 unique images, which we processed to extract information on the children's demographic attributes. We used Optical Character Recognition (OCR) via the Python package PyTesseract [67] to extract demographic details (date and place of disappearance, age, and sex) from the images. This information was not always available in the tweet text, and when it was, it appeared in an unstructured form. Thus, it was more reliable to extract it from the standardized images. At the end of this process, we obtained structured information on 7,800 unique missing children. Of these, 6,875 correspond to tweets from 2018 onward and are therefore included in our study.

[64] Official website available at: ***https://www.albakeneth.gob.gt/***

[65] ***https://developer.twitter.com/en/products/twitter-api/academic-research***

[66] ***https://docs.python.org/3/library/hashlib.html***

[67] ***https://pypi.org/project/pytesseract/***


**Figure 6.2:** Example of a tweet from the *Alerta Alba-Keneth* Twitter account (@alba_keneth).

To extract the coordinates of the place of disappearance, we used the address reported in the "place of disappearance" field as input for two geocoding services: Bing [68] and ArcGIS. [69] We calculated the Haversine distance between the two results to provide an additional measure of the accuracy of the address geolocation.

Table 6.1 summarizes the main differences in data availability between the *Alerta Alba-Keneth* Twitter data and the official records shared by the Guatemalan National Civilian Police. All data used in this study were made publicly available on Twitter by the *Alerta Alba-Keneth* account. Nevertheless, data protection was a major concern when collecting and processing the Twitter data. We recognize that these data are sensitive, as they concern minors in distressing situations

[68] https://docs.microsoft.com/en-us/bingmaps/rest-services/

[69] https://geocode.arcgis.com/arcgis/

| | Guatemalan National Police | Alerta Alba-Keneth Twitter |
|---|---|---|
| Individual-level data | | ✓ |
| Real-time | | ✓ |
| Easily accessible | | ✓ |
| Sex (aggregated) | ✓ | ✓ |
| Age (aggregated) | ✓ | ✓ |
| Ethnicity | | ✓ |
| Place of disappearance | | ✓ |

**Table 6.1:** Overview of data availability in the two data sources used for this study.

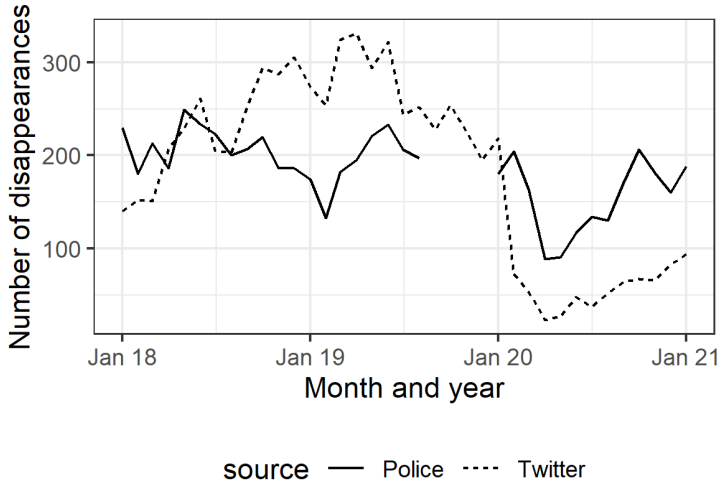



**Figure 6.3:** Monthly counts of missing children: comparison between Guatemalan National Police data and *Alerta Alba-Keneth* Twitter data. Note that National Police data for May–December 2019 were not provided.

that also affect their families. For this reason, all data were stored on an encrypted drive, and although individual-level analyses were conducted, this article reports only aggregated results. Additionally, personally identifying information has been removed from the text to ensure that no child or relative can be uniquely identified.

## 6.4 Results

In this section, we examine temporal and demographic trends in reports of missing children in Guatemala for the 2018–2020 period. The primary data source is Twitter, specifically the *Alerta Alba-Keneth* account. Whenever possible, we compare these findings with official records from the National Police. It is important to note that although both sources track the same phenomenon – reports of missing children – we do not expect perfect correspondence. Each institution has independent, albeit complementary, data collection systems. Nevertheless, we anticipate that both sources will reflect similar temporal trends, even if absolute numbers differ.

As a first step, we compare the number of missing children reported in the Twitter data with the official police data for the study period. Figure 6.3 shows the number of children reported as missing according to both data sources. The image shows a general correspondence in the general trends of child disappearances over time. Although both series show similar fluctuations over time, the number of disappearances is higher in the Twitter data for most of 2019 and lower for 2020. This decline in 2020 does not necessarily indicate a true decrease in disappearances, as it may be linked to the Covid-19 pandemic (Martinez-Folgar et al., 2021). According to a recent report by the Ombudsman of Guatemala (Rodas Andrade, 2021), strict lockdowns, restricted business hours, and limited public transportation likely reduced the number of missing children reports received by *Alerta Alba-Keneth* in 2020.

The measure described above provides a rough proxy for incidence, representing the monthly number of children reported missing. A key advantage of the individual-level Twitter data is the ability to disaggregate the missing children population by age and sex. In contrast, the police data do not provide information on the joint distribution of age and sex. This granularity is crucial for identifying the groups at highest risk of disappearance and is particularly relevant given established links between missing children and human trafficking (Rodas Andrade, 2021).

Figure 6.4 illustrates the distribution of missing minors across three age subgroups: young children (0–4), children (4–12), and adolescents (13–17). The top panels show the absolute number of monthly reports, while the bottom panels show the age composition of missing children each month. The first thing to note is that the age distributions remain relatively constant over time (bottom panels) despite visible fluctuations in the total number of reported cases. This pattern holds even during 2020, when reports were strongly affected by the Covid-19 pandemic (Rodas Andrade, 2021). The figure also shows that adolescents constitute the largest group of missing children in the study period, whereas the proportions of young children and children are both considerably lower and similar to each other.

A striking finding, shown in Figure 6.4, is the sex-based disparity. The bottom panels reveal important differences between the age distributions of missing boys and girls. Specifically, missing girls tend to be older than missing boys. Approximately 75% of missing girls were adolescents, compared to about 60% of missing boys. Female adolescents thus represent the largest subpopulation of missing children and were nearly twice as likely to be reported as missing compared to boys of the same age.

Figure 6.5 summarizes the stock of missing children in the 2018-2020 period by age and sex. The image confirms the patterns discussed above and shows the large disparities in the number of missing male and female children. The number of women aged between 10 and 17 years was 4,200, more than twice as high as the number of men reported as missing in the same age group (1,800). As expected, the ratio between the number of adolescent women and men is largest for older ages. There are 2.7 times more women aged 15-17 reported as missing than men of the same age. This large gap between sexes is absent for younger age groups (e.g., 870 girls younger than 10 years were reported as missing, compared to 850 boys).

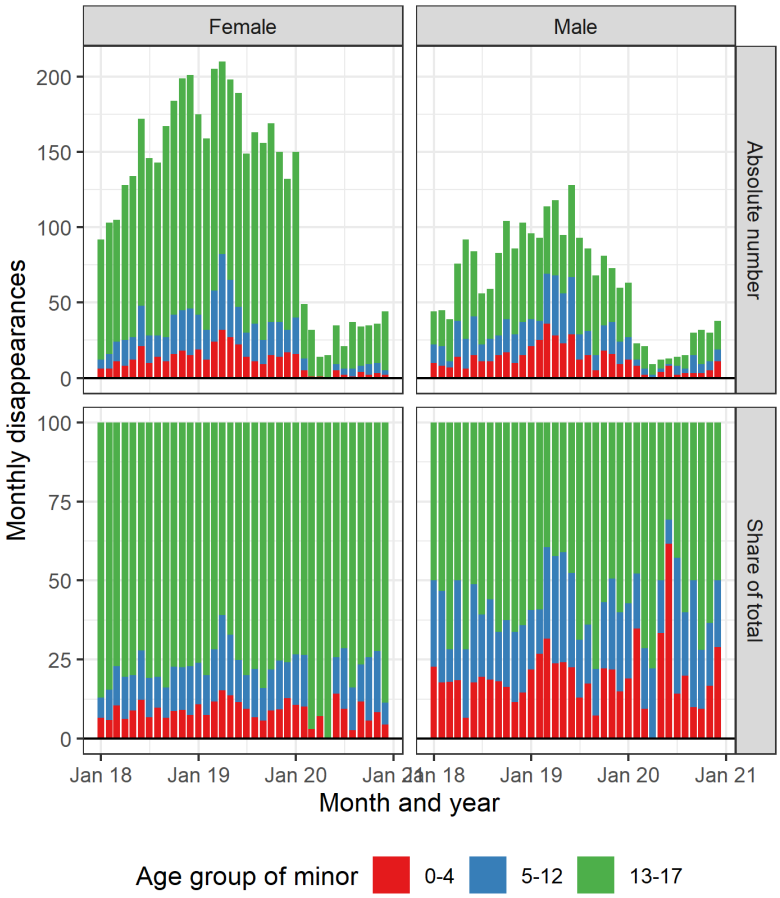



**Figure 6.4:** Age and sex distribution of missing children by month of reported disappearance (2018–2020) according to the *Alerta Alba-Keneth* Twitter data.

Next, we analyze the geographic distribution of disappearance events. This information is available from the Twitter data, as equivalent individual-level location data are not provided by the National Civilian Police. Figure 6.6 displays the locations of missing children reported by the *Alerta Alba-Keneth* Twitter account between 2018 and 2020. Each disappearance is represented as a dot. To avoid overplotting in high-density areas, we applied a small random noise to the locations (less than 0.001°), so the dots may not correspond to the exact locations.

We highlight four important spatial patterns from this figure. First, reports are concentrated in urban centers, particularly in the capital city, which houses about 20% of the country's population. Second, a smaller concentration of reports occurs in the Highlands, located in the central and western regions. This area has a large indigenous population, roughly half of the country's total, and also exhibits the highest poverty rates according to the National Institute of Statistics. [70] Third, higher concentrations of missing children are found in the south, which is mostly populated by non-indigenous groups and generally has higher levels of violence. Fourth,

[70] https://www.ine.gob.gt/ine/vitales/

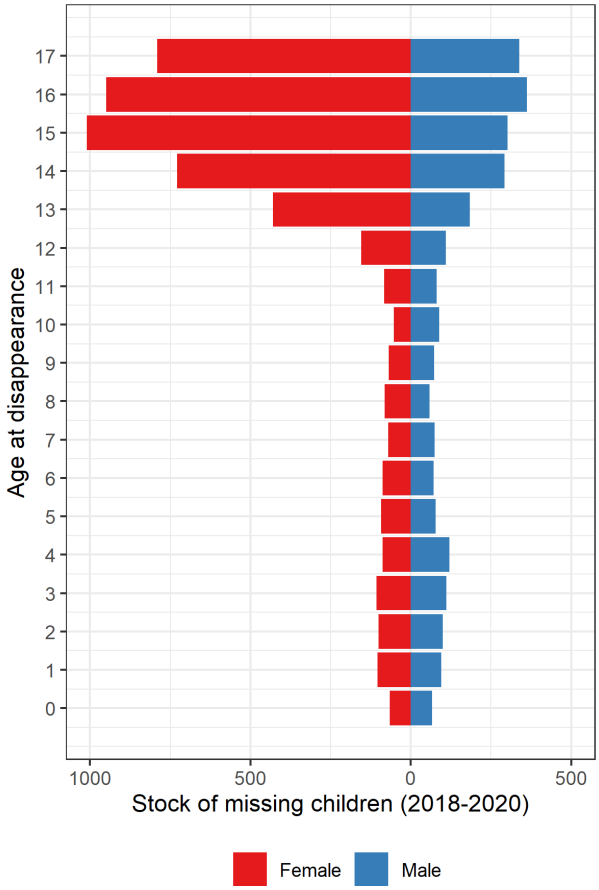



**Figure 6.5:** Cumulative number of missing children (2018–2020) according to the *Alerta Alba-Keneth* Twitter data.

we observe a high incidence of disappearance events near national borders: the eastern border with Mexico, the northwestern border with Belize, and the southeastern border with Honduras and El Salvador. Additionally, the country's only Atlantic port, Puerto Barrios, also exhibits a high number of disappearances. These regions are known for high insecurity and human trafficking activity, highlighting concerning patterns that require further attention.

We do not have data on the number of children who are found after being reported missing by *Alerta Alba-Keneth*. Nevertheless, there are reasons to believe that the number of children that are found after having gone missing is very small. According to police data obtained from this study, the number of children that are found constitutes around 5% of the cases reported every year, on average. In addition, parents may not notify the authorities of a child's "re-appearance" for a number of reasons, including fear of losing custody of the child (Evelyn Espinoza, personal communication, 19 September 2019).

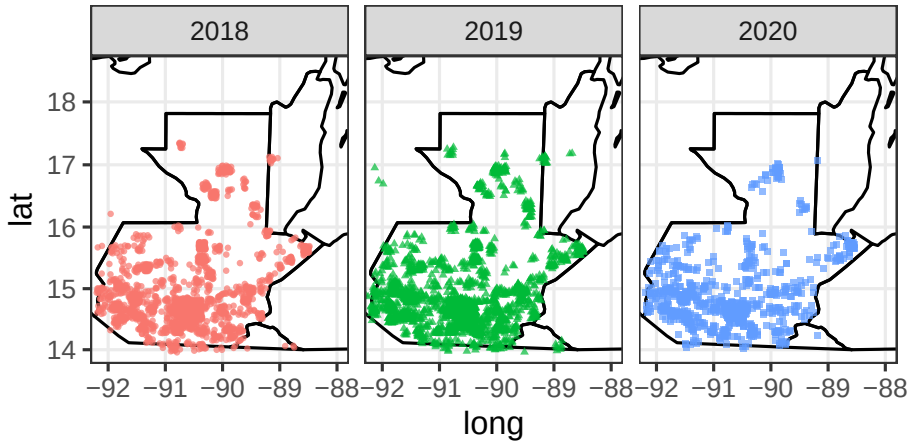



**Figure 6.6:** Geographic distribution of reported missing children (2018–2020) according to the *Alerta Alba-Keneth* Twitter data. Location information was extracted from the "place of disappearance" field in each tweet.

## 6.5 Discussion

The population of missing children is a significant social issue in many societies, particularly in contexts with high levels of violence, drug trafficking, and human trafficking. Due to the limited availability of detailed information on this population, our study leveraged Twitter data to characterize missing children by age, sex, and place of disappearance. While we present results from Guatemala, the methodology for collecting and processing data from tweets can be replicated for other countries.

Our findings reveal remarkable differences in disappearances by sex and age. Female adolescents are the most likely group to go missing. The number of missing girls was approximately twice that of boys, and around 75% of missing girls were between 13 and 17 years old. This study provides the first estimates in Guatemala of the sex and age of individuals most likely to be reported missing, using real-time, individual-level social media data. The high number of missing girls aged 13–17 suggests that sex trafficking may be a potential mechanism behind these disappearances, consistent with known links between disappearances and human trafficking (Rodas Andrade, 2021). In Guatemala, 45% of detected trafficking victims in 2019 were girls, and of these, 70% were between 14 and 17 years old (United Nations Office on Drugs and Crime, 2021).

However, due to the high levels of violence, particularly among men which, according to the National Police in Guatemala, correspond to 86% of the victims of homicide in 2020, [71] one might see our results by sex as surprising. If violence was the main mechanism behind missing children, boys would have gone missing more than girls. Our results showed the opposite. A possible explanation for this might be the age range we analyzed here. Violent deaths are less

[71] *https://infosegura.org/seccion/guatemala/*

concentrated among children (10-17 years old) when compared to individuals aged 18-30. [72] Thus, our results suggest that violence might not be the main mechanism behind disappearances in Guatemala.

This study also showed the spatial patterns of disappearance. Our findings revealed a higher number of reported cases in urban centers, while regions with higher poverty and larger indigenous populations exhibited fewer reported cases. This may reflect a limitation of our study: reporting rates likely vary across the country, and lower state presence in poorer areas could reduce the number of officially reported missing children.

Our study has several limitations. First, both the Twitter and official data rely on reported cases, so unreported disappearances remain unobserved. Second, we cannot perform individual-level comparisons between datasets, as only the Twitter data are available at that level. Thus, we cannot determine which cases appear in both datasets or those reported on Twitter but not to the police. Third, the *Alerta Alba-Keneth* Twitter account may not have tweeted all cases reported to the authorities. Fourth, our automated geocoding is subject to minor errors due to ambiguous or misspelled place names. Future work should expand data collection using regular expressions to identify other disappearance cases discussed online, including those not reported to authorities or tweeted by *Alerta Alba-Keneth*.

## 6.6 Conclusion

In this paper, we demonstrated how digital data can be leveraged to gain a better understanding of pressing social issues. Our study provided a detailed overview of missing children in Guatemala, showing that female adolescents are the most vulnerable group. This finding suggests that sex trafficking may be a significant mechanism behind these disappearances, consistent with the high concentration of human trafficking among girls between 14 and 17 years old. To reduce the incidence of child disappearances, we propose that authorities enhance protection measures specifically for young girls. Additionally, greater efforts are needed to systematically collect data on missing children in low-income settings, where this issue is particularly prevalent and socially detrimental.

[72] https://infosegura.org/2022/02/04/homicidios-guatemala-2021/

CHAPTER 7

# Investigating the Predictability of User-watching Behavior on TikTok via Data Donation

## 7.1 Introduction

In a world where attention spans are shrinking and scrolling is accelerating, TikTok has emerged as a viral social media platform. By popularizing short-format videos as a means of interaction, TikTok has reshaped the way users consume content and has become one of the world's most successful social media platforms. TikTok has captured people's attention and redefined digital engagement, raising important questions about the mechanisms that drive user interaction and engagement on the platform. From a research perspective, many open questions remain regarding users' perceptions and behaviors, particularly with respect to their watching and engagement patterns.

Previous research has shown that TikTok's algorithm learns rapidly and improves its recommendations over time, presenting users with increasingly personalized content that aligns with their interests (Karizat et al., 2021; Lee et al., 2022; Schellewald, 2024; Taylor and Chen, 2024). Personalization on TikTok is based on several factors, including following new accounts, exploring hashtags, sounds, effects, and trending topics, user engagement (e.g., likes and comments), and user watching behavior, such as whether videos are watched until the end or skipped. [73] However, while one might expect that learning users' interests would result in an increasing share of videos watched until the end, prior work found the opposite: the fraction of videos that users watch until the end remains relatively stable over time (Zannettou et al., 2024). This suggests that user watching behavior on TikTok may be more complex and less predictable than commonly assumed.

[73] https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you Accessed on August 19, 2024.

## Chapter 7. Investigating the Predictability of User-watching Behavior on TikTok via Data Donation

In this paper, we investigate the predictability of TikTok users' watching behavior, measured by whether they watch videos until the end. Our investigation adopts a twofold approach—classification and recommendation—to address the following research questions:
**RQ1:** Can we predict, and which features most effectively predict, whether a user will watch a video until the end?
**RQ2:** Can we recommend videos that users are likely to watch?

Answering these research questions requires reliable data on user watching behavior using a shared set of short-format videos. However, obtaining such data is challenging due to the highly dynamic and personalized nature of TikTok, where content exposure varies substantially across users. To address this limitation, we conducted a controlled experiment over Zoom with TikTok users recruited through Prolific. [74] Participants interacted with a curated playlist of 258 TikTok videos over a 30-minute session, during which we recorded their interactions and collected digital trace data from the TikTok accounts used in the experiment. In addition, we gathered demographic information through a survey. This experimental setup allowed us to minimize personalization effects and external distractions, enabling a controlled analysis of how video metadata and user characteristics relate to user watching behavior.

To answer our research questions, we applied two distinct approaches to the data collected. First, we treat the task as a binary classification problem in which, given a user and a video, we predict whether the user will watch the video until the end. We also identify the features that are most predictive of user watching behavior. Our results show that video metadata features are the strongest predictors, with video duration emerging as the most important factor. User demographics contribute only marginal improvements to classification accuracy. We further compare our findings from the experimental setting with real-world TikTok data (Zannettou et al., 2024) and observe similar patterns: video duration plays a key role in determining whether users watch videos until the end.

Second, we approach the task from a recommendation perspective using matrix factorization. In this setting, we model user watching behavior on a rating scale from 1 to 5, where the rating is determined by the percentage of the video watched (e.g., videos watched almost entirely receive higher ratings, while skipped videos receive lower ratings). We evaluate the performance of recommending videos to users in our experimental dataset and compare it with two additional datasets: a real-world dataset of TikTok watching behavior (Zannettou et al., 2024), also transformed into a 1–5 rating scale based on percentage watched, and the MovieLens 100K dataset (Harper and Konstan, 2015), where ratings naturally range from 1 to 5 based on user evaluations of movies. Our findings reveal that recommending short-format videos is considerably more error-prone than recommending movies. Within the experimental TikTok dataset, recommendation errors increase even further for very short videos (under 13 seconds), indicating that shorter content is particularly difficult to recommend accurately.

[74] https://www.prolific.com/

Overall, our findings provide important insights for the design of recommendation systems on short-format video platforms like TikTok. They highlight the challenges of accurately predicting user watching behavior and point to the central role of video duration in shaping watching patterns. These insights have practical implications not only for improving recommender systems to enhance user retention but also for content creators aiming to optimize viewer attention and maximize reach and virality. Additionally, our results raise important considerations for content moderation, emphasizing the unpredictability of user behavior with short-format videos and the need for improved prioritization strategies when moderating content.

## 7.2 Related work

Short-form video platforms have emerged as viral social media platforms, and many researchers have sought to understand the success of short videos, particularly on TikTok. However, questions regarding TikTok's recommendation algorithm and users' behavior remain open. In this section, we review existing literature and highlight the limited understanding of user behavior on short-form video platforms, specifically TikTok.

As one of the first studies assessing engagement with videos, Park et al. (2016) investigated the relationship between YouTube video view duration and various engagement metrics. Their findings indicate that longer view durations are positively associated with higher view counts, more likes per view, and more negative sentiment in comments. Conversely, shorter videos tend to have a higher proportion of their content viewed. Violot et al. (2024) compared user engagement and content creation between short and regular videos on YouTube, finding that short videos are primarily entertainment-focused, while regular videos cover a wider variety of categories. Regarding engagement, short videos receive more views and likes per view than regular videos but attract fewer comments per view. Similar to these studies, we explore the role of video duration in our analysis. However, we focus on TikTok users' behavior as measured by watching a video until the end.

TikTok is one of the pioneers and most popular short-video platforms, with a growing body of literature examining its use in the context of health and social support (Armin et al., 2024; Milton et al., 2023; Schluchter, 2024; Stephenson et al., 2024). Researchers have also studied factors driving user engagement and content popularity on TikTok (Dekker et al., 2025; Lee et al., 2022; Schellewald, 2023). For example, Schellewald (2023) conducted ethnographic fieldwork with young adults in the United Kingdom during 2020–2021, revealing that users view TikTok as a convenient gateway to content aligned with their interests. Dekker et al. (2025) conducted an experiment comparing user interactions with a personalized versus a less personalized TikTok "For You" feed (i.e., the personalized feed that appears when users open the app, showing content based on their interactions and preferences). They found that both frequency and duration of use declined significantly when participants engaged with the less personalized feed over a week. Finally, Lee et al. (2022) conducted 24 semi-structured interviews and proposed the

algorithmic crystal framework, comparing the algorithm's characteristics to those of a crystal: reflective, multifaceted, and refractive. TikTok's algorithm reflects and represents users' multiple interests in their "For You" feed and allows users to recognize interests refracted from other users' engagement. Users are thus exposed to different groups sharing specific interests, a finding confirmed by other researchers (Karizat et al., 2021; Schellewald, 2024; Taylor and Chen, 2024).

Unlike approaches relying on qualitative analyses of user experiences, our study measures the predictability of user-watching behavior on TikTok using survey questionnaires and Zoom-based experiments. Data were collected through surveys and TikTok data donation, following methods in previous research (Boeschoten et al., 2020; Mousavi et al., 2024; Vombatkere et al., 2024; Zannettou et al., 2024).

Vombatkere et al. (2024) used TikTok data donations to propose a framework detecting and characterizing factors contributing to feed personalization. The framework distinguishes content resulting from exploring new interests versus exploiting known interests. They identified key factors affecting TikTok feed personalization, such as following other accounts and liking videos, confirming prior results (Boeker and Urman, 2022; Klug et al., 2021).

Despite evidence that TikTok's algorithm quickly adapts to users' interests, user behavior measured by watching videos until the end does not increase. Zannettou et al. (2024) found that the fraction of videos watched until the end never exceeds 60% of those recommended by TikTok. More strikingly, most users consistently watch only 30–50% of videos until the end, with all users ranging between 10–60%.

In our research, we investigate the predictability of user-watching behavior on TikTok. To the best of our knowledge, this is the first study assessing the predictability of such behavior, contributing new insights into the dynamics of short-form video consumption.

## 7.3 Methodology

We developed a methodology for capturing realistic traces of user-watching behavior on TikTok through controlled experiments, designed to collect data that is difficult to obtain directly from real-world interactions. Our approach involves four main steps: (i) generating a TikTok playlist consisting of a realistic sequence of videos; (ii) recruiting TikTok users via crowdsourcing platforms; (iii) conducting controlled experiments in which users watch and engage with the playlist simulating TikTok's "For You" feed; and (iv) comparing observations from the experimental setting with real-world data. Each methodological step and the resulting dataset are described in detail below.

### 7.3.1 Playlist generation

To understand users' watching behavior with TikTok's short-format videos, we developed a method to present actual TikTok users with a sequence of videos generated by the platform's

recommendation algorithm. To gather a large number of samples of users viewing the same content, we created a static snapshot of the recommendations produced by TikTok.

Following previous work (Boeker and Urman, 2022; Mousavi et al., 2024), we employed an automated bot to interact with TikTok's "For You" feed to obtain video recommendations for a brand-new account. The account did not provide personal information such as gender or location, specifying only that the user was 27 years old and operating from New York via VPN. The bot exhibited random behavior: after loading the "For You" feed, it watched each video for a random duration (between 1 second and the video's full length) before swiping to the next. This process continued until the bot had viewed 300 TikTok videos. All 300 videos were then compiled into a TikTok playlist in the exact order recommended by the algorithm, which we later used to simulate the "For You" feed for our participants.

After creating the playlist, three of the paper's authors manually checked all videos to ensure that no inappropriate or harmful content was included. No video was removed due to inappropriate or harmful content. However, we checked the playlist daily and removed videos that became unavailable on TikTok. In total, 42 videos were removed due to unavailability by the end of the experiments. Although 258 videos remained in the playlist, the maximum number watched by a participant was 221. For the analysis, however, we considered only 105 videos that were watched by at least 15 users, in accordance with our threshold, totaling 127.57 minutes (7,654 seconds). For an overview of video durations in our playlist, see Figure D.2 in the Appendix.

We augmented the dataset by annotating all videos with categories based on a list of interests on TikTok [75] and on the TikTok Ads platform. [76] Three annotators independently labeled each video with a binary variable indicating whether it was related to each topic. A video was assigned to a topic if at least two annotators agreed. Inter-annotator agreement was measured using Fleiss's Kappa (Davies and Fleiss, 1982), implemented in the Python package NLTK, [77] yielding a moderate score of 0.45. These annotations allow us to investigate the relationship between video topics and user watching behavior.

### 7.3.2 User recruitment and screening survey

We used Prolific to recruit participants for our study, targeting workers over 18 years old living in the U.S. who had been using TikTok for at least one month. Our recruitment also followed

[75] Memes, Football, Food and Drink, Celebrities, Fashion, Music, Entertainment, Gaming, Beauty, Sports, Tutorials, Travel, Learning, Anime and Cartoons, Oddly Satisfying, Art and Design, Vlogs, Health and Fitness, Auto, Technology and Science, DIY, Extreme Sports, Dance

[76] Tech & Electronics, Baby, Kids & Maternity, Life Services, E-Commerce (Non-app), Beauty & Personal Care, Education, Financial Services, Games, Vehicles & Transportation, Business Services, Travel, Sports & Outdoors, Food & Beverage, Apparel & Accessories, Home Improvement, Household Products, Apps, Pets, Appliances, News & Entertainment

[77] https://tedboy.github.io/nlps/generated/generated/nltk.AnnotationTask.html

the age–sex distribution of TikTok users in October 2023, according to Statista demographics. [78] Participants provided informed consent and completed a five-minute screening survey, which included questions on demographic characteristics (e.g., age, gender, education, spoken languages, and income) and social media usage. Participants were compensated one pound for completing the survey. At the end of the screening survey, participants had the option to agree to participate in our follow-up study and, if so, book a time slot. The controlled experiments were conducted in September 2023. Of the 1,000 participants recruited on Prolific, 245 (24.5%) booked a time slot for the follow-up experiments, and eventually 108 (10.8%) completed the experiments. Figure 7.1 illustrates the recruitment procedure flow and shows the number of participants who completed or were lost at each stage.

### 7.3.3 Controlled experiments

Participants who booked a time slot for the follow-up study were invited to join a Zoom call for the controlled experiment. The controlled setup ensured consistency and comparability in the data collected. By having all participants watch the same playlist under controlled conditions, we ensured that each participant experienced the same content exposure and viewing duration, thereby enhancing the reliability of our findings. Conducting the experiment via Zoom also allowed researchers to monitor participants in real-time, provide instructions, and address any technical issues or questions. This level of oversight enabled a more accurate assessment of user interactions with the videos, leading to robust conclusions. Zoom was chosen not only for control but also for flexibility, allowing participants to join without commuting to a physical location. Previous studies have successfully used Zoom for experiments (Archibald et al., 2019; Falter et al., 2022; Reñosa et al., 2021), particularly during the COVID-19 pandemic.

In total, 108 participants joined 21 Zoom calls, with up to 15 participants per call and a maximum duration of 90 minutes. Participants were compensated 11 pounds per hour via bonus payment on Prolific. During the experiment, each participant received credentials for a dedicated TikTok account. Participants logged in using the provided credentials and were guided to the experimental playlist. The playlist was displayed as a TikTok "For You" feed, simulating the interface users encounter in their own accounts. Participants were instructed to watch the playlist for 30 minutes, mimicking typical TikTok usage. After interacting with the playlist, participants logged out of the TikTok account and completed a survey regarding their experience and their TikTok interests, which were used for further analysis.

In addition to survey responses, our main analysis relied on digital trace data collected from each TikTok account (Boeschoten et al., 2020). This included browsing history, such as what videos they watched and for how long. Using these traces, we get access to the time spent on each video and whether the videos were watched until the end or skipped by each participant.

[78] *https://www.statista.com/statistics/1299771/tiktok-global-user-age-distribution/*

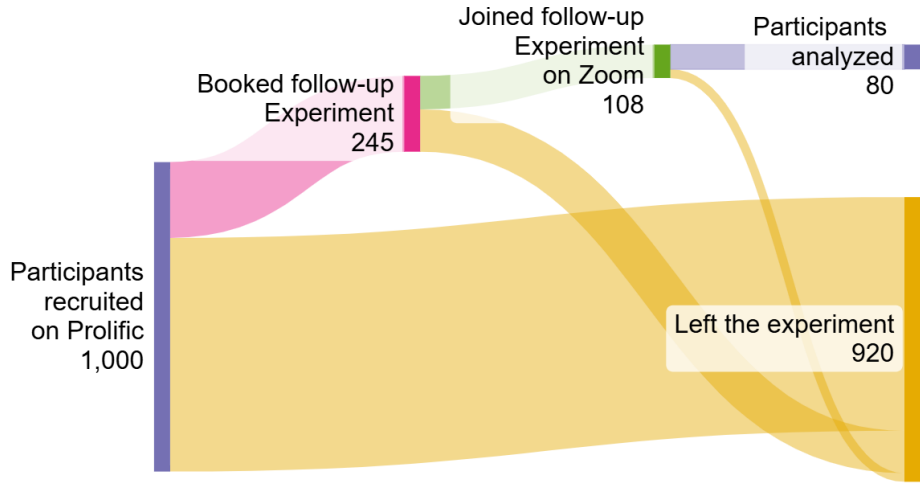



**Figure 7.1:** Experiment flow. Nodes represent the stages of the experiment, and the flows indicate the number of participants transitioning between stages.

After the experiment, the digital traces were requested from TikTok and pre-processed. Of the 108 participants, some were excluded based on the following criteria: $(i)$ not watching videos in the order presented in the playlist, and $(ii)$ watching fewer than 15 videos. We set 15 videos as the lower bound since a 30-minute session corresponds to watching the first 23 videos until the end. After applying these criteria, 80 participants remained for analysis, as shown in Figure 7.1.

Prior to the main experiment, we conducted two pilot studies to test the procedure. The first pilot involved two male participants, both researchers from a co-author's laboratory, with differing ages and TikTok familiarity: one was 25–34 years old with over a year of experience, and the other was 45–54 years old and new to the platform. The second pilot recruited two participants via Prolific, one male and one female, aged 18–24, both with over a year of TikTok experience. These pilot studies helped refine our methodology. Based on participant feedback, we increased the playlist watching time from 15 to 30 minutes, clarified survey questions, and validated the overall Zoom experiment procedure. During the pilots, participants provided feedback on survey clarity and shared their TikTok usage experiences, ensuring that all instructions and questions were well understood.

**Demographics:** We recruited participants to match the distribution of TikTok users by sex and age group in October 2023, according to Statista Demographics. Figure D.1 in the Appendix shows the percentage of participants included in our analysis compared to TikTok users by age and sex. Tables D.1 and D.2, also in the Appendix, provide an overview of participants' demographics and TikTok usage characteristics, respectively. For each demographic group, we report the most prevalent category, emphasizing the most common category in bold. Overall, similar to the official TikTok user distribution by sex and age, our participants are mostly young adults born in the 1990s. Most participants self-identified as women, the majority were White, held a bachelor's degree, were employed full-time, identified as Democrats, and had been using TikTok for over a year, typically several times a day for about 10–30 minutes.

### 7.3.4 Real-world datasets

To achieve meaningful results, our methodology relies on a synthetic and curated dataset, as understanding user-watching behavior on TikTok is challenging with typical unstructured datasets. The controlled nature of our data collection allows us to capture specific user interactions and viewing patterns that are difficult to observe in real-world datasets, enabling clear and interpretable results. However, to enhance reproducibility and contextualize our findings, we compared results from our experimental dataset with those from real-world datasets.

**TikTok real-world data:** First, we used a TikTok dataset created by Zannettou et al. (2024), which was collected via data donation and includes the watch history of approximately 350 TikTok users covering more than 9.2M TikTok videos. To align the real-world dataset with our experimental setting for classification, we focused on users from North/Central America and limited each user to their first 105 videos, corresponding to the maximum watched in the experiment. For matrix factorization, we used the entire dataset from Zannettou et al. (2024), which includes users from multiple regions. We preprocessed the data by retaining only users who watched at least 15 videos and videos watched by at least 15 users. This approach mirrors the preprocessing of the experimental dataset, reduces extreme sparsity, and makes the dataset more comparable to other recommendation datasets such as MovieLens.

**MovieLens 100K data:** Second, to benchmark the performance of matrix factorization on a more standard recommendation dataset, we used the MovieLens 100K dataset (Harper and Konstan, 2015). This dataset consists of 100,000 movie ratings from 943 users on 1,682 movies. Ratings are given on a scale from 1 to 5, which we use as an analog to the range of a video watched in our TikTok datasets. Including MovieLens allows us to compare recommendation performance on a platform with longer-format, highly rated items to our short-format TikTok videos, highlighting the additional challenges of predicting watching behavior with very short videos.

### 7.3.5 Ethical considerations

Before recruiting participants and collecting data, we obtained approval from the Ethical Review Board at the institution where the first author is based. We submitted a detailed document outlining the data collection process, anonymization methods, and participant consent procedures. Participants were recruited through Prolific, where they were provided with a consent form and informed about the study's goals, potential privacy implications, and the types of data collected. All participants gave explicit consent to participate and were compensated via Prolific. Those who participated in the follow-up experiment received additional compensation as a bonus, proportional to the time spent in the study.

In line with ethical standards, all sensitive content in users' data was anonymized or removed (e.g., IP addresses in TikTok data were removed before uploading to secure servers). Data were coded with randomized identifiers and stored on secure, password-protected servers. Reported

results are aggregated, and no attempt was made to track individual users. The real-world dataset from (Zannettou et al., 2024) was anonymized before being shared with us, and the MovieLens dataset (Harper and Konstan, 2015) is also anonymized. No individual-level data will be shared with third parties. For reproducibility, we provide access to the aggregated data and the code used for our analyses and plots in a public web repository. [79]

## 7.4 Results

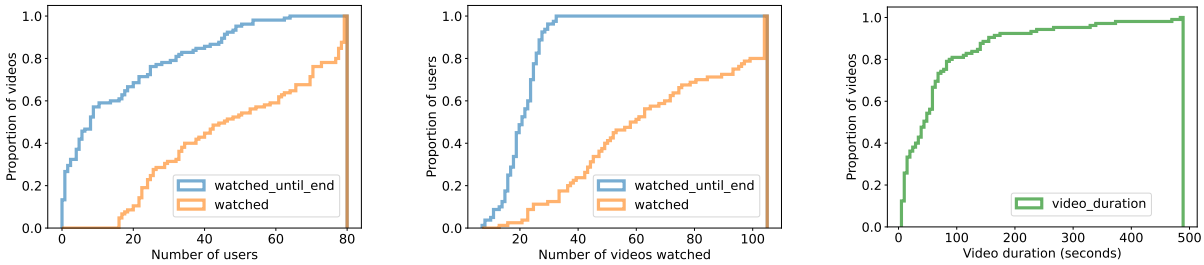
Our analyses focus on understanding TikTok user-watching behavior, specifically whether users watch videos until the end. We first conducted exploratory data analysis on our dataset, which combines TikTok trace data, screening survey responses, and the controlled experiment. Next, we address our first research question (**RQ1**): Can we predict, and which features most effectively predict, whether a user will watch a video until the end? To answer this, we developed a classification model to assess predictability and identify the features most influential in determining whether a user watches a video until the end. We then focus on the second research question (**RQ2**): Can we recommend videos that users are likely to watch? To answer this, we applied a matrix factorization approach to decompose the user–video interaction matrix into two low-rank matrices: a user-factor matrix and a video-factor matrix. These latent representations capture user preferences and video characteristics, enabling the prediction of new videos that users are likely to watch. Finally, we compare results from our experimental dataset with those from a real-world dataset, evaluating both classification and matrix factorization analyses.

### 7.4.1 Descriptive statistics

Our dataset combines two main sources: digital trace data collected from TikTok via data donation and API, and survey responses from the participant screening and the controlled experiment. During the exploratory analysis, we carefully merged and examined these data sources for subsequent analyses.

We explored the characteristics of the playlist created for the experiment and the interactions between participants and the videos. Figure 7.2 presents several cumulative distribution functions (CDFs) related to the videos watched by participants. Figure 7.2a shows the proportion of videos watched and watched until the end by participants. The figure includes two CDFs: (i) the proportion of videos watched by participants, and (ii) the proportion of videos watched until the end. The first line (i) begins after 15 participants because the dataset was constructed so that each video was watched by at least 15 users. This threshold ensures sufficient data for analyzing participants' watching behavior. We observe that half of the videos were watched by at least 46 users, up to the total of 80 participants. The second line (ii) shows that 50% of the videos were watched until the end by 8 or fewer participants, with a maximum of 64 users. Next, Figure 7.2b

[79] https://github.com/carolcoimbra/unpredictable-tiktok




**(a)** Users **(b)** Videos **(c)** Video duration

**Figure 7.2:** CDF of video durations in the playlist, along with the number of videos and users who watched them. Colors indicate the following: orange represents videos that were watched, blue represents videos watched until the end, and green represents the video duration.

shows the proportion of users who watched videos from the playlist. The figure presents two CDFs: (i) the proportion of users who watched videos, and (ii) the proportion of users who watched videos until the end. The first line (i) starts at 15 videos, reflecting that our dataset includes only participants who watched at least 15 videos. This threshold was defined based on the cumulative distribution of video durations, where the first 23 videos sum to 30 minutes (see Figure D.2 in the Appendix). This ensures sufficient data to analyze participants' watching behavior. Half of the participants watched at least 59 videos, with a maximum of 105 videos. The second line (ii) shows that 50% of users watched until the end only 21 or fewer videos, while the maximum number of videos watched until the end by any participant was 33. Finally, Figure 7.2c shows the cumulative distribution of video durations in the playlist. While most videos range between 5 and 81 seconds, 25% of the videos have durations between 82 and 489 seconds.

Figure 7.3 shows the percentage of each video's duration that participants watched. Columns represent videos in playlist order, while rows represent users, sorted by the number of videos each user watched. Several patterns are apparent in the heatmap. Most participants watched the first two videos in full, whereas the next seven videos were typically watched for less than 20% of their duration. From the 10th to the 20th video, the percentage watched increased, with many participants watching these videos until the end. To better understand this variation, we manually examined videos that were not watched until the end by most participants. Although we found no clear pattern in terms of topic or content, these videos tended to be longer, specifically, longer than one minute, providing initial evidence that video duration may be a key factor influencing user-watching behavior.

Figure 7.4a shows the duration of the videos in our playlist. We observe that the third video in the playlist is almost 500 seconds long. We contrasted video duration with the percentage of each video that participants watched. Figure 7.4b shows a moderate negative correlation (Pearson's r = -0.41) between video duration and the percentage watched, indicating that participants tend to watch a smaller portion of longer videos.

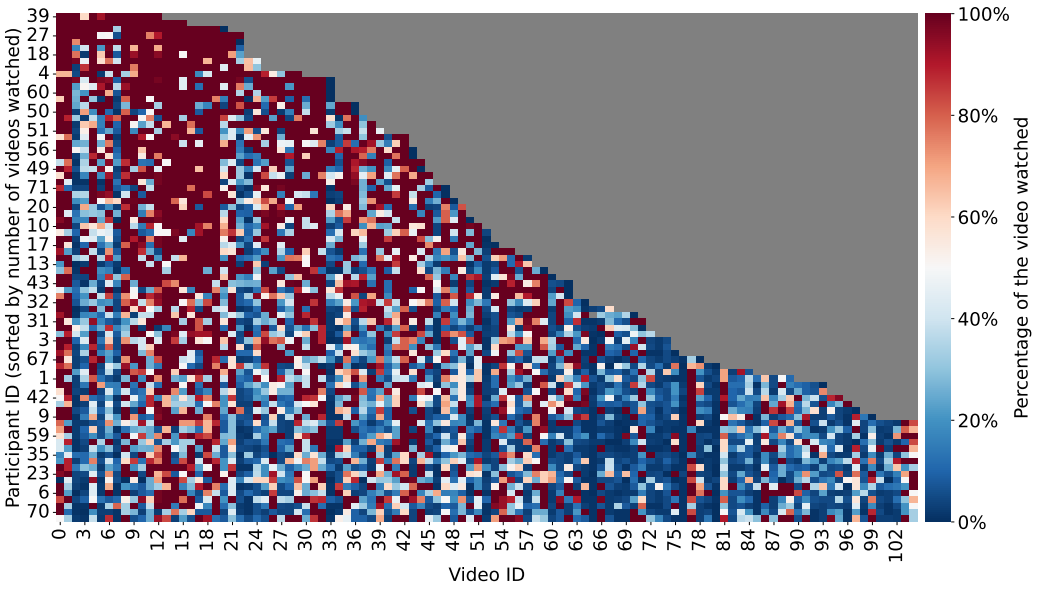
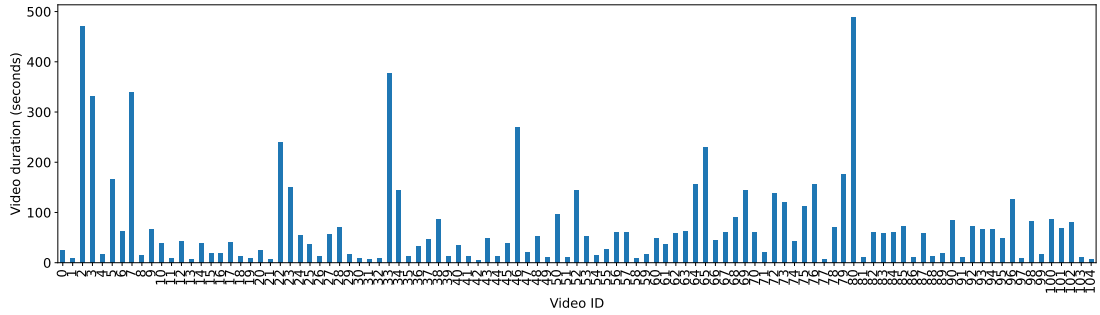



**Figure 7.3:** User-watching behavior as the percentage of each video's duration (columns) that participants (rows) watched. Cell colors indicate the proportion watched, ranging from 0% (dark blue) to 100% (dark red), while gray cells represent videos that the participant did not reach in the playlist. Columns are ordered according to the order in which the videos appear in the playlist, and rows are sorted by the number of videos each participant watched.

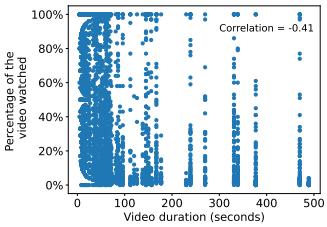
We also examined the relationship between video duration and the percentage of participants who watched the videos until the end. Figure 7.4c shows one point per video and reveals a moderate negative correlation (Pearson's r = -0.5). This suggests that the proportion of participants watching a video until the end decreases as video duration increases. Additionally, the variability in the percentage of participants who watch videos until the end decreases for longer videos, indicating a more consistent early drop-off for longer videos.

Finally, we analyzed the percentage of videos each user watched until the end. Figure 7.4d shows that most users watched between 23% (25th percentile) and 56% (75th percentile) of the videos until the end. This aligns with Zannettou et al. (2024), who reported that 70% of participants watched between 30% and 50% of videos in their viewing history until the end. However, while Zannettou et al. (2024) found no participant who watched more than 65% of videos until the end, we observed some participants who watched all videos until the end.

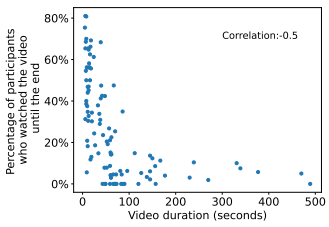
Our preliminary analysis reveals that more than half of the videos were watched by over 50% of participants, yet only 8 or fewer participants watched these videos until the end. The number of videos watched by participants in our study ranges from 15 to 105. Overall, participants watched many videos (half of them watched 60 or more), but relatively few were watched until the end. In fact, most users watched between 23% and 56% of the videos until the end. We also observed a clear pattern in participants' watching behavior: the percentage of participants who watched a video until the end is negatively correlated with both the average proportion of



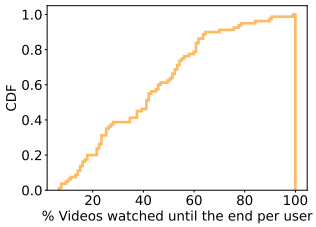

**(a)** Video duration across the playlist




**(b)** Relationship between video duration and the percentage of each video watched by participants




**(c)** Relationship between video duration and the percentage of users who watched each video until the end




**(d)** CDF showing the percentage of videos watched until the end per participant

**Figure 7.4:** Overview of video duration and user watching behavior. The majority of TikTok users in the dataset watch until the end between 20% and 60% of all videos they watched.

the video duration watched and the video duration. Next, we investigate the predictability of user-watching behavior as a classification task.

### 7.4.2 RQ1: Can we predict, and which features most effectively predict, whether a user will watch a video until the end?

Our first research question focuses on predicting whether a user will watch a video until the end. We simplify the problem as a binary classification task, where the outcome is 1 if the video is watched until the end and 0 otherwise. The classification task consists of three steps. First, we evaluated a set of classification algorithms to select the most suitable model. Next, we tested different combinations of features to identify the best-performing model and determine which features are most predictive of user-watching behavior. Finally, we compare the performance of our model, trained on data collected from the controlled experiment, with a model applied to real-world TikTok data (Zannettou et al., 2024).

## Features

Our dataset includes features grouped into three main categories: video metadata, user demographics, and TikTok usage patterns.

**Video metadata:** Features describing the content and popularity of each video, including duration, number of likes, comments, shares, plays, and the order in which the video appears in the playlist.

**User demographics:** Features providing information about the user's background, such as year of birth, gender, language proficiency, education level (i.e., the highest school degree), political leaning, income, race/ethnicity, and employment status.

**Usage patterns:** Features capturing how users interact with TikTok, including the frequency and duration of usage, the number of videos engaged with, and account type (e.g., personal or business, viewer or content creator) as reported in the screening survey.

In addition, we created a feature to capture the similarity between the topics annotated for each video and the participant's reported interests. In the screening survey, participants indicated their interests on TikTok using the same list of topics. We then computed an Interest Similarity score between each participant $p$ and video $v$ using a variation of the Jaccard similarity. Let $T_p$ denote the set of topics reported by participant $p$ and $T_v$ denote the set of topics annotated for video $v$. The Interest Similarity is calculated as: $IS(T_p, T_v) = \frac{|T_p \cap T_v|}{|T_v|}$ This score ranges from 0 (no overlap between the participant's interests and the video topics) to 1 (all topics annotated for the video match the participant's interests).

All the features used in our models, along with their type and a brief description, are listed in Table 7.1. Categorical features, such as gender, race/ethnicity, and employment status, were preprocessed using one-hot encoding to allow the model to interpret non-numeric data. All features were standardized by removing the mean and scaling to unit variance.

## Model selection

We tested six algorithms for classification: Logistic Regression, K Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random forest, and Multi-Layer Perceptron (MLP). Table 7.2 shows the performance of each model. For all models, we used a random search to select hyperparameters that optimize performance. We also applied 5-fold cross-validation to validate the model's generalizability, splitting the dataset into 80% training and 20% testing. Model performance was evaluated based on accuracy, precision, recall, and F1 score.

The Random forest classifier achieved the highest accuracy among the models tested. Therefore, we focus on this model for the remainder of the analyses due to its superior performance, robustness in handling high-dimensional datasets, and interpretability via feature importance scores, which provide insights into the key factors driving users' behavior of watching videos until the end.

| Feature name | Type | Description |
|---|---|---|
| Video ID | Numerical | Identifier for the video considering the order of its inclusion in the playlist (in the experimental setting) or the order in which the video is watched by the user (in the real-world setting). |
| Video duration | Numerical | Length of the video in seconds. |
| Video num. likes | Numerical | Number of likes the video has received. |
| Video num. shares | Numerical | Number of times the video has been shared. |
| Video num. comments | Numerical | Number of comments the video has received. |
| Video num. plays | Numerical | Number of times the video has been played. |
| User ID | Categorical | Identification number for each user. |
| Year born | Numerical | Birth year of the user. |
| Gender | Categorical | Gender reported by the user. |
| Race/Ethnicity | Categorical | Race/ethnicity reported by the user. |
| Language (e.g., English) | Numerical | For each language the value represents the proficiency level on a scale of 1 to 5, where 1 represents basic proficiency and 5 represents native. |
| School degree | Numerical | Highest level of school reported by the user. |
| Employment status | Categorical | Employment status reported by the user. |
| Political leaning: Republican | Numerical | The value represents the degree of Republican-leaning. |
| Political leaning: Democrat | Numerical | The value represents the degree of Democratic-leaning. |
| Income (annual) | Numerical | The user's annual income level. |
| Interest Similarity | Numerical | Similarity (measured as a variation of the Jaccard Similarity) between the video's topics and the participants' topics of interest. |
| How long use TikTok | Numerical | Duration in months of how long the user has a TikTok account. |
| How often access TikTok | Numerical | Frequency of accessing TikTok. |
| How many videos engage with | Numerical | Number of videos with which the user interacts. |
| Avg time per day using TikTok | Numerical | Average daily usage time in the past week the user spent on TikTok. |
| TikTok viewer vs. creator: Viewer | Categorical | Whether the user views content on TikTok (Yes/No). |
| TikTok viewer vs. creator: Creator | Categorical | Whether the user creates content on TikTok (Yes/No). |
| TikTok personal vs. business: Personal | Categorical | Whether the user uses TikTok for personal purposes (Yes/No). |
| TikTok personal vs. business: Business | Categorical | Whether the user uses TikTok for business purposes (Yes/No). |
| When access TikTok | Categorical | Moment when the participants watch TikTok. |

**Table 7.1:** Description of the features used in our models.

## Findings: Classification

The Random forest classifier works by building multiple decision trees during training, and outputs the class corresponding to the final prediction, in this case, whether a video is watched until the end. We tested a variety of models to provide a comprehensive view of the features that could influence users' behavior of watching a video until the end.

The **Model using all features** includes all available features in our dataset, as listed in Table 7.1. By including all features, we aim to identify which factors are most significant in predicting whether a user will watch a video until the end. The F1 score for this model is 0.74. Figure 7.5 shows the confusion matrix and ranks all features by their importance in the model. For categorical features, which were preprocessed through one-hot encoding, feature importances were aggregated using the median. The most important features are related to video metadata, the similarity between the video's topics and the participant's interests, the participant's year of birth, and income, followed by several TikTok usage patterns, the highest school degree, political leaning, gender, and race/ethnicity.

| Model | F1 Score | Accuracy | Precision | Recall |
|---|---|---|---|---|
| Logistic Regression | 0.72 | 0.74 | 0.72 | 0.75 |
| K Nearest Neighbors | 0.68 | 0.74 | 0.71 | 0.67 |
| SVM | 0.72 | 0.73 | 0.71 | 0.74 |
| Decision Tree | 0.7 | 0.72 | 0.7 | 0.72 |
| **Random forest** | **0.74** | **0.78** | **0.75** | **0.74** |
| MLP | 0.72 | 0.76 | 0.72 | 0.72 |

**Table 7.2:** Evaluation of models' performance on our experimental dataset using all the features.

Given the importance of video metadata features, we implemented a model to assess the predictability of user-watching behavior using only video metadata. To account for multiple views of the same video by different participants, we included a categorical feature representing the user ID, which was preprocessed via one-hot encoding. The **Model using video metadata** considers only video metadata features to predict the likelihood of a TikTok video being watched until the end. The F1 score for this model drops slightly from 0.74 (Model using all features) to 0.73.

As a baseline for comparison, we evaluated a **Random Model**. This model predicts whether a video will be watched until the end based on the overall fraction of videos watched until the end in the dataset. In our data, 32% of videos were watched until the end, while 68% were skipped. The random model assigns a class to each video in the test set with probabilities reflecting this distribution, in our case, a 0.32 probability of being watched until the end and 0.68 of being skipped. The F1 score for the random model is 0.49, and all other evaluation metrics are lower than those obtained by the classification models, confirming that our models perform substantially better than chance.

In the experimental setting, our models outperformed a simple random model, highlighting the predictive value of video metadata for classification. To evaluate the generalizability of these findings, we extended our analyses to the real-world dataset. Video metadata is particularly suitable for this comparison, as it is the only feature set consistently available in both datasets. For the comparison, we selected video metadata features present in both datasets: video duration and the number of likes, shares, comments, and plays. In the real-world dataset, the order in which videos appear in the playlist was replaced by a numerical variable representing the order in which each user watched the videos. Additionally, we included a categorical feature representing the user ID.

To better approximate the overlap of videos watched by multiple participants, we divided the real-world dataset into regional sub-samples: Africa, North/Central America (N/C America), Europe, and South America. Since our experiment targeted U.S. participants, we focused on the North/Central America subset. To align with the experimental dataset, we limited each user to their first 105 videos, corresponding to the number of videos considered in our study. After this processing, the real-world subset remains much sparser, with 9,001 observations from 108 users

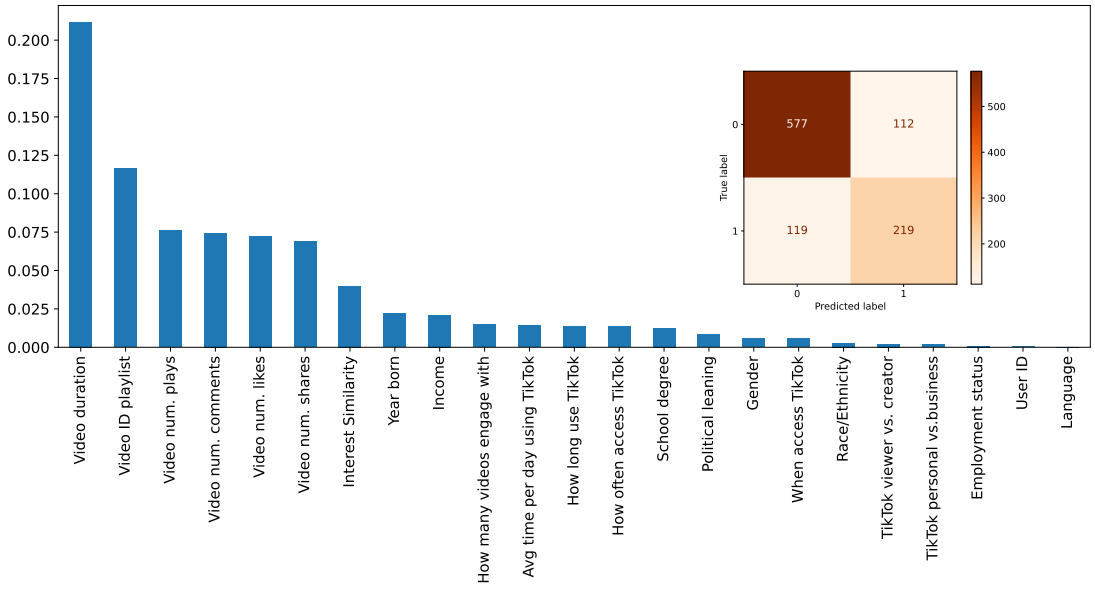
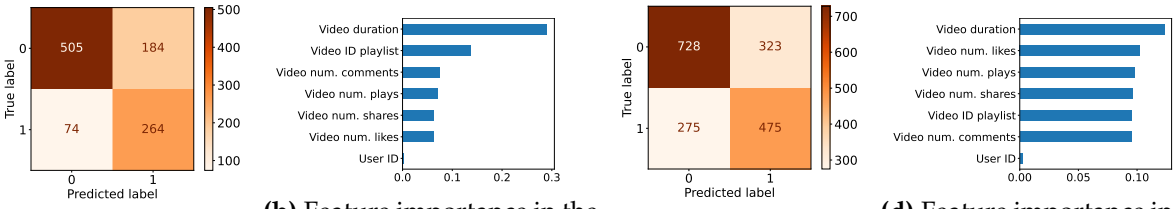



**Figure 7.5:** Feature importance and confusion matrix for predicting whether a TikTok video will be watched until the end. The confusion matrix illustrates the performance of the classification model, while the feature importance bar plot highlights the key factors influencing the model's predictions. Higher values indicate greater significance in determining the likelihood of a TikTok video being watched until the end.




**(a)** CM in the experimental setting

**(b)** Feature importance in the experimental setting

**(c)** CM in the real-world setting

**(d)** Feature importance in the real-world setting

**Figure 7.6:** Confusion matrix (CM) and feature importance for predicting whether a TikTok video will be watched until the end using only video metadata. The model is evaluated on two datasets: the experimental dataset and the real-world dataset from North/Central America.

across 7,541 videos, compared to the experimental dataset, which contains 5,135 observations from 80 users and 105 videos. Figure D.3 in the Appendix compares the distribution of key features between the two datasets.

Figure 7.6 shows the confusion matrix and feature importance for the model applied to both our experimental dataset and the real-world dataset. Overall, the magnitude of feature importance in the real-world setting is smaller than in the experimental setting. However, the most notable result is the consistently high importance of video duration for predicting whether a video will be watched until the end in both datasets.

| Model | F1 Score | Accuracy | Precision | Recall | AUC-ROC |
|---|---|---|---|---|---|
| Random (based on the experiment) | 0.49 | 0.55 | 0.49 | 0.49 | 0.49 |
| All features (experiment) | 0.74 | 0.78 | 0.75 | 0.74 | 0.83 |
| Video metadata (experiment) | 0.73 | 0.75 | 0.73 | 0.76 | 0.82 |
| Video metadata (real-world data in N/C America) | 0.66 | 0.67 | 0.66 | 0.66 | 0.71 |

**Table 7.3:** Evaluation of the Random forest models' performance.

Table 7.3 summarizes the performance of Random forest models trained on the experimental and real-world datasets, with a random model included as a baseline. The model using all features from the experimental dataset achieved the highest F1 score of 0.74. However, including demographic features added only marginal improvement, as the model using only video metadata reached an F1 score of 0.73, indicating that video metadata alone is a strong predictor of user-watching behavior. In contrast, the video metadata–only model trained on the real-world dataset achieved a lower F1 score of 0.66. The reduced performance is likely due to the sparsity of the dataset and the more complex, uncontrolled engagement patterns in real-world settings.

Overall, our classification analysis highlights that the most important predictors of whether a user watches a video until the end are features related to video metadata. User demographics contribute only marginally to model performance. Among video metadata features, video duration consistently emerges as the most influential factor. The importance of video duration persists when the model is applied to the real-world dataset, although predictive accuracy is lower outside the controlled experimental setting.

### 7.4.3 RQ2: Can we recommend videos that users are likely to watch?

Our second research question examines TikTok user-watching behavior from an algorithmic perspective, aiming to assess whether we can recommend videos that a user is likely to watch until the end. We frame this as a recommendation task and employ matrix factorization, a widely used collaborative filtering technique that predicts user preferences by leveraging observed user–item interactions (Bokde et al., 2015).

In the matrix factorization approach, a typically sparse user–item interaction matrix is approximated as the product of two lower-rank matrices capturing latent user and item factors. By embedding users and items in a shared latent space, matrix factorization can infer unobserved interactions and estimate the likelihood of a user engaging with an unseen item. This method also allows us to evaluate recommendation performance across specific segments of the dataset, such as subsets of videos stratified by duration.

To operationalize the approach, we first constructed a user–video interaction matrix, where each entry represents the range of video duration watched by the user. We then applied matrix factorization to the experimental dataset, systematically tuning hyperparameters to identify the best-performing configuration. Given prior evidence that video duration strongly influences

user watching behavior, we further assessed performance across subsets of videos grouped by duration.

To contextualize our findings, we applied the same matrix factorization procedure to two real-world datasets. The first is a large-scale TikTok dataset collected via data donations (Zannettou et al., 2024), capturing user–video interactions on TikTok. The second is the MovieLens 100K dataset (Harper and Konstan, 2015), [80] which contains explicit ratings from 1 to 5 stars for movies. MovieLens is a widely used benchmark in recommendation systems research and allows for performance comparison across datasets with explicit versus implicit feedback. While TikTok data captures implicit engagement through watch behavior, MovieLens provides explicit ratings, enabling an assessment of the challenges in short-format video recommendation relative to traditional recommendation scenarios.

## Matrix Construction

We constructed a user-item interaction matrix $R$, where each row corresponds to a user and each column corresponds to a video (or movie, for MovieLens). For the TikTok datasets, the interaction value reflects the percentage of the video's duration watched, discretized into ordinal levels as follows:

$$R_{u,v} = \begin{cases} 0 & \text{if user } u \text{ did not watch video } v \\ 1 & \text{if user } u \text{ watched 0–20\% of } v \\ 2 & \text{if user } u \text{ watched 21–40\% of } v \\ 3 & \text{if user } u \text{ watched 41–60\% of } v \\ 4 & \text{if user } u \text{ watched 61–80\% of } v \\ 5 & \text{if user } u \text{ watched 81–100\% of } v \end{cases}$$

We applied the same interaction matrix formulation to both the controlled experiment dataset and the real-world TikTok dataset obtained via data donation (Zannettou et al., 2024). In the controlled experiment, all users were exposed to the same set of videos, resulting in relatively dense interactions across users. In contrast, the real-world TikTok dataset is highly sparse because users interact with videos through personalized feeds, leading to limited overlap in the videos watched by different users. This dataset includes donors from multiple regions—primarily North and Central America and parts of Africa—introducing additional heterogeneity in user behavior. To reduce extreme sparsity, we retained only users who watched at least 15 videos and videos that were watched by at least 15 users. This preprocessing mirrors the construction of the experimental dataset, where the same thresholds were applied.

[80] https://grouplens.org/datasets/movielens/100k/

| Dataset | Users | Items | Interactions | Sparsity |
|---|---|---|---|---|
| TikTok Experiment | 80 | 105 | 5,135 | 0.3887 |
| TikTok Real-World | 334 | 21,416 | 598,251 | 0.9164 |
| MovieLens 100K | 943 | 1,682 | 100,000 | 0.9370 |

**Table 7.4:** Summary statistics of the datasets used for the matrix factorization analysis.

As a benchmark, we used the MovieLens 100K dataset, which contains explicit ratings from 1 to 5 stars. In this case, each matrix entry $R_{u,v}$ represents the rating given by user $u$ to movie $v$, or 0 if no rating was provided.

Table 7.4 summarizes key statistics for all three datasets, including the number of users, items (videos or movies), observed interactions, and sparsity. The table highlights that the controlled TikTok experiment is relatively dense (sparsity = 0.39), whereas the real-world TikTok and MovieLens datasets are highly sparse (0.92 and 0.94, respectively), illustrating the challenges of recommendation tasks in real-world settings.

## Matrix Factorization Model

We implement matrix factorization using the Truncated Singular Value Decomposition (SVD) algorithm from scikit-learn [81] in Python. The model is applied to a user–video interaction matrix, where each entry represents a rating from 1 to 5 or 0 if there is no user-item interaction. Prior to factorization, we normalize the interaction matrix by subtracting the global mean rating and incorporating user- and item-specific bias terms. This baseline correction accounts for systematic tendencies, such as users who consistently interact more with popular items, allowing the latent factor model to focus on residual interaction patterns that are more informative for personalization.

Following the procedure used for our classification model, we conduct a grid search to identify the optimal hyperparameters for SVD, including the number of latent components and the maximum number of iterations. The model is trained on a subset of the data and evaluated on a held-out test set. To ensure a robust evaluation, we adopt a user- and item-aware train–test split. First, the data is split into 80% training and 20% testing. To avoid cold-start issues, any user or item appearing in the test set but not in the training set has one of their interactions reassigned to the training set. This guarantees that every user and video in the test set has at least one corresponding interaction in the training set, allowing the model to learn latent representations for all entities.

[81] *https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html*

Model selection is based on the Root Mean Square Error (RMSE), [82] which quantifies the deviation between observed and predicted ratings. Lower RMSE values indicate a more accurate reconstruction of the interaction matrix and better recommendation performance. To facilitate comparisons across datasets with different rating distributions, we also report the Normalized RMSE (NRMSE), computed by dividing RMSE by the standard deviation of the observed ratings (Shani and Gunawardana, 2010). In addition, we evaluate the model on the test set using standard classification measures, including accuracy, precision, recall, and F1 score, to assess how well the model predicts the true values of user-item interaction.

Finally, we compute the Normalized Discounted Cumulative Gain (NDCG) [83] to evaluate the quality of ranked recommendations. Using a leave-one-out (LOO) approach, one user is left out at each iteration while the model is trained on the remaining N-1 users and tested on the left-out user. We calculate the average NDCG among the top 10 items recommended to each user (NDCG@10). NDCG scores range from 0 to 1, with 1 indicating an ideal ranking where all relevant items appear at the top of the recommendation list. The same LOO procedure is applied to compute RMSE and NRMSE for unseen users, serving as an additional check of model performance using a different validation approach.

## Findings: Matrix Factorization

To evaluate the performance of matrix factorization in modeling user-item interactions, we applied this technique to both TikTok datasets and the MovieLens 100K benchmark. This analysis allows us to quantify how effectively latent factor models can predict whether users watch videos until the end and to compare performance between short-format video watching behavior and traditional recommendation scenarios. We evaluate model performance using multiple metrics, including accuracy, precision, recall, F1 score, RMSE, NRMSE, and NDCG@10. In addition, we use leave-one-out cross-validation (LOOCV) to assess the model's generalization to unseen users.

Table 7.5 reports the performance of the matrix factorization models across the three datasets using multiple evaluation metrics. In the TikTok datasets, user-item interaction is represented by an ordinal number reflecting the proportion of the video watched, whereas in MovieLens, it is captured through explicit ratings. The results indicate that user engagement with short-form videos, measured through watching behavior, is considerably more variable than engagement with long-form content such as movies, which is based on ratings. Despite similar levels of sparsity in the real-world TikTok dataset and MovieLens 100K (0.92 and 0.94, respectively), the NRMSE is considerably higher for TikTok (0.934 versus 0.835 for MovieLens), indicating greater uncertainty and noise in modeling watch behavior for short-form content. Even in the controlled TikTok experiment dataset, which is comparatively dense (sparsity = 0.39), NRMSE remains

[82] *https://scikit-learn.org/stable/modules/generated/sklearn.metrics.root_mean_squared_error.html*

[83] *https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ndcg_score.html*

| Data | Accuracy | Precision | Recall | F1 score | RMSE | NRMSE | NDCG@10* | RMSE* | NRMSE* |
|---|---|---|---|---|---|---|---|---|---|
| TikTok Experiment | 0.373 | 0.348 | 0.301 | 0.287 | 1.333 | 0.809 | 0.807 | 1.320 | 0.788 |
| TikTok Real-World | 0.185 | 0.305 | 0.211 | 0.145 | 1.547 | 0.934 | 0.055 | 1.548 | 0.951 |
| MovieLens 100K | 0.422 | 0.386 | 0.271 | 0.278 | 0.939 | 0.835 | 0.002 | 1.022 | 0.908 |

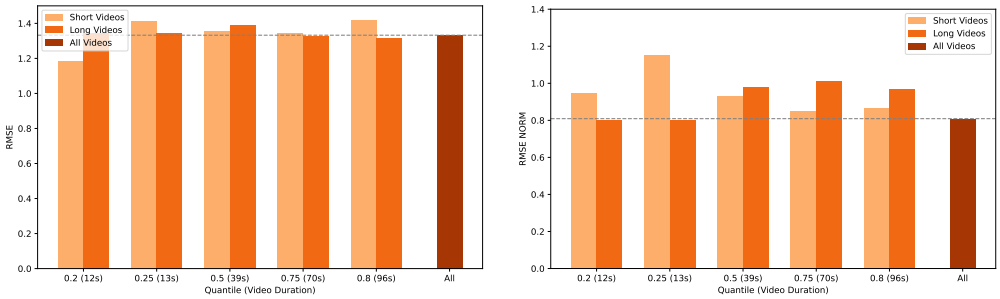
**Table 7.5:** Comparison of matrix factorization performance across datasets. $*$ Metrics computed using leave-one-out cross-validation (LOOCV), which provides an estimate of model performance for predicting unseen user-item interactions.

relatively elevated (0.809), suggesting that short-form video watching behavior is difficult to predict even under stable exposure conditions.

Accuracy, precision, recall, and F1 scores are consistently lower for the real-world TikTok data, reflecting the additional challenge of predicting watch behavior in personalized and heterogeneous feed environments with limited user–video overlap. Metrics calculated using leave-one-out cross-validation (LOOCV), marked with an asterisk (*), including RMSE, NRMSE, and NDCG@10, provide additional insight into generalization to unseen users. The NDCG@10 results show a particularly strong contrast: while the TikTok experiment dataset achieves high ranking performance (0.807), NDCG@10 drops sharply in both the real-world TikTok dataset (0.055) and MovieLens dataset (0.002). This decline is largely explained by the substantial increase in the number of candidate items. The MovieLens dataset contains over 1,600 movies, the real-world TikTok dataset includes more than 21,000 videos, and the experimental TikTok dataset contains only 105 videos. With a much larger item space, relevant items are less likely to appear among the top-10 ranked recommendations, which severely lowers NDCG. Limited overlap in user–item interactions further compounds this challenge, as the model has fewer shared observations from which to learn meaningful latent representations. Overall, these findings demonstrate that recommender models based on matrix factorization face greater prediction uncertainty when applied to short-form video platforms like TikTok compared to traditional recommendation domains such as movies, as exemplified by MovieLens 100K.

For matrix factorization in the TikTok experimental dataset, we further examined the role of video duration in the recommendation task. We partitioned the dataset into two subsets based on a duration threshold, separating shorter from longer videos, to analyze prediction errors across different video duration. Our hypothesis is that longer videos are more likely to be watched until the end by users with a strong interest in the content, whereas shorter videos attract a broader audience with more heterogeneous watching behavior patterns, making it harder to identify latent features and increasing the likelihood of errors in the recommendation task. This analysis allows us to assess whether latent user preferences differ between short and long videos and whether predicting the proportion of a video watched is more variable for short videos.

Figure 7.7 reports the RMSE and NRMSE for matrix factorization applied to subsets of the dataset split by video duration thresholds at 12, 13, 39, 70, and 96 seconds, which correspond to the 20th, 25th, 50th, 75th, and 80th percentiles of the duration distribution. We observe that RMSE decreases monotonically as video duration increases, indicating that the model makes




**(a)** RMSE **(b)** NRMSE

**Figure 7.7:** RMSE and NRMSE for videos below or above specific duration thresholds in the matrix factorization analysis applied to the TikTok experimental dataset.

smaller absolute prediction errors for longer videos. In contrast, NRMSE exhibits the opposite trend and increases with video duration. This occurs because watching behavior with longer videos shows lower variance: most users watch only a small proportion of long videos, leading to interaction values concentrated near one. Since NRMSE normalizes RMSE by the standard deviation of observed watching behavior, the reduced variability in the long-video subset inflates the normalized error. This also reflects a bias in the data, as relatively few users watch long videos until the end, which makes watching behavior with long videos appear less predictable in relative terms despite lower absolute error. At the same time, very short videos (below 13 seconds) exhibit high RMSE and NRMSE, indicating substantial unpredictability in user watching behavior. Short videos tend to attract a wide audience regardless of topical interest, resulting in highly variable watch behavior. This reinforces the challenge of modeling watching behavior for short-form content.

Overall, our findings indicate that modeling watching behavior with short-format videos is substantially more challenging than modeling preferences for traditional items such as movies. In the TikTok experimental dataset, despite its relatively high density, the NRMSE is comparable to that of the sparse MovieLens 100K dataset, highlighting considerable variability in watch behavior even under controlled conditions. Performance further deteriorates in the real-world TikTok dataset, where personalized feeds and limited user–video overlap introduce additional uncertainty. Examining recommendation errors by video duration in the experimental setting, we observe that RMSE decreases for longer videos, indicating more stable watching patterns, whereas NRMSE increases due to the reduced variance in long-video watching behavior. Very short videos consistently produce high errors across both RMSE and NRMSE, reflecting inconsistent and volatile watching behavior. Together, these results demonstrate that predicting user watching behavior with short-format videos is particularly challenging at the extremes of video duration, where watching behavior is either highly unstable for short videos or strongly biased toward low watch proportions for long videos.

## 7.5 Discussion

In this study, we investigated the predictability of whether users will watch a video until the end. Our research aimed to address two key questions: (i) can we predict, and which features most effectively predict, whether a user will watch a video until the end? and (ii) can we recommend videos that users are likely to watch? To answer these questions, we conducted an experiment via Zoom, which allowed for a controlled setting where participants interacted with a curated playlist of TikTok videos. This experiment enabled us to closely analyze the impact of video metadata and user demographics on TikTok user-watching behavior.

We applied classification models to assess the predictability of user-watching behavior, focusing on video metadata and user demographics as potential predictors. Our results show that video duration plays a central role in predicting whether users will watch a video until the end. We also compared the results obtained in the experimental setting with those in a real-world setting, confirming the critical role of video duration for watching behavior prediction.

Additionally, we leveraged matrix factorization techniques to assess whether we can recommend videos that a user would watch. Our results indicate that the error is higher for recommending short videos under 13 seconds. This increase in error arises from the high variability in the proportion of users watching short videos. Watching behavior for these videos appears nearly random, making accurate recommendations challenging. We compared the recommendation performance with real-world datasets based on TikTok and MovieLens 100K. Our findings indicate that recommending short-format videos results in substantially higher errors than modeling preferences for traditional items such as movies.

Our study began with the premise that TikTok's recommendation algorithm quickly adapts to users' preferences, producing a personalized feed that is widely perceived as being precisely curated for each individual. Yet, our findings reveal a notable gap in the predictability of user-watching behavior, especially for short videos, where recommendation errors are significantly higher. Longer videos exhibit smaller recommendation errors, suggesting that they may be better suited for personalization and reinforcing a sense of relevance for users. Another possible explanation for the gap between perceived and actual personalization could lie in TikTok's extensive use of user behavioral data, which we could not fully replicate in our experiment.

While video duration is a strong predictor of user-watching behavior, especially for short videos, the unpredictability of watching behavior on platforms like TikTok remains a significant challenge. This unpredictability highlights the evolving nature of short-form content and presents opportunities for improving recommendation systems and content creation strategies. Marketers, content creators, and educators need to consider video duration when designing content optimized for watching until the end. For example, creators should balance brevity with content that encourages curiosity or builds anticipation to improve completion rates. Longer videos tend to appeal to more dedicated viewers, providing opportunities for in-depth storytelling or niche content without sacrificing watching behavior. This allows creators to tailor content for broader,

fleeting audiences or for more loyal and engaged viewers. Platforms like TikTok could also use personalization to make educational content more accessible and engaging, broadening the social benefits of algorithmic curation.

The unpredictability of user watching behavior with short videos also has implications for content moderation. Platforms often rely on popularity metrics such as view counts, and shares, as well as user reports, to determine which content should be moderated first (Zannettou, 2021). Our findings suggest these metrics are less reliable for short videos due to erratic watching behavior. Harmful or inappropriate content in short videos might not generate the engagement signals necessary for timely moderation. Consequently, platforms need to incorporate additional factors beyond popularity metrics, such as AI-based content risk assessment or increased reliance on user reports. Overall, this highlights the importance of considering video format when prioritizing content moderation, particularly for short videos that may require intervention.

Finally, it is important to acknowledge the limitations of this study. First, the experimental setting involved a pre-defined playlist rather than personalized TikTok feeds, so participants' behavior may differ from typical usage. To mitigate this, we tested our model in a real-world setting and observed a similar pattern regarding video duration, reinforcing its predictive importance. Second, we only analyzed participants' watching behavior for 105 videos, which may not reflect the full diversity of TikTok content. Finally, while video duration is a robust predictor, our findings may not capture all aspects of TikTok's recommendation engine or proprietary features.

Future research could simulate TikTok's recommendation dynamics more closely, examining how different markers of user interaction or various content types impact perceived personalization. Additionally, investigating whether watching until the end is the most effective metric, or whether earlier viewing thresholds better capture personalization, could provide further insights into user satisfaction with short-form content.

## 7.6 Conclusion

This study examined the predictability of watching behavior in short-form video platforms through two complementary tasks: classification and recommendation. Using an experimental dataset and real-world validation, we showed that in the classification setting, video metadata, particularly video duration, is the strongest predictor of whether users watch a video until the end, while demographic attributes contribute marginally. In contrast, in the recommendation setting, the high variability in watching behavior for short videos makes it difficult to accurately recommend videos users are likely to watch. As a result, recommendation errors are substantially higher for short-form videos than for long-form content such as movies. These findings highlight a fundamental distinction: while short video engagement can be partially explained at the individual video level through classification, it remains difficult to model and predict in a user–item interaction framework typical of recommendation systems.

CHAPTER 8

# Discussion, Limitations & Future work

In this chapter, we discuss the work presented in the prior chapters, as well as their limitations, and highlight avenues for future work.

This thesis focuses on the use of digital trace data to advance research on migration, culture, inequalities, and online behavior in algorithmically mediated platforms. Each chapter demonstrates both the opportunities and challenges of working with such data in contexts that are often difficult to study with traditional sources.

Chapter 2 demonstrates the potential of digital trace data, specifically Facebook Advertising data on users' interests in food and drink, to measure cultural similarity between countries. The findings show that digital traces of everyday preferences capture meaningful aspects of culture, aligning with traditional survey-based measures such as the World Values Survey, while providing scalable, timely, and cost-effective alternatives. Building on this, Chapter 3 incorporates measures of cultural similarity into migration models. In particular, it demonstrates that cultural similarity derived from Facebook data has predictive power comparable to traditional covariates such as shared language and history.

Chapter 4 explores the use of digital traces in the context of information-seeking behavior during forced migration. Using data from Wikipedia Pageviews, the study shows that the increase in views on Wikipedia pages about cities correlates with migration flows during crises, as illustrated by the case of the Ukrainian refugees crises followed by the Russian invasion of Ukraine. The findings suggest that digital behavior may serve as a timely indicator of migration intentions, offering opportunities for policymakers and humanitarian agencies to anticipate population movements in crisis situations.

Chapters 5 and 6 focus on inequalities by analyzing gender balance and vulnerable populations. Chapter 5 presents a large-scale analysis of the global STEM gender gap using Facebook Ads data, revealing consistent cross-country patterns and validating results against the Global Gender Gap Report. A case study of Brazil illustrates how digital trace data can extend coverage and uncover demographic variation within a country. Chapter 6 examines missing children in Guatemala, drawing on data from Twitter's Alerta Alba-Keneth account. By extracting demographic and geographic information from digital alerts, the study provides the first systematic description of missing children in Guatemala, highlighting age and gender disparities as well as

spatial concentration in urban centers. Together, these chapters demonstrate how digital traces can complement scarce or delayed official statistics to shed light on persistent inequalities and vulnerable populations, particularly in the Global South, often overlooked in existing research.

Finally, Chapter 7 shifts the focus to data donation as a mode of data collection, exploring online behavior on algorithmically mediated platforms through the case of TikTok. This chapter provides a full methodology for collecting data via user donation and introduces a new dataset capturing user watching behavior on short-format video platforms. Using this experimental dataset, the study demonstrates that video metadata, particularly video duration, is the strongest predictor of whether a video is watched until the end, while demographic attributes contribute only marginally to predicting watching behavior. Short videos are especially difficult to predict due to high variability in watching behavior, whereas longer videos show more stable patterns but are biased toward low watch proportions. Matrix factorization analyses further show that recommending short videos is more error-prone than long videos and traditional items such as movies, highlighting the challenges of modeling user behavior in highly dynamic, short-form content environments. This chapter not only offers insights into the limits of algorithmic predictability on short-format video platforms but also emphasizes the potential of data donation as a complementary approach in response to increasingly restricted access to social media APIs, enabling researchers to study fine-grained user interactions with digital content.

Collectively, the four chapters underscore the value of digital trace data as an important complement to traditional data sources. By leveraging large-scale, passively collected behavioral data, this thesis provides insights into cultural diffusion, migration dynamics, inequality, and online engagement that would be difficult or costly to capture with conventional methods alone. The methodological innovations introduced, such as asymmetric measures of cultural similarity, the use of Wikipedia as a proxy for crisis-related information-seeking behavior, and data donation for user-behavior analysis, demonstrate how computational approaches can overcome both substantive and practical challenges in computational social science research.

Despite the contributions of this work, several limitations must be acknowledged. Each chapter discusses constraints specific to the data and methods applied. Here, however, we summarize the limitations that are common to most studies relying on digital trace data.

First, issues of representativity are inherent to digital trace data. Users of platforms such as Facebook, Wikipedia, TikTok, and Twitter do not perfectly reflect the underlying populations. Users may differ systematically from the general population in terms of gender, age, geographic location, or socio-economic status (Araujo et al., 2017; Gil-Clavel and Zagheni, 2019). Moreover, the use of online platforms depends heavily on internet penetration, which varies considerably across countries and may bias the observed behaviors and patterns.

Second, this work rely on self-reported information or platform-generated inferences. In the case of Facebook, for instance, some interests are explicitly declared by users, while others are inferred through black-box algorithms. Although demographic attributes such as age, gender, and location are generally accurate (Grow et al., 2022), the validity of inferred interests, such

as food and drink preferences or declared STEM-related interests, has not been systematically validated. Similarly, the report of missing children on Twitter may not accurately represent all actual cases, which limits the generalizability of the findings and should be considered when interpreting the results.

Third, because digital trace data are not originally designed for research purposes, repurposing them often requires methodological creativity and the use of proxies. For example, Facebook users' interests in food and drink are used as proxies for cultural attributes, while interests in STEM majors serve as proxies for gender balance in education and the labor market. On Wikipedia, the language edition is used as a proxy for the origin of searches, and articles about specific locations can serve as indicators of destination interests. While these proxies represent the best available options, they cannot fully capture the multidimensional nature of culture, subject-specific interests, or migration intentions. In particular, the use of language as a proxy for origin in Wikipedia data introduces potential misclassification, especially for languages spoken across multiple countries.

Fourth, data coverage and availability impose additional constraints. Several analyses were limited to specific countries or populations with sufficient digital activity, which restricts the ability to examine historical trends or capture the effects of major disruptions such as wars or pandemics. The Facebook Advertising Platform, for instance, only provides data available at the time of collection and does not allow retrospective analysis, whereas Wikipedia offers historical data on page views. However, even Wikipedia Pageview data are only available from 2015 onward, constraining the study of earlier events of interest.

Fifth, the studies are primarily cross-sectional and correlational, which limits causal inference. For instance, while cultural similarity measures correlate with migration patterns, establishing bidirectional or causal relationships would require longitudinal or experimental designs. Likewise, analyses of user behavior on TikTok identify strong predictors, such as video duration, but cannot fully disentangle causality from correlation.

Sixth, digital trace data are sensitive to platform-specific dynamics and evolving digital ecosystems. Traffic patterns on Wikipedia, for example, may be influenced by search engine rankings, recommendation algorithms, or AI-driven tools, which can change over time and affect the stability and generalizability of findings.

Finally, it is important to note that official data used for validation also present certain limitations when combined with or used to validate digital trace data. While traditional data sources remain indispensable for providing reliable, structured, and representative information, they often suffer from gaps, reporting delays, or limited temporal and spatial granularity. These constraints can restrict the extent to which they can be directly aligned with the real-time and high-frequency nature of digital trace data.

These limitations highlight that digital trace data are powerful but imperfect tools. Issues of representativity, measurement validity, and data coverage must be carefully considered. Nevertheless, even with these caveats, digital trace data provide scalable, timely, and complementary

insights to traditional data sources. Future work can mitigate these limitations by expanding geographic and temporal coverage, validating inferred attributes, integrating multi-platform data, adopting longitudinal designs, and developing methods to adjust for biases, improving the robustness and applicability of digital trace approaches in computational social science.

# Chapter 9

# Conclusion

In conclusion, this thesis makes methodological, substantive, and empirical contributions to computational social science. In particular, it advances research on migration, culture, inequalities, and online user behavior by demonstrating the potential of digital trace data. Methodologically, the contributions include measuring cultural similarity from online interests, analyzing online information-seeking during forced migration, and collecting engagement data through data donation. These innovations show how computational approaches can expand the toolkit of social scientists and help overcome the limitations of traditional data sources. Empirically and substantively, the thesis demonstrates that digital trace data can reveal meaningful cultural patterns, provide near real-time insights into migration flows during crises, and document gender balance in STEM fields as well as the demographics of vulnerable populations. It also highlights the use of data donation as a valuable method for studying user behavior on algorithmically mediated platforms. Overall, the findings highlight the potential of digital trace data to bridge social science and computational methods, offering new perspectives and practical tools for studying complex societal issues in a scalable, timely, and cost-effective way.

# Appendices

Appendix A

# Evaluating the Impact of Cultural Similarity on Migration Prediction

### Biases on Social Media Data

A line of research has focused on identifying the different types of errors and biases in studies that use digital trace data, and on organizing them in a framework (Olteanu et al., 2019; Sen et al., 2021). For instance, Sen et al. (2021) proposed a categorization based on the total survey error (TSE) framework to identify several types of errors that may occur in studies that use digital traces. As a consequence, these frameworks also contribute to creating a common vocabulary between researchers using digital trace data. In addition to that, Drouhot et al. (2023) provide an overview of how some innovative datasets and methodological tools can enrich migration research. Despite all the advantages and promises of using digital trace data for migration research (e.g., less time and costs needed to leverage data for a large sample size), the authors point out some of the challenges when working with these data. Since digital trace data are not generated for research purposes, they require some extra care to repurpose their meaning. Otherwise, they may be superficial and inappropriate for many central research questions. Besides the data quality, some of the key challenges involve ethical considerations and selection bias.

Zagheni and Weber (2015) considered the problem of selection bias in non-representative samples, such as digital trace data, and proposed two main approaches to reduce bias: calibration and the difference-in-differences approach. The idea of the calibration approach is based on adjusting the online data based on reliable official statistics, including the generation of correction factors (Ribeiro et al., 2020; Zagheni and Weber, 2012; Zagheni et al., 2017). For instance, Ribeiro et al. (2020) compared data from Facebook Ads and the USA Census and calculated correction factors for some demographic dimensions, such as age, gender, education, and income. However, for contexts where ground truth data are not available, the authors suggested a difference-in-differences approach to evaluate relative changes pre- and post-event (Alexander et al., 2019; Flores, 2017). Alexander et al. (2019) used Facebook Ads data and the difference-in-differences approach to monitor flows of outmigrants from Puerto Rico before and after Hurricane Maria in

2017. The difference-in-differences approach assumes a constant relationship between estimates from digital trace data and official data, at least over relatively short periods of time.

An emerging line of research focuses on Bayesian approaches to combine different sources of data to estimate migration trends (Alexander et al., 2020; Hsiao et al., 2023; Rampazzo et al., 2021). For instance, in a recent study focused on nowcasting stocks of migrants in the US, Alexander et al. (2020) demonstrated that a Bayesian hierarchical model combining data from both Facebook and the American Community Survey outperforms alternative models using only Facebook data or solely relying on time series data from the American Community Survey. Recently, Leasure et al. (2023) built a real-time monitoring system to estimate subnational population sizes and internal displacement in Ukraine by leveraging data from Facebook Ads in combination with pre-conflict population data in Ukraine.

## Correlations between variables and Gravity Model

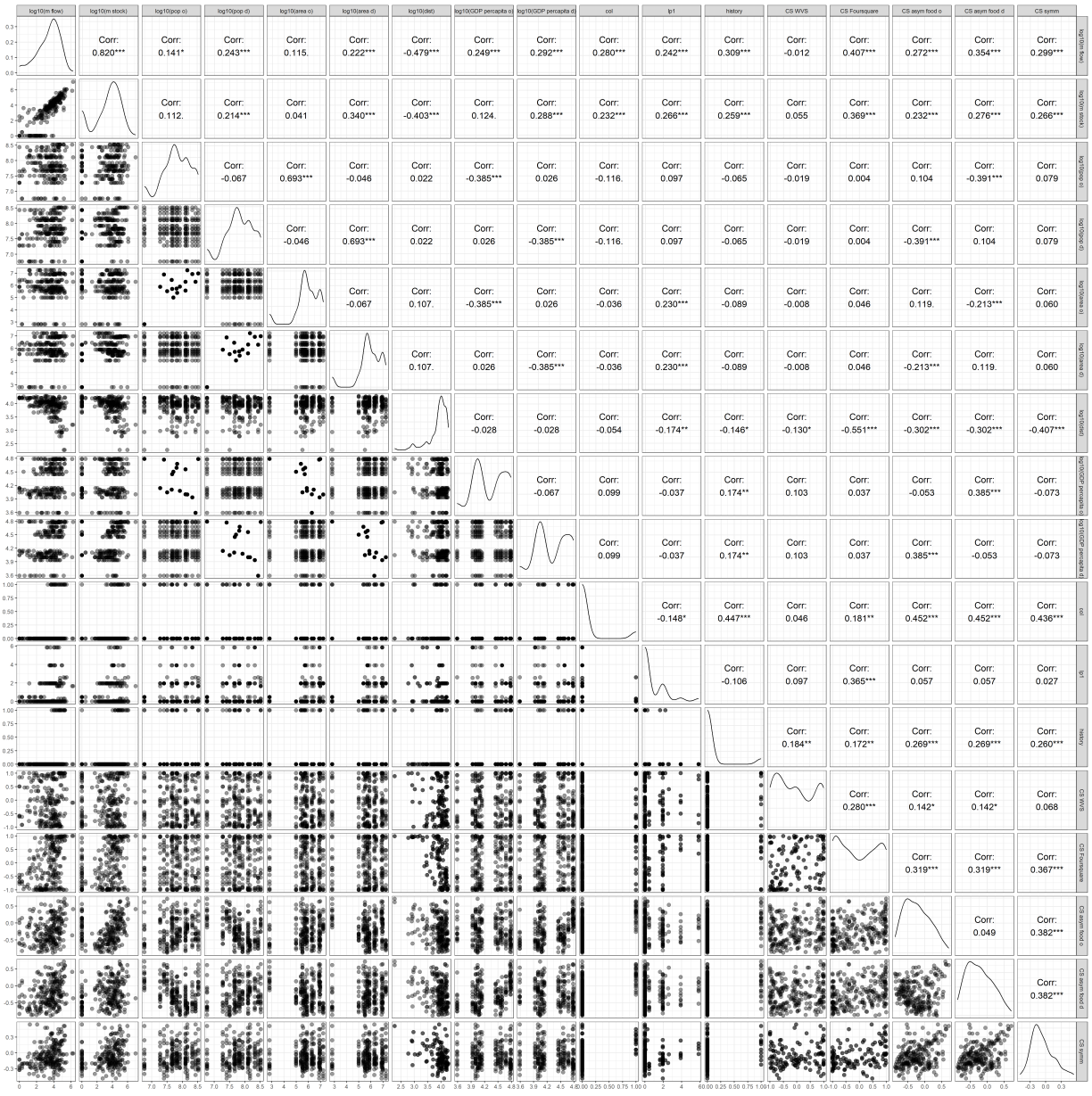



**Figure A.1:** Distribution and correlations between all the variables in our dataset. Each dot represents a pair of countries within the 16 countries we analyzed.
$***p < 0.001$; $**p < 0.01$; $*p < 0.05$

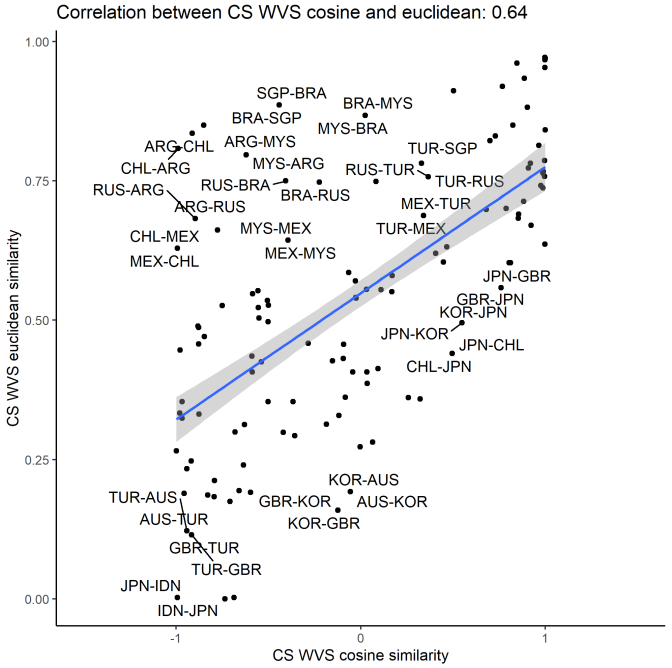



**Figure A.2:** Distribution and correlations between the measure of cultural similarity derived from WVS data calculated using the cosine and Euclidean distance. Each dot represents a pair of countries within the 16 countries we analyzed.

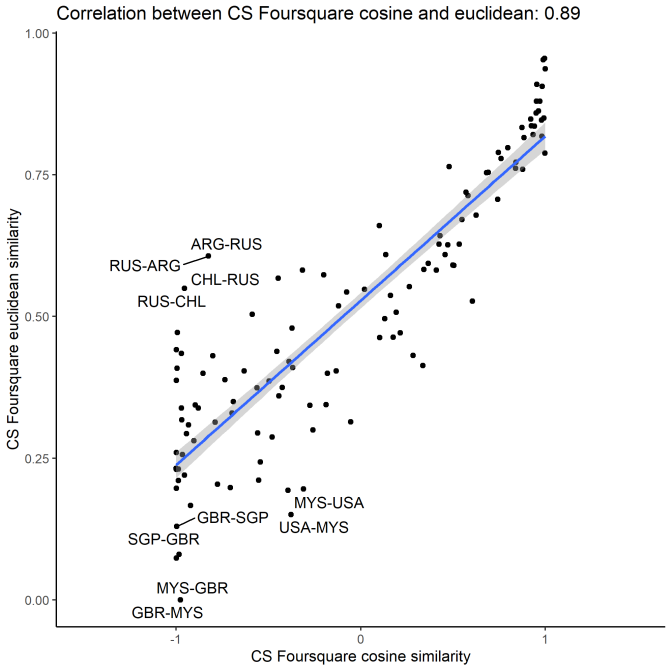



**Figure A.3:** Distribution and correlations between the measure of cultural similarity derived from Foursquare data calculated using the cosine and Euclidean distance. Each dot represents a pair of countries within the 16 countries we analyzed.

| | Model 1 | Model 1 + FB asymmetric | Model 1 + FB symmetric | Model 1 + Foursquare CS | Model 2 | Model 2 + FB asymmetric | Model 2 + FB symmetric | Model 2 + Foursquare CS | Model 3 | Model 3 + Facebook asymmetric | Model 3 + Facebook symmetric | Model 3 + Foursquare CS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -10.48*** | -11.95*** | -10.92*** | -10.88*** | -12.12*** | -12.30*** | -12.04*** | -12.17*** | -12.90*** | -13.23*** | -12.83*** | -13.31*** |
| | (1.71) | (1.71) | (1.72) | (1.74) | (1.67) | (1.69) | (1.68) | (1.69) | (1.58) | (1.60) | (1.59) | (1.60) |
| log10_pop_o | 0.23 | 0.39** | 0.23 | 0.25* | 0.35** | 0.40** | 0.36** | 0.35** | 0.37** | 0.43*** | 0.38** | 0.38** |
| | (0.12) | (0.13) | (0.12) | (0.12) | (0.12) | (0.13) | (0.12) | (0.12) | (0.11) | (0.12) | (0.12) | (0.11) |
| log10_area_o | 0.23*** | 0.19*** | 0.23*** | 0.22*** | 0.20*** | 0.20*** | 0.21*** | 0.20*** | 0.21*** | 0.20*** | 0.21*** | 0.21*** |
| | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.06) | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) | (0.05) |
| log10_pop_d | 0.64*** | 0.75*** | 0.64*** | 0.66*** | 0.74*** | 0.73*** | 0.75*** | 0.74*** | 0.74*** | 0.75*** | 0.75*** | 0.75*** |
| | (0.12) | (0.13) | (0.12) | (0.12) | (0.12) | (0.13) | (0.12) | (0.12) | (0.11) | (0.12) | (0.11) | (0.11) |
| log10_area_d | 0.02 | -0.01 | 0.02 | 0.01 | 0.03 | 0.03 | 0.04 | 0.03 | 0.07 | 0.06 | 0.07 | 0.06 |
| | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) | (0.06) |
| log10_distwces | -1.00*** | -0.82*** | -0.93*** | -0.92*** | -1.04*** | -1.01*** | -1.08*** | -1.03*** | -1.15*** | -1.08*** | -1.19*** | -1.06*** |
| | (0.13) | (0.14) | (0.14) | (0.15) | (0.13) | (0.15) | (0.14) | (0.14) | (0.12) | (0.14) | (0.13) | (0.14) |
| log10_m_stock | 0.40*** | 0.38*** | 0.39*** | 0.40*** | 0.34*** | 0.34*** | 0.34*** | 0.34*** | 0.32*** | 0.32*** | 0.32*** | 0.31*** |
| | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) |
| log10_GDP_percapita_o | 1.03*** | 0.93*** | 1.05*** | 1.03*** | 1.02*** | 0.99*** | 1.02*** | 1.02*** | 1.10*** | 1.05*** | 1.10*** | 1.10*** |
| | (0.12) | (0.13) | (0.12) | (0.12) | (0.12) | (0.13) | (0.12) | (0.12) | (0.11) | (0.12) | (0.11) | (0.11) |
| log10_GDP_percapita_d | 0.84*** | 0.80*** | 0.87*** | 0.84*** | 0.91*** | 0.92*** | 0.91*** | 0.91*** | 1.03*** | 1.03*** | 1.03*** | 1.04*** |
| | (0.15) | (0.15) | (0.15) | (0.15) | (0.14) | (0.15) | (0.14) | (0.14) | (0.13) | (0.14) | (0.13) | (0.13) |
| CS_nonsymm_food_o | | 0.26* | | | | -0.01 | | | | 0.05 | | |
| | | (0.12) | | | | (0.14) | | | | (0.13) | | |
| CS_nonsymm_food_d | | 0.38** | | | | 0.13 | | | | 0.21 | | |
| | | (0.12) | | | | (0.14) | | | | (0.13) | | |
| CS_symm | | | 0.35 | | | | -0.15 | | | | -0.14 | |
| | | | (0.20) | | | | (0.22) | | | | (0.21) | |
| CS_foursquare | | | | 0.08 | | | | 0.01 | | | | 0.10 |
| | | | | (0.07) | | | | (0.07) | | | | (0.06) |
| col | | | | | 0.60*** | 0.54** | 0.64*** | 0.60*** | 0.60*** | 0.48** | 0.64*** | 0.57*** |
| | | | | | (0.14) | (0.18) | (0.15) | (0.14) | (0.13) | (0.17) | (0.15) | (0.13) |
| lp1 | | | | | 0.03 | 0.03 | 0.03 | 0.03 | 0.05 | 0.04 | 0.05 | 0.03 |
| | | | | | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) | (0.03) |
| shared_hist | | | | | 0.16 | 0.16 | 0.17 | 0.16 | 0.29* | 0.29* | 0.30* | 0.28* |
| | | | | | (0.15) | (0.15) | (0.15) | (0.15) | (0.14) | (0.14) | (0.14) | (0.14) |
| CS_wvs | | | | | | | | | -0.29*** | -0.30*** | -0.29*** | -0.31*** |
| | | | | | | | | | (0.05) | (0.05) | (0.05) | (0.05) |
| $R^2$ | 0.80 | 0.81 | 0.80 | 0.80 | 0.82 | 0.82 | 0.82 | 0.82 | 0.84 | 0.84 | 0.84 | 0.84 |
| Adj. $R^2$ | 0.79 | 0.80 | 0.80 | 0.79 | 0.81 | 0.81 | 0.81 | 0.81 | 0.83 | 0.84 | 0.83 | 0.84 |
| Num. obs. | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 | 240 |

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

**Table A.1:** Linear models using the full input dataset (240 pairs of countries).

APPENDIX B

# Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

## Appendix B. Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

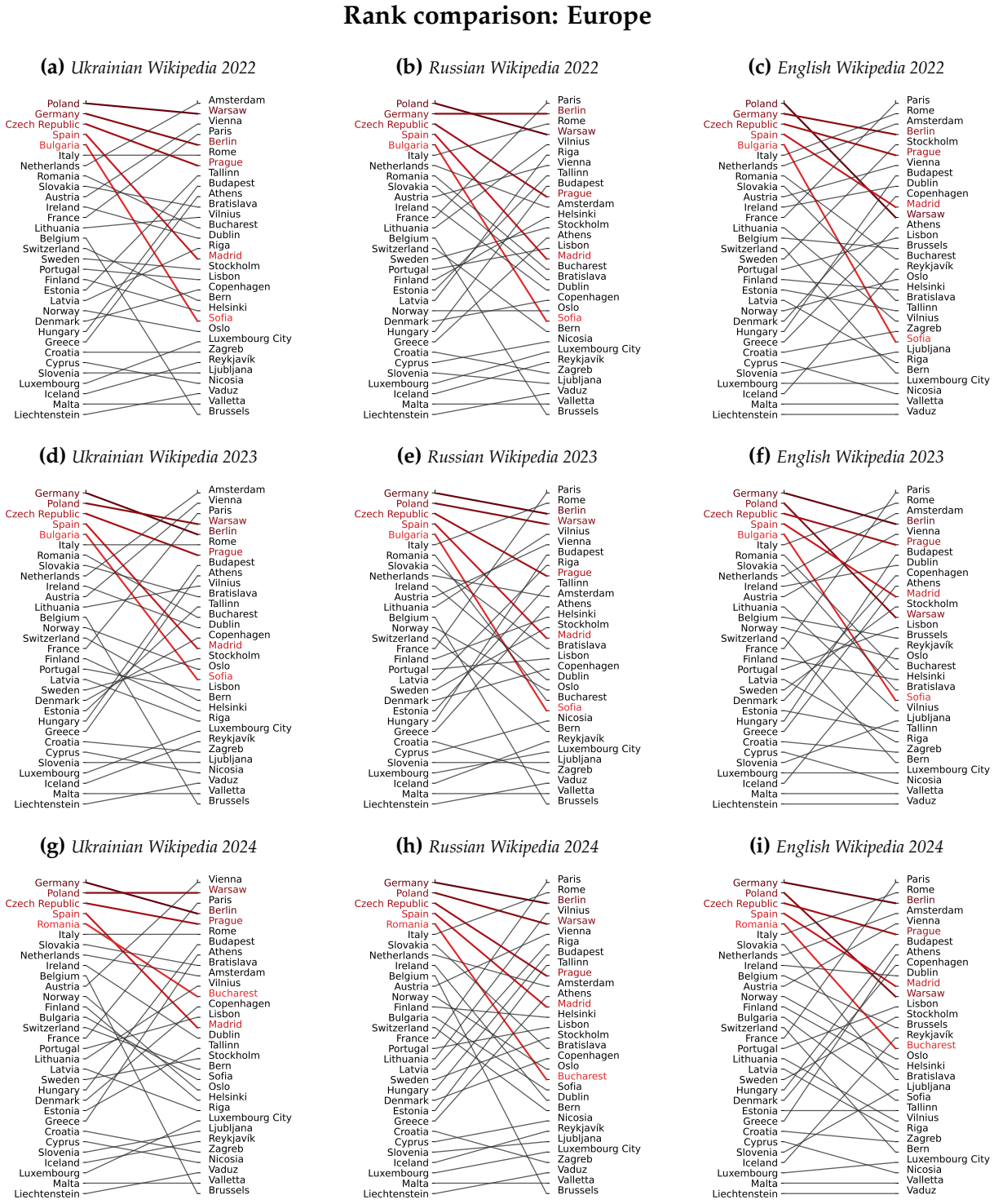



**Figure B.1:** Correlation between rankings: stocks of Ukrainian refugees in EU countries (left) and the proportion of views of Wikipedia articles about EU capitals (right), by year. The five countries hosting the largest numbers of Ukrainian refugees are shown in color, while the remaining countries are shown in gray.

## Appendix B. Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

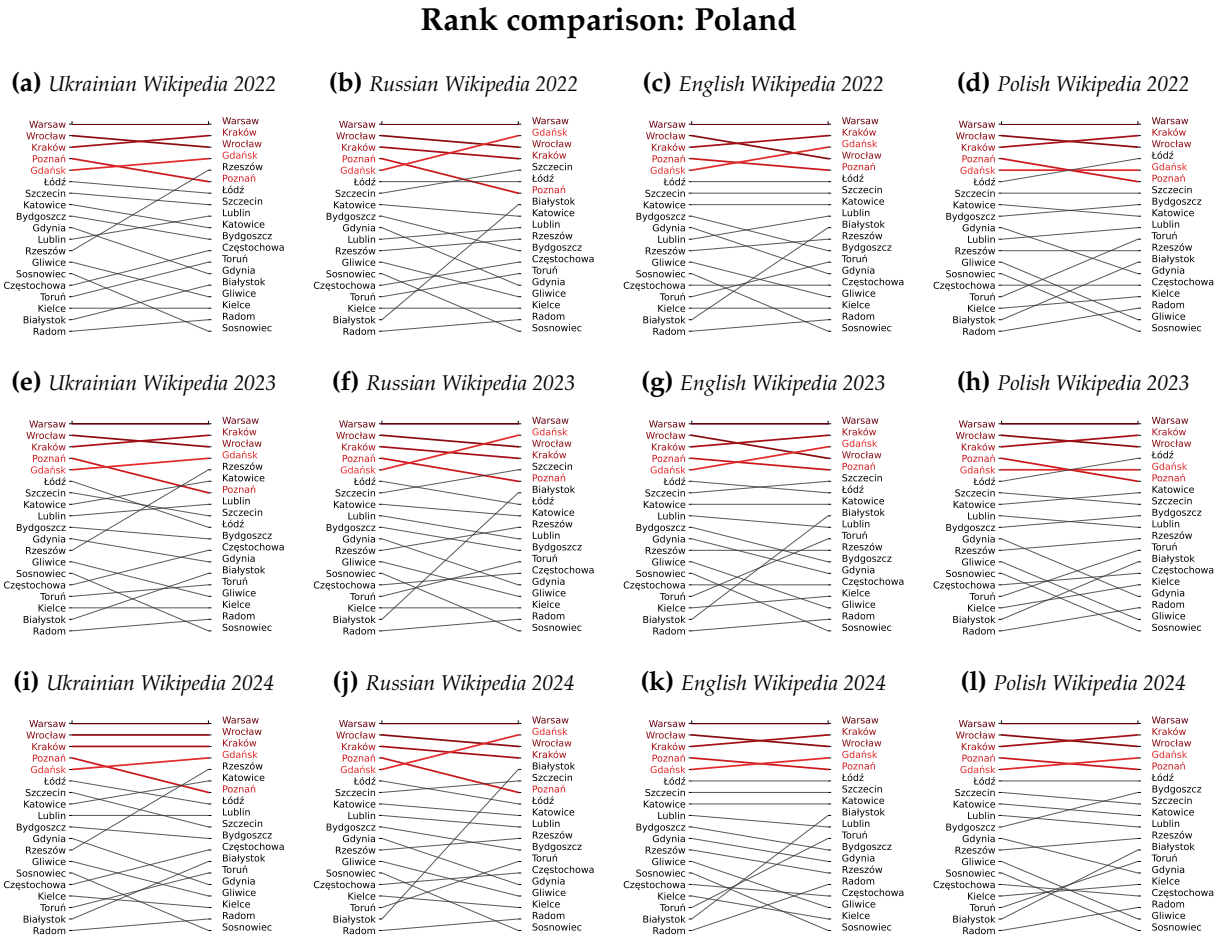



**Figure B.2:** Correlation between rankings: stocks of Ukrainian refugees who have been assigned a PESEL number in Polish cities (left) and the proportion of views of Wikipedia articles about the 19 most populous cities in Poland (right), by year. The five cities hosting the largest numbers of PESEL-registered Ukrainian refugees are shown in color, while the remaining cities are shown in gray.

## Appendix B. Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

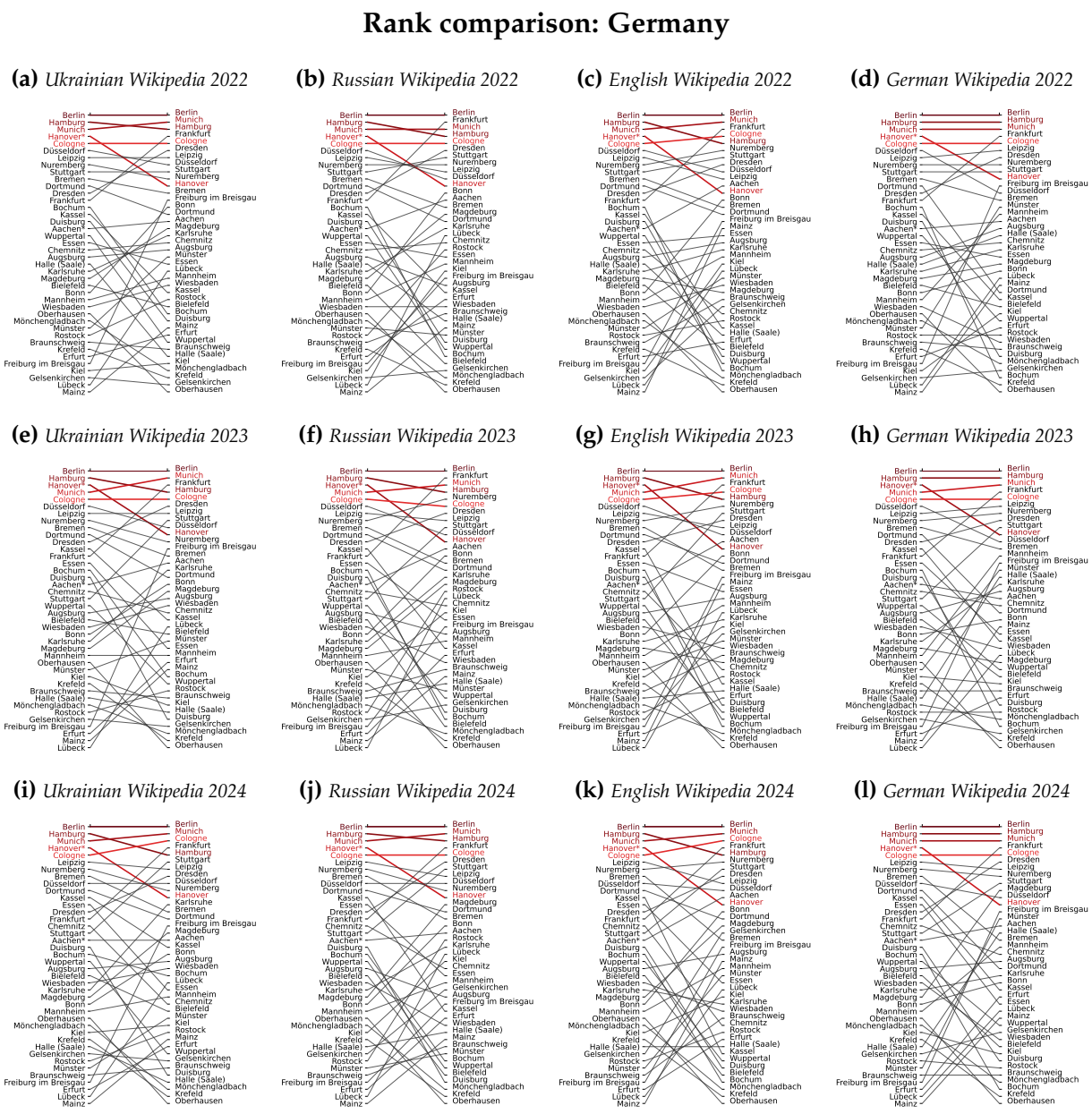



**Figure B.3:** Correlation between rankings: stocks of Ukrainian refugees with temporary protection status in German cities (left) and the proportion of views of Wikipedia articles about the 40 most populous German cities (right), by year. The five cities hosting the largest numbers of Ukrainian refugees with temporary protection are shown in color, while the remaining cities are shown in gray. For Hanover and Aachen, data on Ukrainians under temporary protection are available only at the city-regional level (*Städteregion*), rather than at the independent city level (*kreisfreie Stadt*).


**Figure B.4:** Relative change in the proportion of weekly views, compared to the same period in the previous year, of Wikipedia articles about the 19 most populous Polish cities across four languages (English, Polish, Russian, and Ukrainian) from August 24, 2020, to August 24, 2023.

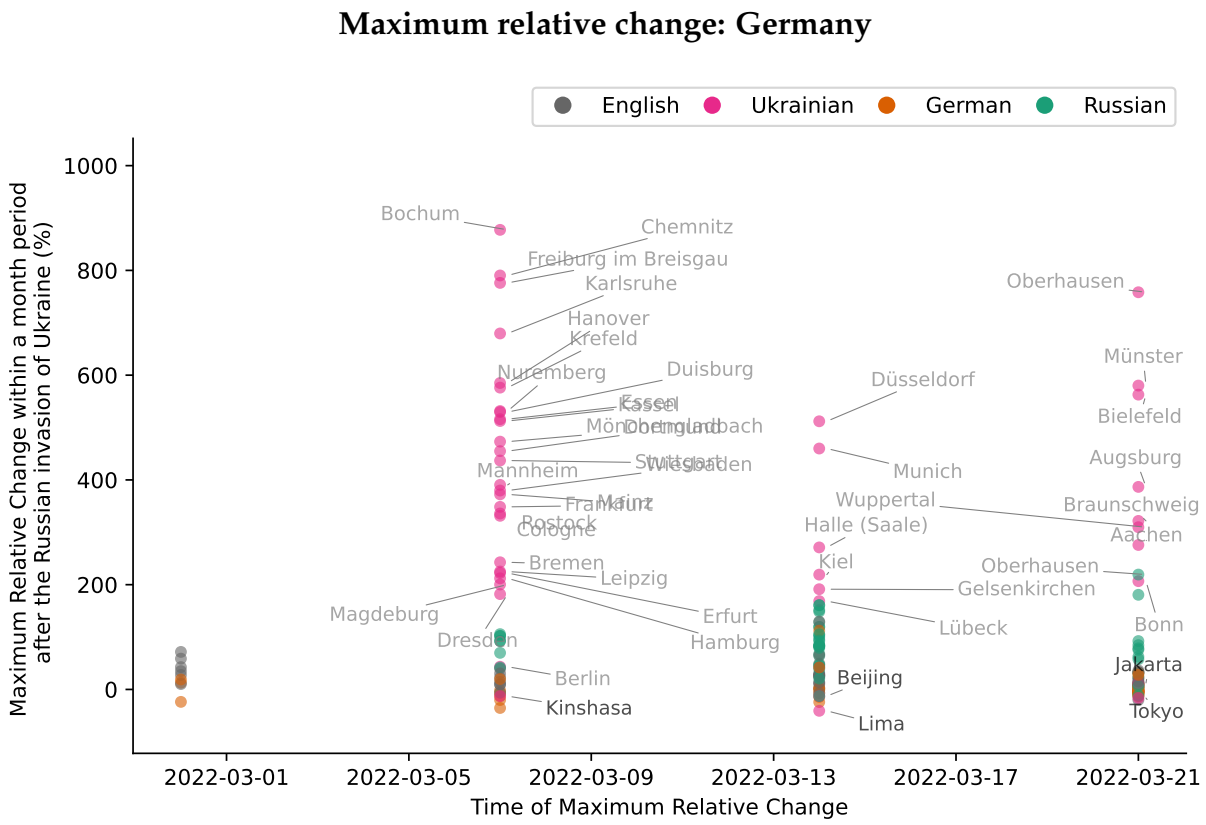



**Figure B.5:** Maximum relative change in the proportion of weekly views over the month following the Russian invasion of Ukraine, compared to the same period in the previous year. Results are shown for Wikipedia articles about the 40 most populous German cities and five of the most populous cities in the world (Beijing, Jakarta, Kinshasa, Lima, and Tokyo) across four languages (English, German, Russian, and Ukrainian).

**Structural breaks: Poland**

| City | Break Date | CI lower | CI upper |
|---|---|---|---|
| Bydgoszcz | **2022-02-24** | 2022-02-14 | 2022-03-10 |
| | 2022-08-31 | 2022-08-16 | 2022-09-06 |
| Częstochowa | **2022-03-01** | 2022-02-07 | 2022-03-08 |
| | 2022-08-28 | 2022-08-22 | 2022-09-16 |
| Gdańsk | **2022-03-01** | 2022-02-11 | 2022-03-06 |
| | 2022-08-23 | 2022-08-15 | 2022-09-03 |
| Gdynia | **2022-03-01** | 2022-02-20 | 2022-03-07 |
| | 2022-09-04 | 2022-08-22 | 2022-09-08 |
| Gliwice | **2022-02-16** | 2022-02-07 | 2022-03-05 |
| | 2022-08-31 | 2022-08-18 | 2022-09-12 |
| Katowice | **2022-02-27** | 2022-02-22 | 2022-03-12 |
| | 2022-09-04 | 2022-08-07 | 2022-09-15 |
| Kielce | **2022-02-26** | 2022-02-20 | 2022-03-17 |
| | 2022-09-05 | 2022-08-17 | 2022-09-17 |
| Kraków | **2022-02-25** | 2022-02-20 | 2022-03-02 |
| | 2022-09-04 | 2022-08-10 | 2022-09-17 |
| Łódź | 2022-06-11 | 2022-06-01 | 2022-07-12 |
| | 2022-11-22 | 2022-11-02 | 2023-01-21 |
| Lublin | 2022-06-03 | 2022-04-26 | 2022-06-04 |
| | 2022-11-19 | 2022-11-18 | 2023-01-15 |
| Poznań | **2022-02-27** | 2022-02-16 | 2022-03-29 |
| | 2022-09-04 | 2022-07-09 | 2022-10-31 |
| Radom | **2022-02-25** | 2022-02-21 | 2022-02-28 |
| | 2022-08-31 | 2022-08-18 | 2022-09-11 |
| Kraków | **2022-02-25** | 2022-02-20 | 2022-03-02 |
| | 2022-09-04 | 2022-08-10 | 2022-09-17 |
| Rzeszów | **2022-03-14** | NA | NA |
| | 2022-09-04 | 2022-08-22 | 2022-10-29 |
| Sosnowiec | **2022-03-02** | 2022-02-22 | 2022-03-09 |
| | 2022-09-18 | 2022-09-06 | 2022-10-02 |
| Szczecin | **2022-02-25** | 2022-02-12 | 2022-03-12 |
| | 2022-09-04 | 2022-08-21 | 2022-09-12 |
| Toruń | **2022-03-01** | 2022-02-21 | 2022-03-06 |
| | 2022-08-31 | 2022-08-20 | 2022-09-11 |
| Warsaw | **2022-02-28** | 2022-02-23 | 2022-03-03 |
| | 2022-09-04 | 2022-08-12 | 2022-10-01 |
| Wrocław | **2022-02-27** | 2022-02-09 | 2022-05-17 |

**Table B.1:** Results of the structural break analysis using an autoregressive model (AR(1)) on the time series of the proportion of daily views of Ukrainian-language Wikipedia articles about the 19 most populous Polish cities. Only break points detected in 2022 are reported. For each city, the table reports the estimated break date in the second column, with the third and fourth columns indicating the lower and upper bounds of the corresponding confidence interval. Structural breaks identified within one month before or after the start of the Russian invasion of Ukraine (February 24, 2022) are shown in bold. Confidence intervals for break points are calculated by examining how the model fit improves when the relevant break point is shifted. For Rzeszów, the time series before the first estimated break point is highly monotonous, resulting in a distribution of the estimated break point with excessive probability at the boundary. When statistically meaningful confidence intervals cannot be computed, they are reported as NA.

**Structural breaks: World**

| City | Break Date | CI lower | CI upper |
|---|---|---|---|
| Beijing | 2021-08-25 | 2021-08-05 | 2021-09-13 |
| | **2022-02-05** | 2022-01-15 | 2022-04-17 |
| | 2022-08-02 | 2022-05-03 | 2022-09-11 |
| Jakarta | 2022-01-14 | 2021-10-04 | 2022-01-23 |
| | 2022-06-29 | 2022-06-27 | 2022-10-04 |
| Kinshasa | 2022-11-04 | 2022-10-12 | 2022-12-04 |
| Lima | NA | NA | NA |
| Tokyo | 2021-02-13 | 2021-01-24 | 2021-03-11 |
| | 2021-07-28 | 2021-07-14 | 2021-08-09 |
| | 2022-12-30 | 2022-12-09 | 2023-02-03 |

**Table B.2:** Sensitivity checks in the structural break analysis. We conducted a structural break analysis using an autoregressive model (AR(1)) on the time series of the proportion of daily views of Ukrainian-language Wikipedia articles corresponding to five of the most populous capitals in the world. For each city, the table reports the estimated break date in the second column, with the third and fourth columns indicating the lower and upper bounds of the corresponding confidence interval. Structural breaks occurring within one month before or after the start of the Russian invasion of Ukraine (February 24, 2022) are shown in bold. A structural break was detected only for Beijing, about three weeks before the invasion (February 5, 2022). However, this break was followed by a decline rather than an increase in the proportion of Ukrainian-language views. For Lima, no statistically meaningful structural breaks were detected within the observed period, and the results are reported as NA.

## Structural breaks: Germany

| City | Break Date | CI lower | CI upper |
|---|---|---|---|
| Aachen | 2022-05-20 | 2022-02-25 | 2022-05-21 |
| Augsburg | **2022-03-02** | 2022-02-25 | 2022-03-04 |
| | 2022-08-29 | 2022-07-13 | 2022-09-16 |
| Berlin | **2022-03-12** | 2022-01-29 | 2022-04-02 |
| Bielefeld | **2022-03-02** | 2022-02-26 | 2022-03-05 |
| | 2022-09-01 | 2022-08-23 | 2022-09-12 |
| Bochum | **2022-03-03** | 2022-02-25 | 2022-03-04 |
| | 2022-09-09 | 2022-08-06 | 2022-10-25 |
| Bonn | **2022-03-05** | 2022-02-28 | 2022-03-07 |
| | 2022-09-04 | 2022-08-31 | 2022-09-12 |
| Braunschweig | **2022-03-03** | 2022-02-27 | 2022-03-06 |
| | 2022-09-05 | 2022-08-27 | 2022-09-13 |
| Bremen | **2022-03-02** | 2022-02-27 | 2022-03-03 |
| | 2022-08-31 | 2022-08-27 | 2022-09-06 |
| Chemnitz | **2022-03-04** | 2022-02-07 | 2022-03-14 |
| | 2022-09-04 | 2022-08-29 | 2022-09-16 |
| Cologne | **2022-03-02** | 2022-02-05 | 2022-03-12 |
| | 2022-09-04 | 2022-08-25 | 2022-09-16 |
| Dortmund | **2022-03-03** | 2022-02-22 | 2022-03-13 |
| | 2022-09-04 | 2022-08-23 | 2022-09-15 |
| Dresden | **2022-02-28** | 2022-02-24 | 2022-03-04 |
| | 2022-09-06 | 2022-08-14 | 2022-09-21 |
| Duisburg | **2022-03-05** | 2022-02-27 | 2022-03-07 |
| | 2022-10-30 | 2022-10-04 | 2022-11-11 |
| Düsseldorf | **2022-03-03** | 2022-02-13 | 2022-03-10 |
| | 2022-09-01 | 2022-08-15 | 2022-09-21 |
| Erfurt | **2022-03-02** | 2022-02-26 | 2022-03-07 |
| | 2022-09-01 | 2022-08-19 | 2022-09-11 |
| Essen | **2022-03-02** | 2022-02-26 | 2022-03-04 |
| | 2022-09-04 | 2022-08-25 | 2022-09-16 |
| Frankfurt | **2022-03-20** | 2022-02-25 | 2022-03-21 |
| | 2022-09-01 | 2022-08-30 | 2022-10-11 |
| Freiburg | **2022-02-28** | 2022-02-23 | 2022-03-01 |
| im Breisgau | 2022-09-04 | 2022-08-25 | 2022-09-25 |
| Gelsenkirchen | **2022-03-08** | 2022-02-20 | 2022-03-15 |
| Halle | **2022-03-05** | 2022-03-01 | 2022-03-08 |
| | 2022-09-01 | 2022-08-23 | 2022-09-30 |
| Hamburg | **2022-03-02** | 2022-02-26 | 2022-03-04 |
| | 2022-09-04 | 2022-08-25 | 2022-09-21 |
| Hanover | **2022-03-01** | 2022-02-25 | 2022-03-06 |
| | 2022-08-29 | 2022-08-03 | 2022-09-26 |
| Karlsruhe | **2022-03-02** | 2022-02-25 | 2022-03-06 |
| | 2022-09-04 | 2022-08-24 | 2022-09-11 |
| Kassel | **2022-03-03** | 2022-02-25 | 2022-03-05 |
| | 2022-09-04 | 2022-08-31 | 2022-09-13 |
| Kiel | **2022-03-03** | 2022-02-22 | 2022-03-10 |
| | 2022-09-04 | 2022-08-31 | 2022-09-17 |
| Krefeld | **2022-02-23** | 2022-02-13 | 2022-02-26 |
| | 2022-09-18 | 2022-09-04 | 2022-10-17 |
| Leipzig | **2022-03-01** | 2022-02-03 | 2022-03-05 |
| | 2022-09-07 | 2022-08-31 | 2022-09-24 |
| Lübeck | **2022-03-09** | 2022-03-02 | 2022-03-13 |
| | 2022-09-06 | 2022-09-01 | 2022-09-15 |
| ⋮ | ⋮ | ⋮ | ⋮ |

| City | Break Date | CI lower | CI upper |
|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ |
| Magdeburg | **2022-03-03** | 2022-02-25 | 2022-03-05 |
| | 2022-09-04 | 2022-08-28 | 2022-09-20 |
| Mainz | **2022-03-03** | 2022-02-27 | 2022-03-06 |
| | 2022-09-03 | 2022-08-26 | 2022-09-18 |
| Mannheim | **2022-03-05** | 2022-02-27 | 2022-03-07 |
| | 2022-09-04 | 2022-08-30 | 2022-09-18 |
| Mönchengladbach | **2022-03-04** | 2022-01-22 | 2022-09-18 |
| Munich | **2022-02-28** | 2022-02-18 | 2022-03-15 |
| | 2022-08-28 | 2022-08-07 | 2022-09-08 |
| Münster | **2022-03-02** | 2022-02-25 | 2022-03-04 |
| | 2022-08-31 | 2022-08-17 | 2022-09-16 |
| Nuremberg | **2022-02-28** | 2022-02-24 | 2022-03-06 |
| | 2022-09-04 | 2022-08-27 | 2022-09-13 |
| Oberhausen | **2022-03-03** | 2022-02-25 | 2022-03-06 |
| | 2022-09-07 | 2022-05-28 | 2022-12-09 |
| Rostock | **2022-03-02** | 2022-02-21 | 2022-03-14 |
| | 2022-09-03 | 2022-08-20 | 2022-09-22 |
| Stuttgart | **2022-03-02** | 2022-02-27 | 2022-03-04 |
| | 2022-09-03 | 2022-08-28 | 2022-09-10 |
| Wiesbaden | **2022-03-06** | 2022-03-02 | 2022-03-09 |
| | 2022-09-04 | 2022-08-21 | 2022-09-20 |
| Wuppertal | **2022-03-01** | 2022-02-21 | 2022-03-03 |
| | 2022-09-11 | 2022-08-20 | 2022-10-06 |

**Table B.3:** Results of the structural break analysis using an autoregressive model (AR(1)) on the time series of the proportion of daily views of Ukrainian-language Wikipedia articles about the 40 most populous German cities. Only break points detected in 2022 are reported. For each city, the table reports the estimated break date in the second column, with the third and fourth columns indicating the lower and upper bounds of the corresponding confidence interval. When a structural break occurred within one month before or after the start of the Russian invasion of Ukraine (February 24, 2022), the date is shown in bold.

## Appendix B. Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

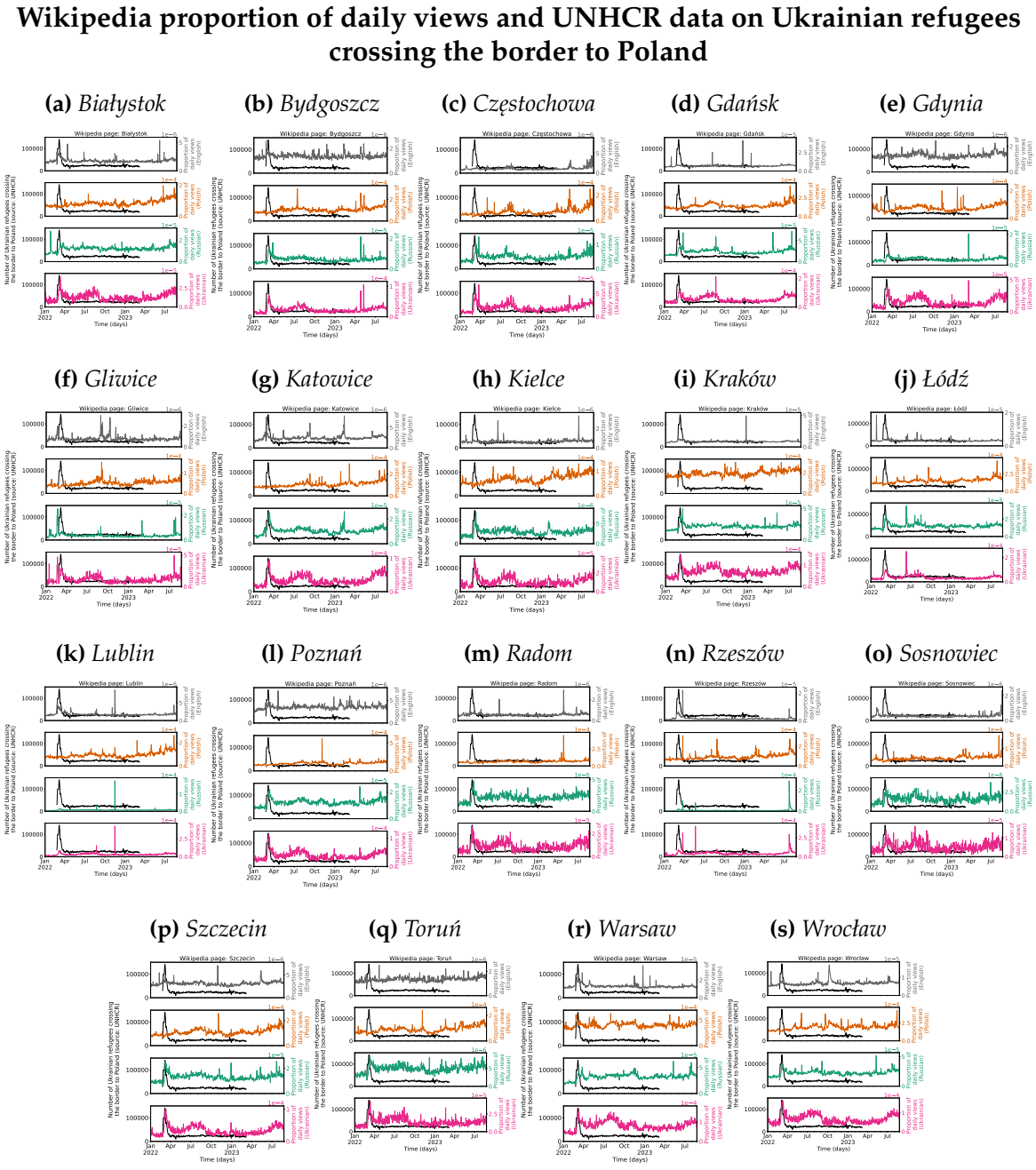



**Figure B.6:** Time series of the daily number of Ukrainian refugees crossing the border from Ukraine to Poland (from February 24, 2022 to March 3, 2023) and the proportion of daily views of Wikipedia articles about the 19 most populous Polish cities in four languages (English, Polish, Russian, and Ukrainian).

## Granger causality

| City | Relationship | Optimal lag | F-statistic (p-value) |
|---|---|---|---|
| Białystok | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in English | 8 | 1.98 (p = 0.0484) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 8.45 (p = 0.0000) |
| | Wikipedia views in English → Ukrainian refugees in Poland (UNHCR) | 8 | 2.83 (p = 0.0046) |
| Bydgoszcz | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in English | 8 | 2.00 (p = 0.0462) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 10.31 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 10.37 (p = 0.0000) |
| Częstochowa | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 5.79 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 5.49 (p = 0.0000) |
| Gdańsk | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in English | 8 | 2.03 (p = 0.0425) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 23 | 3.97 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 9.11 (p = 0.0000) |
| | Wikipedia views in English → Ukrainian refugees in Poland (UNHCR) | 8 | 6.06 (p = 0.0000) |
| | Wikipedia views in Russian → Ukrainian refugees in Poland (UNHCR) | 23 | 1.66 (p = 0.0317) |
| | Wikipedia views in Ukrainian → Ukrainian refugees in Poland (UNHCR) | 8 | 2.48 (p = 0.0125) |
| Gdynia | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 3.09 (p = 0.0022) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 6.51 (p = 0.0000) |
| Gliwice | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 2.73 (p = 0.0062) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 4.32 (p = 0.0001) |
| Katowice | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 5.99 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 8.73 (p = 0.0000) |
| | Wikipedia views in English → Ukrainian refugees in Poland (UNHCR) | 8 | 2.56 (p = 0.0101) |
| Kielce | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 4.19 (p = 0.0001) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 4.80 (p = 0.0000) |
| Kraków | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in English | 9 | 4.43 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 8.01 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 8.50 (p = 0.0000) |
| | Wikipedia views in English → Ukrainian refugees in Poland (UNHCR) | 9 | 9.67 (p = 0.0000) |
| Poznań | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 10.24 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 10.93 (p = 0.0000) |
| | Wikipedia views in English → Ukrainian refugees in Poland (UNHCR) | 8 | 2.65 (p = 0.0078) |
| Radom | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 5.67 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 15 | 2.48 (p = 0.0018) |
| | Wikipedia views in Ukrainian → Ukrainian refugees in Poland (UNHCR) | 15 | 1.98 (p = 0.0160) |
| Rzeszów | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in English | 8 | 3.00 (p = 0.0029) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 2.44 (p = 0.0139) |
| Sosnowiec | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 3.53 (p = 0.0006) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 2.86 (p = 0.0043) |
| Szczecin | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 10 | 6.70 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 10.82 (p = 0.0000) |
| | Wikipedia views in Russian → Ukrainian refugees in Poland (UNHCR) | 10 | 5.32 (p = 0.0000) |
| Toruń | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 6.21 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 5.25 (p = 0.0000) |
| | Wikipedia views in Ukrainian → Ukrainian refugees in Poland (UNHCR) | 8 | 2.90 (p = 0.0039) |
| Warsaw | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in English | 17 | 2.94 (p = 0.0001) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Polish | 8 | 2.54 (p = 0.0107) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 23 | 2.06 (p = 0.0035) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 6.60 (p = 0.0000) |
| | Wikipedia views in English → Ukrainian refugees in Poland (UNHCR) | 17 | 3.81 (p = 0.0000) |
| | Wikipedia views in Russian → Ukrainian refugees in Poland (UNHCR) | 23 | 1.59 (p = 0.0452) |
| Wrocław | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 5.94 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 17 | 2.83 (p = 0.0002) |
| | Wikipedia views in Ukrainian → Ukrainian refugees in Poland (UNHCR) | 17 | 3.00 (p = 0.0001) |
| Łódź | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Russian | 8 | 4.51 (p = 0.0000) |
| | Ukrainian refugees in Poland (UNHCR) → Wikipedia views in Ukrainian | 8 | 2.71 (p = 0.0065) |

**Table B.4:** Granger causality relationships between Ukrainian refugee flows crossing the border into Poland and the proportion of daily views of Wikipedia articles about the 19 most populous Polish cities in Ukrainian. For each relationship, the table reports the optimal lag length (in days) selected by the model, the associated F-statistic, and the p-value. Only statistically significant relationships (p-value < 0.05) are included.

## Appendix B. Assessing Online Information-Seeking Behavior During Forced Migration: Evidence from Wikipedia

### 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 (Avramescu and Wiśniowski, 2021; Böhme et al., 2020; Sanliturk and Billari, 2024) and Yandex (Anastasiadou et al., 2024). Data provided by the Google Trends tool by Google are often preferred due to the widespread use of Google worldwide. While the Google Trends data are proven to be a useful indicator of interest in moving and the 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 instead 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 with a sample size that is unknown to the users. If the query cannot pass a threshold level of interest, which is determined by Google and is unknown to the users, the level of interest is deemed negligible and is reported as zero. Google Trends reports the daily search popularity index for a time period shorter than nine months. For longer periods, researchers may stitch together multiple datasets of nine months and rescale. While the Google Trends Index has proven useful and consistent despite these limitations, it must be acknowledged that the 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 introduced the use of Wikipedia data as an indicator of migration and mobility. Wikipedia data have certain advantages over Google Trends data, the most important of which is that Wikipedia provides the absolute number of pageviews instead of a normalized index, which 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 make it easier to differentiate between two groups. However, Wikipedia provides information only on the language of the viewed page, and 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, pageviews in the 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 Google Trends data, we collected Google Trends data [84] matching the Wikipedia data in our

[84] Google Trends data were collected using the `gtrendsR` package in R and `pytrends` in Python. Both R and Python were used to accelerate the data collection process.

study for a descriptive analysis. Figure B.7 shows a comparison between the Wikipedia data used in our study and the matching Google Trends data for the 19 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, because in many cases query by keyword resulted in zero-inflated data or no data at all due to low interest. To enable an easier visual comparison, we normalized the Wikipedia views to the same range as the Google Trends data, i.e., 0-100. We then compare and contrast the changes in searches for information about 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, we can see that Google Trends data reports many zero values and noise in the data. This creates an advantage for Wikipedia data over Google Trends data, which can be observed in the cases of Białystok, Bydgoszcz, Częstochowa, Gdynia, Gliwice, Kielce, Radom, Sosnowiec, and Toruń (Figure B.7). Relative to the other cities, for which Google Trends provides good quality data, these cities are less well-known and less populated. Thus, in the case of Wikipedia, the advantage in terms of data quality is more pronounced for smaller cities. Second, for more populated big cities, for which we have better quality (less noisy) Google Trends data, we do not observe important divergences between the patterns of Google Trends data and Wikipedia data. Especially following the start of the Russian invasion of Ukraine, both sources show overlapping increases in interest that are similar in size when good quality data are available from both sources.

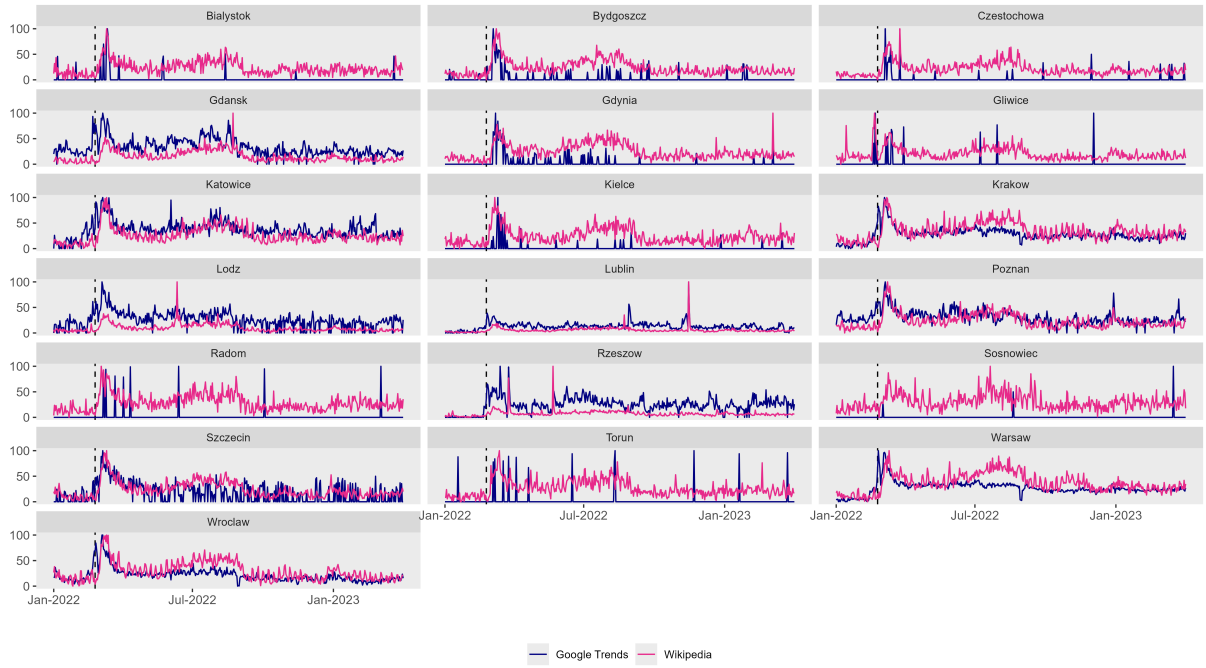



**Figure B.7:** Comparison between the Google Trends Index (GTI) of daily Google searches in Ukraine for Polish cities (as a topic) and the proportion of daily views of the corresponding Wikipedia articles about the 19 most populous Polish cities in Ukrainian. For comparability, Wikipedia views are normalized to the 0–100 range. GTI values are shown in dark blue, and Wikipedia views are shown in pink. The time series cover the period from January 1, 2022, to April 2, 2023, and the vertical dashed line marks the beginning of the Russian invasion of Ukraine (February 24, 2022).

Appendix C

# Mapping Global Gender Balance in STEM: Evidence from Facebook

**STEM and non-STEM interests available per country on Facebook Ads**

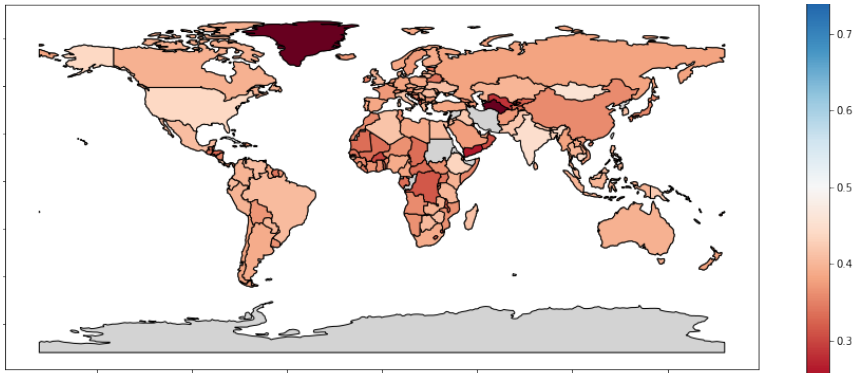



**Figure C.1:** Proportion of STEM majors on Facebook. Countries are colored from red (non-STEM) to blue (STEM), with white tones indicating balance. Gray indicates countries with unavailable information.

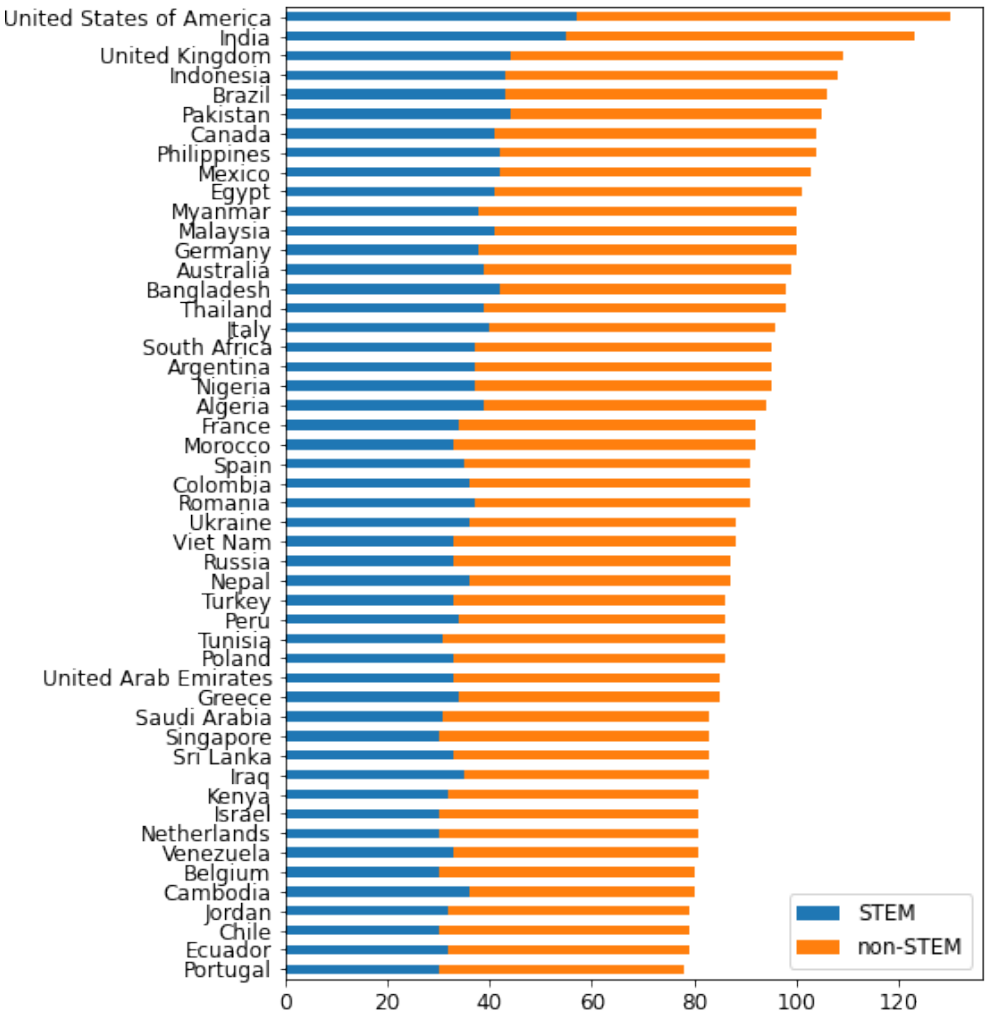
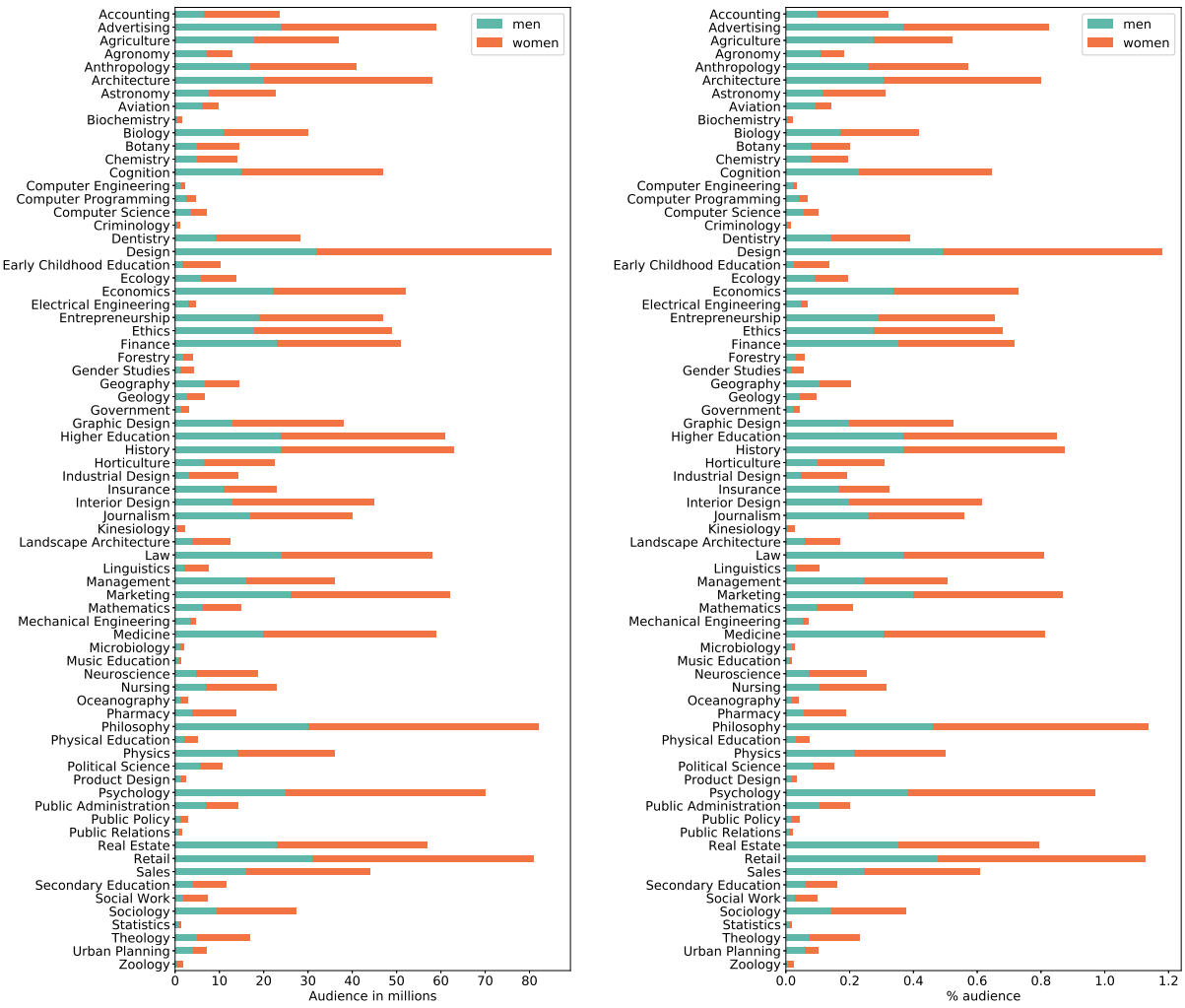



**Figure C.2:** Number of STEM and non-STEM majors on Facebook for the top 50 countries with over 1,000 monthly active users in college majors.

## Demographics attributes used to collect data for Brazil using the Facebook Ads API

| Demographic attribute | Groups | Facebook Attributes |
|---|---|---|
| **Education Level** | High school | HIGH_SCHOOL, HIGH_SCHOOL_GRAD, SOME_HIGH_SCHOOL |
| | College | UNDERGRAD, ALUM, SOME_COLLEGE, ASSOCIATE_DEGREE |
| | Graduate School | IN_GRAD_SCHOOL, SOME_GRAD_ SCHOOL, MASTER_DEGREE, PROFESSIONAL_DEGREE, DOCTORATE_DEGREE |
| **Age** | Adolescent | 13 – 19 years |
| | Young (“Early adulthood”) | 20 – 39 years |
| | Mid-aged (“Adulthood”) | 40 – 64 years |
| | Older (“Maturity”) | 65 years or more |

**Table C.1:** Facebook demographic attributes




**(a)** Number of Facebook users (in millions). **(b)** Normalized proportions of Facebook users.

**Figure C.3:** Facebook audience in Brazil interested in 73 college majors by gender.

Appendix D

# Investigating the Predictability of User-watching Behavior on TikTok via Data Donation

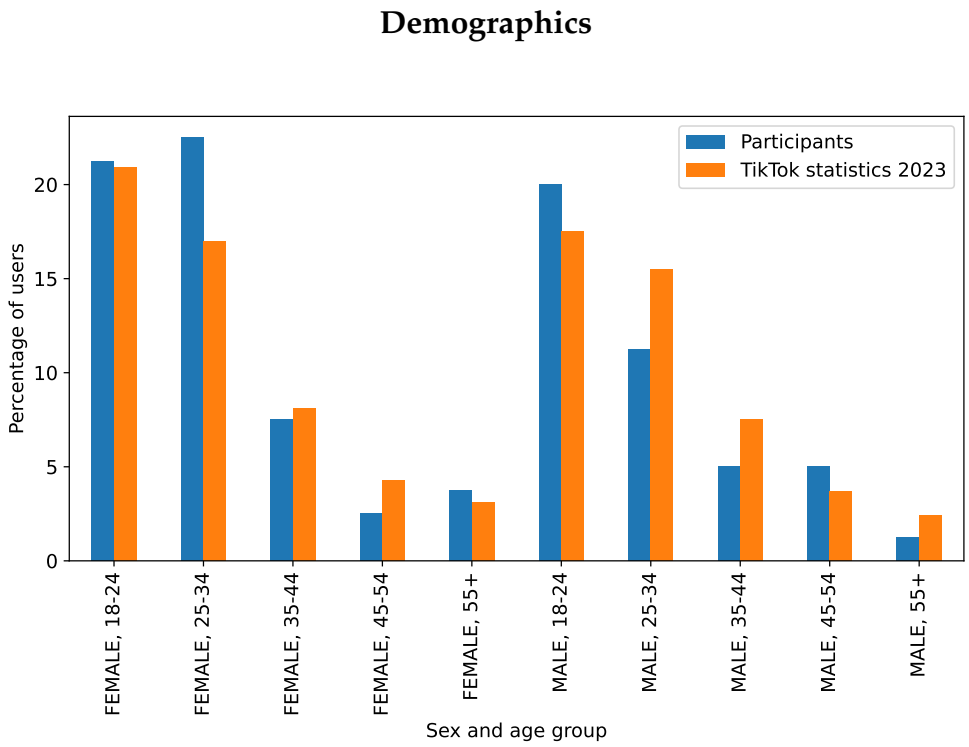



**Figure D.1:** Distribution of participants in our study (N=80) and TikTok users in October 2023 according to Statista demographics by age and sex.

| Group | Categories | Count (%) |
|---|---|---|
| Gender | Women | 46 (57.50%) |
| | Men | **34 (42.50%)** |
| Age (based on Year born) | 18-24 years old | 25 (31.25%) |
| | **25-34 years old** | **34 (42.50%)** |
| | 35-44 years old | 10 (12.50%) |
| | 45-54 years old | 5 (6.25%) |
| | $\geq$ 55 years old | 6 (7.50%) |
| Race/Ethnicity (multiple choice) | **White** | **36 (45.00%)** |
| | Asian or Asian American | 13 (16.25%) |
| | Black or African American | 11 (13.75%) |
| | Hispanic or Latino | 11 (13.75%) |
| | ... | |
| Language (multiple choice) | **English** | **80 (100.00%)** |
| | Spanish | 14 (17.50%) |
| | Chinese | 5 (6.25%) |
| | French | 4 (5.00%) |
| | ... | |
| School degree | High school diploma | 2 (2.50%) |
| | Some college, no degree | 16 (20.00%) |
| | Associate's degree | 10 (12.50%) |
| | **Bachelor's degree** | **39 (48.75%)** |
| | Advanced degree | 13 (16.25%) |
| Employment Status (multiple choice) | **Employed full-time** | **34 (42.50%)** |
| | Not employed | 12 (15.00%) |
| | Student | 11 (13.75%) |
| | Employed part-time | 10 (12.50%) |
| | ... | |
| Political leaning | **Democrat** | **56 (70.00%)** |
| | Independent-Democrat | 12 (15.0%) |
| | Independent-Republican | 5 (6.25%) |
| | Republican | 3 (3.75%) |
| | Strong Republican | 3 (3.75%) |
| | No preference, closer to Democrat | 1 (1.25%) |
| Income (annual) | **< $5,000** | **11 (13.75%)** |
| | $5,000–$10,000 | 9 (11.25%) |
| | $10,000–$20,000 | 7 (8.75%) |
| | $20,000–$30,000 | 5 (6.26%) |
| | $30,000–$40,000 | 8 (10.00%) |
| | $40,000–$50,000 | 10 (12.50%) |
| | $50,000–$65,000 | 10 (12.50%) |
| | ... | |

**Table D.1:** Demographics of participants (N = 80)

| Group | Categories | Count (%) |
|---|---|---|
| How long use TikTok | Less than a month | 3 (3.75%) |
| | 1-6 months | 6 (7.50%) |
| | 6-12 months | 12 (15.00%) |
| | **1–3 years** | **40 (50.00%)** |
| | More than 3 years | 19 (23.75%) |
| How often access TikTok | Almost constantly | 5 (6.25%) |
| | **Several times a day** | **37 (46.25%)** |
| | About once a day | 13 (16.25%) |
| | Several times a week | 18 (22.50%) |
| | Less often | 7 (8.75%) |
| How many videos engage with | Most videos (almost all videos) | 8 (10.00%) |
| | Many (more than every other video) | 6 (7.50%) |
| | Half (every other video) | 7 (8.75%) |
| | **Moderate (few to half)** | **37 (46.25%)** |
| | Few (1–2 videos) | 19 (23.75%) |
| | None of them | 3 (3.75%) |
| Avg time per day using TikTok | < 10 minutes/day | 15 (18.75%) |
| | **10–30 minutes/day** | **23 (28.75%)** |
| | 31–60 minutes/day | 17 (21.25%) |
| | 1–2 hours/day | 17 (21.25%) |
| | 2–3 hours/day | 4 (5.00%) |
| | More than 3 hours per day | 4 (5.00%) |
| TikTok viewer vs. creator | **Content consumer** | **70 (87.50%)** |
| | Equally consumer and creator | 9 (11.25%) |
| | Content creator | 1 (1.25%) |
| TikTok personal vs. business | **Personal account** | **73 (91.25%)** |
| | Both personal and business | 5 (6.25%) |
| | Business account | 2 (2.50%) |
| When access TikTok (multiple choice) | **When bored** | **16 (19.42%)** |
| | Before bed | 13 (15.65%) |
| | During work breaks | 10 (12.46%) |
| | While waiting briefly | 9 (11.59%) |
| | In the restroom | 8 (10.15%) |
| | While eating | 6 (7.54%) |
| | Getting up | 5 (6.67%) |
| | While family watches other content | 5 (6.38%) |
| | While traveling | 5 (6.38%) |
| | While commuting | 3 (3.19%) |
| | ... | |

**Table D.2:** TikTok usage characteristics of participants (N = 80)

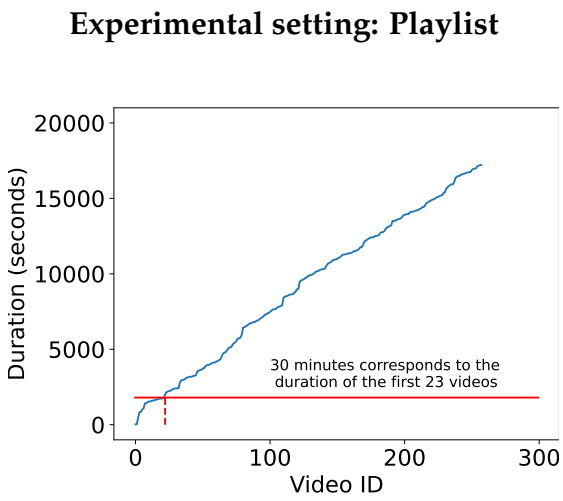
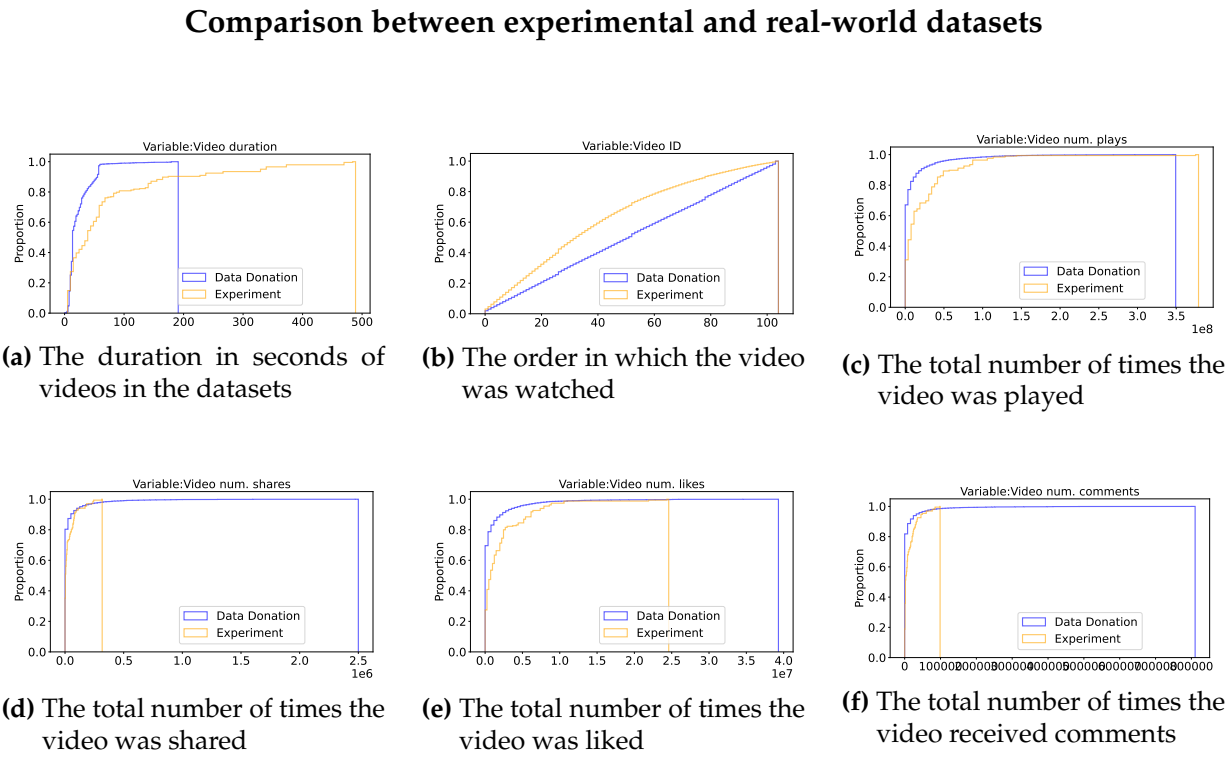



**Figure D.2:** Cumulative duration of the playlist created for the controlled experiment.




**(a)** The duration in seconds of videos in the datasets

**(b)** The order in which the video was watched

**(c)** The total number of times the video was played

**(d)** The total number of times the video was shared

**(e)** The total number of times the video was liked

**(f)** The total number of times the video received comments

**Figure D.3:** CDF of the video duration, the order in which the video was watched, and the total number of times the video was played, shared, liked, and received comments.

# Bibliography


Abel, G. J. and Cohen, J. E. (2019). Bilateral international migration flow estimates for 200 countries. *Scientific Data*, 6(1):1–13. Nature Publishing Group.

Abel, G. J., DeWaard, J., Ha, J. T., and Almquist, Z. W. (2021). The form and evolution of international migration networks, 1990–2015. *Population, Space and Place*, 27(3):e2432. Wiley Online Library.

Abel, G. J., Raymer, J., and Guan, Q. (2019). Driving factors of Asian international migration flows. *Asian Population Studies*, 15(3):243–265. Taylor & Francis.

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3):211–230. Elsevier.

Aguilera, D., Lupiáñez, J., Vílchez-González, J., and Perales-Palacios, F. (2021). In search of a long-awaited consensus on disciplinary integration in STEM education. *Mathematics*, 9(6):597. MDPI.

Al Tamime, R. and Weber, I. (2022). Using social media advertisement data to monitor the gender gap in STEM: opportunities and challenges. *PeerJ Computer Science*, 8(994).

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*, pages 617–630. JSTOR.

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*.

Algan, Y., Bisin, A., Manning, A., and Verdier, T. (2012). *Cultural integration of immigrants in Europe*. Oxford University Press.

Almerico, G. M. (2014). Food and identity: Food studies, cultural, and personal identity. *Journal of International Business and Cultural Studies*, 8:1. Citeseer.

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. Max Planck Institute for Demographic Research.

Anderson, J. E. (2011). The gravity model. *Annual Review of Economics*, 3(1):133–160. Annual Reviews.

Araujo, M., Mejova, Y., Weber, I., and Benevenuto, F. (2017). Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the ACM on Web Science Conference*, pages 253–257.

Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., and Lawless, M. (2019). Using Zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. *International Journal of Qualitative Methods*, 18:1609406919874596. SAGE Publications Sage CA: Los Angeles, CA.

Armin, A., Trybala, J. J., Young, J., and Razi, A. (2024). Support in Short Form: Investigating TikTok Comments on Videos with# Harassment. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Ashley, B., Hollows, J., Jones, S., and Taylor, B. (2004). *Food and cultural studies*. Routledge.

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. JSTOR.

Bai, J. (1997). Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4):551–563. MIT Press.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22. Wiley Online Library.

Bailard, C. S., Tromble, R., Zhong, W., Bianchi, F., Hosseini, P., and Broniatowski, D. (2024). "Keep Your Heads Held High Boys!": examining the relationship between the Proud Boys' online discourse and offline activities. *American Political Science Review*, 118(4):2054–2071. Cambridge University Press.

Baily, J. and Collyer, D. M. (2006). Introduction: Music and Migration. *Journal of Ethnic and Migration Studies*, 32(2):167–182. Routledge _eprint: https://doi.org/10.1080/13691830500487266.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132. American Association for the Advancement of Science.

Beine, M., Bertoli, S., and Fernández-Huertas Moraga, J. (2016). A practitioners' guide to gravity models of international migration. *The World Economy*, 39(4):496–512. Wiley Online Library.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement*, pages 49–62.

Berry, J. W. (2006). Contexts of acculturation. *The Cambridge Handbook of Acculturation Psychology*, pages 27–42. Cambridge University Press.

Beyer, R. M., Schewe, J., and Lotze-Campen, H. (2022). Gravity models do not explain, and cannot predict, international migration dynamics. *Humanities and Social Sciences Communications*, 9(1):56.

Bierbrauer, G. and Pedersen, P. (1996). Culture and migration. *Applied Social Psychology*, pages 399–422. Sage London.

Bijak, J. (2022). Uncertainty and Complexity: Towards Model-Based Demography. In *Towards Bayesian Model-Based Demography*, pages 13–29. Springer.

Boeker, M. and Urman, A. (2022). An Empirical Investigation of Personalization Factors on TikTok. In *Proceedings of the ACM Web Conference*, WWW '22, pages 2298–2309, New York, NY, USA. Association for Computing Machinery.

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., and Oberski, D. L. (2020). Digital trace data collection through data donation. *arXiv preprint arXiv:2011.09851*.

Bokde, D., Girase, S., and Mukhopadhyay, D. (2015). Matrix factorization model in collaborative filtering algorithms: A survey. *Procedia Computer Science*, 49:136–146. Elsevier.

Botella, C., Rueda, S., López-Iñesta, E., and Marzal, P. (2019). Gender Diversity in STEM Disciplines: A Multiple Factor Problem. *Entropy*, 21(1):30. MDPI AG.

Boutaud, J., Becuţ, A., and Marinescu, A. (2016). Food and culture. Cultural patterns and practices related to food in everyday life. *International Review of Social Research*, 6:1–3.

Boyd, D. M. and Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1083-6101.2007.00393.x.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370. Springer-Verlag.

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. Elsevier.

Cantarero, L., Espeitx, E., Gil Lacruz, M., and Martín, P. (2013). Human food preferences and cultural identity: The case of Aragón (Spain). *International Journal of Psychology*, 48(5):881–890. Taylor & Francis.

Caragliu, A., Del Bo, C., de Groot, H. L., and Linders, G.-J. M. (2013). Cultural determinants of migration. *The Annals of Regional Science*, 51(1):7–32. Springer.

CEH (1999). *Capítulo segundo: Las violaciones de los derechos humanos y los hechos de violencia [Chapter II: Human Rights Violations and the Acts of Violence]*. UNOPS - Comisión para el Esclarecimiento Histórico, Guatemala City.

Cheryan, S., Ziegler, S., Montoya, A., and Jiang, L. (2016). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, 143.

Chi, G., Abel, G. J., Johnston, D., Giraudy, E., and Bailey, M. (2025). Measuring global migration flows using online data. *Proceedings of the National Academy of Sciences*, 122(18):e2409418122.

Christie, M., O Neill, M., Rutter, K., Young, G., and Medland, A. (2017). Understanding why women are under-represented in Science, Technology, Engineering and Mathematics (STEM) within Higher Education: a regional case study. *Production*, 27. scielo.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118. National Academy of Sciences.

Citroni, G. (2014). The Pitfalls of Regulating the Legal Status of Disappeared Persons Through Declaration of Death. *Journal of International Criminal Justice*, 12(4):787–803.

Clewell, B. and Campbell, P. (2002). Taking stock: Where we've been, where we are, where we're going. Journal of Women and Minorities in Science and Engineering.

Cohen, J. E., Roig, M., Reuman, D. C., and GoGwilt, C. (2008). International migration beyond gravity: A statistical model for use in population projections. *Proceedings of the National Academy of Sciences*, 105(40):15269–15274. National Academy of Sciences _eprint: https://www.pnas.org/content/105/40/15269.full.pdf.

Cooper, C. R. and Denner, J. (1998). Theories linking culture and psychology: Universal and community-specific processes. *Annual Review of Psychology*, 49(1):559–584. Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA.

Counihan, C., Van Esterik, P., and others (1997). *Food and culture*. Routledge, New York.

Crump, J. (2011). What Are the Police Doing on Twitter? Social Media, the Police and the Public. *Policy & Internet*, 3(4):1–27. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2202/1944-2866.1130.

Cruz, J. M., Tanyu, M., Vorobyeva, Y., Mizrahi, Y., Coombes, A., Sánchez, J., Hill, C., and Campie, P. (2020). A study of gang disengagement in Guatemala. Technical report, American Institutes for Research & Florida International University.

Cuevas, A., Cuevas, R., Desmet, K., and Ortuño-Ortín, I. (2021). The Gender Gap in Preferences: Evidence from 45,397 Facebook Interests. Technical report, National Bureau of Economic Research.

Dai, M., He, W., Tian, X., Giraldi, A., and Gu, F. (2017). Working with communities on social media: Varieties in the use of Facebook and Twitter by local police. *Online Information Review*. Emerald Publishing Limited.

Davies, M. and Fleiss, J. L. (1982). Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051. JSTOR.

De Haas, H. (2021). A theory of migration: the aspirations-capabilities framework. *Comparative Migration Studies*, 9(1):1–35.

De Santis, G., Maltagliati, M., and Salvini, S. (2016). A measure of the cultural distance between countries. *Social Indicators Research*, 126(3):1065–1087. Springer.

De Solier, I. and Duruz, J. (2013). Food cultures: Introduction. *Cultural Studies Review*, 19(1):4–8.

Dekker, C. A., Baumgartner, S. E., and Sumter, S. R. (2025). For you vs. for everyone: The effectiveness of algorithmic personalization in driving social media engagement. *Telematics and Informatics*, 101:102300.

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. SAGE Publications Sage UK: London, England.

DESA, U. (2019). International Migrant Stock 2019. *Population Division New York*.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a):427–431. Taylor & Francis.

Drouhot, L. G., Deutschmann, E., Zuccotti, C. V., and Zagheni, E. (2023). Computational approaches to migration and integration research: promises and challenges. *Journal of Ethnic and Migration Studies*, 49(2):389–407. Taylor & Francis.

Dubois, A., Zagheni, E., Garimella, K., and Weber, I. (2018). Studying migrant assimilation through Facebook interests. In *International Conference on Social Informatics*, pages 51–60. Springer. Backup Springer.

Duggan, M., Brenner, J., and others (2013). The demographics of social media users, 2012. Technical report, Pew Research Center's Internet & American Life Project, Washington, DC.

Duszczyk, M. and Kaczmarczyk, P. (2022). The war in Ukraine and migration to Poland: outlook and challenges. *Intereconomics*, 57(3):164–170. Springer.

Egger, P. H., Von Ehrlich, M., and Nelson, D. R. (2012). Migration and trade. *The World Economy*, 35(2):216–241.

Ertl, B., Luttenberger, S., and Paechter, M. (2017). The Impact of Gender Stereotypes on the Self-Concept of Female Students in STEM Subjects with an Under-Representation of Females. *Frontiers in Psychology*, 8:703.

Esses, V. M. (2018). Immigration, migration, and culture. *Oxford Research Encyclopedia of Psychology*.

Fagiolo, G. and Mastrorillo, M. (2013). International migration network: Topology and modeling. *Phys. Rev. E*, 88(1):012812. American Physical Society.

Falk, A. and Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. *Science*, 362(6412).

Falter, M., Arenas, A. A., Maples, G. W., Smith, C. T., Lamb, L. J., Anderson, M. G., Uzzell, E. M., Jacobs, L. E., Cason, X. L., Griffis, T. A. N., and others (2022). Making room for Zoom in focus group methods: opportunities and challenges for novice researchers (during and beyond COVID-19). *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 23(1).

Fatehkia, M., del Villar, Z., Koebe, T., Letouzé, E., Lozano, A., Al Feel, R., Mrad, F., and Weber, I. (2022). Using Facebook advertising data to describe the socio-economic situation of Syrian refugees in Lebanon. *Frontiers in Big Data*, 5. Frontiers.

Fatehkia, M., Kashyap, R., and Weber, I. (2018). Using Facebook ad data to track the global digital gender gap. *World Development*, 107:189–209. Elsevier.

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, 8(2):195–222. Springer.

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. Common Ground Publishing.

Ferguson, G. M., Iturbide, M. I., and Raffaelli, M. (2020). Proximal and remote acculturation: Adolescents' perspectives of biculturalism in two contexts. *Journal of Adolescent Research*, 35(4):431–460. Sage Publications Sage CA: Los Angeles, CA.

Ferguson, L. and Soave, V. (2021). #Missing to #Found: Exploring police Twitter use for missing persons investigations. *Police Practice and Research*, 22(1):869–885.

Flores, R. D. (2017). Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data. *American Journal of Sociology*, 123(2):333–384. University of Chicago Press Chicago, IL.

Forbes (2018). A Study Finds That Diverse Companies Produce 19% More Revenue.

Forum, W. E. (2020). Global Gender Gap Report.

Garcia, D., Mitike Kassa, Y., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., and Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences*, 115(27):6958–6963. National Academy of Sciences _eprint: https://www.pnas.org/content/115/27/6958.full.pdf.

Garcia-Holgado, A. and Garcia-Penalvo, F. (2022). A Model for Bridging the Gender Gap in STEM in Higher Education Institutions. In *Women in STEM in Higher Education: Good Practices of Attraction, Access and Retainment in Higher Education*, pages 1–19. Springer.

García, J. and Aburto, J. M. (2019). The impact of violence on Venezuelan life expectancy and lifespan inequality. *International Journal of Epidemiology*, 48(5):1593–1601. Oxford University Press.

García-Holgado, A., Mena, J., García-Peñalvo, F. J., Pascual, J., Heikkinen, M., Harmoinen, S., García-Ramos, L., Peñabaena-Niebles, R., and Amores, L. (2020). Gender equality in STEM programs: a proposal to analyse the situation of a university about the gender gap. In *IEEE Global Engineering Education Conference (EDUCON)*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Ghemawat, P. (2001). Distance still matters. *Harvard Business Review*, 79(8):137–147.

Gil-Clavel, S. and Zagheni, E. (2019). Demographic differentials in Facebook usage around the world. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 647–650. Issue: 01.

Gillespie, T. (2014). The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–194. MIT Press.

Gompers, P. and Kovvali, S. (2018). The Other Diversity Dividend. *Harvard Business Review*, 96(4):72–77.

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*, page e2790. Wiley Online Library.

Google (2024). FAQ about Google Trends data.

Granger, C. W. J. and Newbold, P. (2014). *Forecasting economic time series*. Academic Press.

Groves, R. M. and Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, 74(5):849–879. _eprint: https://academic.oup.com/poq/article-pdf/74/5/849/5144458/nfq065.pdf.

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R. D., Ventura, I., and Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement_2):S343–S363. Oxford University Press.

Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., Zagheni, E., Flores, R. D., Ventura, I., Weber, I., and others (2021). How reliable is Facebook's advertising data for use in social science research? Insights from a cross-national online survey. Technical report, Max Planck Institute for Demographic Research, Rostock, Germany.

Guimarães, S., Reis, J., Vasconcelos, M., and Benevenuto, F. (2021). Characterizing political bias and comments associated with news on Brazilian Facebook. *Social Network Analysis and Mining*, 11(1):94.

Gupta, V., Hanges, P. J., and Dorfman, P. (2002). Cultural clusters: Methodology and findings. *Journal of World Business*, 37(1):11–15. Elsevier.

Haranko, K., Zagheni, E., Garimella, K., and Weber, I. (2018). Professional Gender Gaps Across US Cities. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Harper, F. M. and Konstan, J. A. (2015). The MovieLens datasets: history and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19.

Hattingh, M. and Matthee, M. (2016). Using Facebook to Find Missing Persons: A Crowd-Sourcing Perspective. In Dwivedi, Y. K., Mäntymäki, M., Ravishankar, M., Janssen, M., Clement, M., Slade, E. L., Rana, N. P., Al-Sharhan, S., and Simintiras, A. C., editors, *Social Media: The Good, the Bad, and the Ugly*, pages 685–694, Cham. Springer International Publishing.

Herdağdelen, A., State, B., Adamic, L., and Mason, W. (2016). The Social Ties of Immigrant Communities in the United States. In *Proceedings of the ACM Conference on Web Science*, WebSci '16, pages 78–84, New York, NY, USA. ACM. event-place: Hannover, Germany.

Hill, B. M. and Shaw, A. (2014). Consider the redirect: A missing dimension of Wikipedia research. In *Proceedings of the International Symposium on Open Collaboration*, pages 1–4.

Hofstede, G. (1983). National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization*, 13(1-2):46–74. Taylor & Francis.

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*, page 00491241221140144. SAGE Publications Sage CA: Los Angeles, CA.

Huyer, S. (2015). Is the Gender Gap Narrowing in Science and Engineering? incollectionUNESCO Science Report: Towards 2030.

INEP (2018). Sinopses Estatísticas da Educação Superior – Graduação.

Inglehart, R. (1997). *Modernization and Postmodernization: Cultural, Economic, and Political Change in 43 Societies*. Princeton University Press.

Inglehart, R. and Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

Inglehart, R. and Welzel, C. (2010). Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics*, pages 551–567. JSTOR.

Intelligence, E. R. (2017). Gender in the Global Research Landscape. Elsevier.

International Organization for Migration (2022). DTM Germany - Third Country Nationals arriving from Ukraine in Germany.

Jacobs, E., Theile, T., Perrotta, D., Zhao, X., Anastasiadou, A., and Zagheni, E. (2024). Global gender gaps in the international migration of professionals on LinkedIn. *Population and Development Review*.

Jemielniak, D., Rychwalska, A., Talaga, S., and Ziembowicz, K. (2021). WikiProject Tropical Cyclones: the most successful crowd-sourced knowledge project with near real-time coverage of extreme weather phenomena. *Weather and Climate Extremes*, 33:100354. Elsevier.

Johansen, S. (1995). *Likelihood-based inference in cointegrated vector autoregressive models*. OUP Oxford.

Jonas, S. and Rodríguez, N. (2015). *Guatemala-US migration: Transforming regions*. University of Texas Press.

Kaasa, A. and Minkov, M. (2022). Are Different Two-Dimensional Models of Culture Just a Matter of Different Rotations? Evidence From the Analysis Based on the WVS/EVS. *Journal of Cross-Cultural Psychology*, 53(2):127–156. _eprint: https://doi.org/10.1177/00220221211072790.

Karizat, N., Delmonaco, D., Eslami, M., and Andalibi, N. (2021). Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–44. ACM New York, NY, USA.

Kashyap, R., Fatehkia, M., Tamime, R. A., and Weber, I. (2020). Monitoring global digital gender inequality using the online populations of Facebook and Google. *Demographic Research*, 43:779–816. JSTOR.

Kim, K. and Cohen, J. E. (2010). Determinants of international migration flows to and from industrialized countries: A panel data approach beyond gravity. *International Migration Review*, 44(4):899–932. SAGE Publications Sage CA: Los Angeles, CA.

Kittler, P. G., Sucher, K. P., and Nelms, M. (2016). *Food and culture*. Cengage Learning.

Klug, D., Qin, Y., Evans, M., and Kaufman, G. (2021). Trick and Please. A Mixed-Method Study On User Assumptions About the TikTok Algorithm. In *Proceedings of the ACM Web Science Conference*, WebSci '21, pages 84–92, New York, NY, USA. Association for Computing Machinery.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543. American Psychological Association.

Kulyk, V. (2024). Language shift in time of war: the abandonment of Russian in Ukraine. *Post-Soviet Affairs*, 40(3):159–174. Taylor & Francis.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a Social Network or a News Media? In *Proceedings of the International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. Association for Computing Machinery.

Kwantes, C. T. and Glazer, S. (2017). Toward an Operationalization of Culture. In *Culture, Organizations, and Work: Clarifying Concepts*, SpringerBriefs in Psychology, pages 13–43. Springer International Publishing, Cham.

Kämpf, M., Tessenow, E., Kenett, D. Y., and Kantelhardt, J. W. (2015). The detection of emerging trends using Wikipedia traffic data and context networks. *PLOS ONE*, 10(12):e0141892. Public Library of Science San Francisco, CA USA.

Laitin, D. D. (2000). What is a language community? *American Journal of Political Science*, pages 142–155. JSTOR.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., and others (2009). Computational social science. *Science*, 323(5915):721–723. American Association for the Advancement of Science.

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. Wiley Online Library.

Lee, A. Y., Mieczkowski, H., Ellison, N. B., and Hancock, J. T. (2022). The algorithmic crystal: Conceptualizing the self through algorithmic personalization on TikTok. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–22. ACM New York, NY, USA.

Lewer, J. J. and Van den Berg, H. (2008). A gravity model of immigration. *Economics Letters*, 99(1):164–167. Elsevier.

Lin, A. Y., Cranshaw, J., and Counts, S. (2019). Forecasting US domestic migration using internet search queries. In *The World Wide Web Conference*, pages 1061–1072.

Lütkepohl, H. (2013). Vector autoregressive models. In *Handbook of Research Methods and Applications in Empirical Macroeconomics*, pages 139–164. Edward Elgar Publishing.

Magno, G. and Weber, I. (2014). International Gender Differences and Gaps in Online Social Networks. In *International Conference on Social Informatics*.

Manly, C., Wells, R., and Kommers, S. (2018). The influence of STEM definitions for research on women's college attainment. *International Journal of STEM Education*, 5(1):45.

Martinez-Folgar, K., Alburez-Gutierrez, D., Paniagua-Avila, A., Ramirez-Zea, M., and Bilal, U. (2021). Excess Mortality During the COVID-19 Pandemic in Guatemala. *American Journal of Public Health*, pages e1–e8.

Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., and Taylor, J. E. (1993). Theories of International Migration: A Review and Appraisal. *Population and Development Review*, 19(3):431–466. [Population Council, Wiley].

Massey, D. S. and España, F. G. (1987). The social process of international migration. *Science*, 237(4816):733–738. American Association for the Advancement of Science.

Mayer, T. and Zignago, S. (2011a). Notes on CEPII's distances measures: The GeoDist database. *Working Paper*. CEPII.

Mayer, T. and Zignago, S. (2011b). Notes on CEPII's distances measures: The GeoDist database. Working Papers 2011-25, CEPII.

Mayer, T., Zignago, S., and others (2005). *Market access in global and regional trade*. CEPII Paris.

McCormick, T. H., Lee, H., Cesare, N., Shojaie, A., and Spiro, E. S. (2017). Using Twitter for demographic and social science research: Tools for data collection and processing. *Sociological Methods & Research*, 46(3):390–421. SAGE Publications Sage CA: Los Angeles, CA.

McIlwaine, C. and Moser, C. O. (2001). Violence and social capital in urban poor communities: perspectives from Colombia and Guatemala. *Journal of International Development*, 13(7):965–984. John Wiley & Sons, Ltd. Chichester, UK.

McIver, D. J. and Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. *PLoS Computational Biology*, 10(4):e1003581. Public Library of Science San Francisco, USA.

Mejova, Y., Gandhi, H. R., Rafaliya, T. J., Sitapara, M. R., Kashyap, R., and Weber, I. (2018). Measuring subnational digital gender inequality in india through gender gaps in facebook use. In *Proceedings of the ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 1–5.

Melitz, J. and Toubal, F. (2012). Native language, spoken language, translation and trade. Working Papers 2012-17, CEPII.

Meneghel, S. N. and Hirakata, V. N. (2011). Femicides: female homicide in Brazil. *Revista de Saude Publica*, 45:564–574.

Merisalo, M. and Jauhiainen, J. S. (2020). Digital divides among asylum-related migrants: comparing internet use and smartphone ownership. *Tijdschrift voor Economische en Sociale Geografie*, 111(5):689–704. Wiley Online Library.

Mesoudi, A. (2018). Migration, acculturation, and the maintenance of between-group cultural variation. *PLOS ONE*, 13(10):1–23. Public Library of Science.

Milton, A., Ajmani, L., DeVito, M. A., and Chancellor, S. (2023). "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1–17.

Min, S.-J. and Feaster, J. C. (2010). Missing Children in National News Coverage: Racial and Gender Representations of Missing Children Cases. *Communication Research Reports*, 27(3):207–216.

Miz, V., Hanna, J., Aspert, N., Ricaud, B., and Vandergheynst, P. (2020). What is trending on Wikipedia? Capturing trends and language biases across Wikipedia editions. In *Companion Proceedings of the Web Conference 2020*, pages 794–801.

Mohr, J. W., Bail, C. A., Frye, M., Lena, J. C., Lizardo, O., McDonnell, T. E., Mische, A., Tavory, I., and Wherry, F. F. (2020). *Measuring Culture*. Columbia University Press.

Montanari, M. (2006). *Food is culture*. Columbia University Press.

Moss-Racusin, C., Pietri, E., Toorn, J., and Ashburn-Nardo, L. (2021). Boosting the Sustainable Representation of Women in STEM With Evidence-Based Policy Initiatives. *Policy Insights from the Behavioral and Brain Sciences*, 8(1):50–58.

Mousavi, S., Gummadi, K. P., and Zannettou, S. (2024). Auditing Algorithmic Explanations of Social Media Feeds: A Case Study of TikTok Video Explanations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1110–1122.

Mucciardi, M. and De Santis, G. (2017). Cultural versus objective distances: the DBS-EM approach. *Social Indicators Research*, 130(3):867–882. Springer.

Munoz-Boudet, A. M. and Revenga, A. (2017). Breaking the STEM ceiling for girls.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., and Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, 31(6):678–701. _eprint: https://doi.org/10.1177/0956797620916782.

Myers, S. A., Sharma, A., Gupta, P., and Lin, J. (2014). Information network or social network? The structure of the Twitter follow graph. In *Proceedings of the International Conference on World Wide Web*, pages 493–498.

Najafikhah, S. and Shamizanjani, M. (2018). Examining the Motivations of LinkedIn Users and Their Demographics. In *European Conference on Social Media (ECSM) 2018*, page 171.

Obradovich, N., Özak, O., Martín, I., Ortuño Ortín, I., Awad, E., Cebrián, M., Cuevas, R., Desmet, K., Rahwan, I., and Cuevas, A. (2020). Expanding the measurement of culture with a sample of two billion humans. Technical report, National Bureau of Economic Research.

Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13. Frontiers Media SA.

Opare-Obisaw, C., Fianu, D. A. G., and Awadzi, K. (2000). Changes in family food habits: the role of migration. *Journal of Consumer Studies & Home Economics*, 24(3):145–149. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2737.2000.00127.x.

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Garcia Herranz, M., Al-Asad, M., and Weber, I. (2020). Monitoring of the Venezuelan exodus through Facebook's advertising platform. *PLOS ONE*, 15(2):1–15. Public Library of Science.

Park, M., Naaman, M., and Berger, J. (2016). A data-driven study of view duration on YouTube. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 651–654. Issue: 1.

Pedersen, P. J., Pytlikova, M., and Smith, N. (2004). Selection or network effects? Migration flows into 27 OECD countries, 1990–2000. *IZA Discussion Papers, No. 1104, Institute for the Study of Labor (IZA), Bonn*.

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, pages 759–769.

Pietka, A. and Sielska, A. (2025). The urban hierarchy: how the war in Ukraine has changed the status of the Polish city of Rzeszów. *The Annals of Regional Science*, 74(2):47–? Springer.

Poot, J. (1996). Information, communication and networks in international migration systems. *The Annals of Regional Science*, 30(1):55–73.

Rama, D., Mejova, Y., Tizzoni, M., Kalimeri, K., and Weber, I. (2020). Facebook Ads as a Demographic Tool to Measure the Urban-Rural Divide. In *Proceedings of The Web Conference*, WWW '20.

Ramos, R. (2016). Gravity models: A tool for migration analysis. *IZA World of Labor*.

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., and Zagheni, E. (2021). A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom. *Demography*, 58(6):2193–2218. Duke University Press.

Rapoport, H., Sardoschau, S., and Silve, A. (2020). Migration and Cultural Change. Working Papers 2020-10, CEPII.

Ravenstein, E. G. (1889). The laws of migration. *Journal of the Royal Statistical Society*. JSTOR.

Recchi, E. and Favell, A. (2019). *Everyday Europe: Social transnationalism in an unsettled continent*. Policy press.

Reñosa, M. D. C., Mwamba, C., Meghani, A., West, N. S., Hariyani, S., Ddaaki, W., Sharma, A., Beres, L. K., and McMahon, S. (2021). Selfie consents, remote rapport, and Zoom debriefings: collecting qualitative data amid a pandemic in four resource-constrained settings. *BMJ Global Health*, 6(1):e004193. BMJ Specialist Journals.

Ribeiro, F. N., Benevenuto, F., and Zagheni, E. (2020). How Biased is the Population of Facebook Users? Comparing the Demographics of Facebook Users and Census Data to Generate Correction Factors. In *Proceedings of the ACM Web Science Conference*, WebSci '20, pages 325–334. Place: Southampton, UK.

Ribeiro, F. N., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., and Gummadi, K. P. (2018). Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale. In *AAAI Conference on Web and Social Media*, ICWSM'18, Stanford, USA.

Ribeiro, M. H., Gligorić, K., Peyrard, M., Lemmerich, F., Strohmaier, M., and West, R. (2021). Sudden Attention Shifts on Wikipedia During the COVID-19 Crisis. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:12.

Rivers, C. M. and Lewis, B. L. (2014). Ethical research standards in a world of big data. *F1000Research*, 3(38):38. F1000 Research Limited.

Robinson, V. (1998). The importance of information in the resettlement of refugees in the UK. *Journal of Refugee Studies*, 11(2):146–160. Oxford University Press.

Rodas Andrade, J. (2021). *Informe: Situacion de la trata de personas en Guatemala 2020 [Report: State of human trafficking in Guatemala 2020]*. Defensoria de las Personas Victimas de trata, Guatemala City.

Rufener, M.-C., Ofli, F., Fatehkia, M., and Weber, I. (2024). Estimation of internal displacement in Ukraine from satellite-based car detections. *Scientific Reports*, 14(1):31638. Nature Publishing Group.

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.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sanliturk, E. and Billari, F. C. (2024). Search for a New Home: Refugee Stock and Google Search. *International Migration Review*, page 01979183241275452. SAGE Publications Sage CA: Los Angeles, CA.

Schellewald, A. (2023). Understanding the popularity and affordances of TikTok through user experiences. *Media, Culture & Society*, 45(8):1568–1582. SAGE Publications Sage UK: London, England.

Schellewald, A. (2024). Discussing the role of TikTok sharing practices in everyday social life. *International Journal of Communication*, 18:909–926. USC Annenberg Press/University of Southern California.

Schluchter, T. (2024). Investigating User Perceptions of Mental Health Content on TikTok: A Comprehensive Exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Schwartz, S. H. (1994). Beyond Individualism/Collectivism: New Cultural Dimensions of Values. In Kim, U., Triandis, H. C., Kâğitçibaşi, Ç., Choi, S.-C., and Yoon, G., editors, *Individualism and Collectivism: Theory, Method, and Applications*, pages 85–119. Sage Publications, Inc., Thousand Oaks, CA.

Sen, I., Flöck, F., Weller, K., Weiß, B., and Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1):399–422. Oxford University Press.

Shani, G. and Gunawardana, A. (2010). Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. Springer.

Sibal, V. (2018). Food: Identity of culture and religion. *Food and Culture*, 6:10908–10915.

Silva, M., de Oliveira, L. S., Andreou, A., de Melo, P. O. V., Goga, O., and Benevenuto, F. (2020). Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook. *arXiv Preprint arXiv:2001.10581*.

Silva, T. H., de Melo, P. O. V., Almeida, J. M., Musolesi, M., and Loureiro, A. A. (2014). You are what you eat (and drink): Identifying cultural boundaries by analyzing food and drink habits in foursquare. In *International AAAI Conference on Weblogs and Social Media*, pages 466–475.

Soares Filho, A. M. (2011). Homicide victimization according to racial characteristics in Brazil. *Revista de Saude Publica*, 45:745–755.

Social Progress Imperative (2022). Ukraine Refugee Pulse.

Solymosi, R., Petcu, O., and Wilkinson, J. (2021). Exploring public engagement with missing person appeals on Twitter. *Policing and Society*, 31(7):798–821.

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F., Arvanitakis, G., Benevenuto, F., Gummadi, K., Loiseau, P., and Mislove, A. (2018). Potential for discrimination in online targeted advertising. In *FAT-Conference on Fairness, Accountability, and Transparency*, volume 81, pages 1–15. PMLR.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2018). Migration data using social media: a European perspective. Technical report, Publications Office of the European Union. JRC Technical Reports.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., and Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLOS ONE*, 14(10):e0224134.

Steinke, J., Lapinski, M., Long, M., Maas, C. V. D., Ryan, L., and Applegate, B. (2009). Seeing Oneself as a Scientist: Media Influences and Adolescent Girls' Science Career-Possible Selves. *Journal of Women and Minorities in Science and Engineering*, 15(4):279–301.

Steinke, J. and Tavarez, P. P. (2018). Cultural Representations of Gender and STEM: Portrayals of Female STEM Characters in Popular Films 2002-2014. *International Journal of Gender, Science and Technology*, 9(3):244–277.

Stephenson, S., Page, C. N., Wei, M., Kapadia, A., and Roesner, F. (2024). Sharenting on TikTok: Exploring Parental Sharing Behaviors and the Discourse Around Children's Online Privacy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Stewart, I., Flores, R. D., Riffe, T., Weber, I., and Zagheni, E. (2019). Rock, Rap, or Reggaeton?: Assessing Mexican Immigrants' Cultural Assimilation Using Facebook Data,. In *The World Wide Web Conference*, WWW '19, pages 3258–3264, New York, NY, USA. ACM. event-place: San Francisco, CA, USA.

Suchman, E. A. (1962). An analysis of "bias" in survey research. *Public Opinion Quarterly*, pages 102–111. JSTOR.

Tandrayen-Ragoobur, V. and Gokulsing, D. (2021). Gender gap in STEM education and career choices: what matters? *Journal of Applied Research in Higher Education*.

Tang, J., Hu, X., and Liu, H. (2013). Social recommendation: a review. *Social Network Analysis and Mining*, 3:1113–1133. Springer.

Taras, V., Rowney, J., and Steel, P. (2009). Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management*, 15(4):357–373. Elsevier.

Taylor, S. H. and Chen, Y. A. (2024). The lonely algorithm problem: the relationship between algorithmic personalization and social connectedness on TikTok. *Journal of Computer-Mediated Communication*, 29(5):zmae017. Oxford University Press.

Tinbergen, J. (1962). *Shaping the world economy: suggestions for an international economic policy*. Twentieth Century Fund.

Tizzoni, M., Panisson, A., Paolotti, D., and Cattuto, C. (2020). The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic. *PLoS Computational Biology*, 16(3):e1007633. Public Library of Science San Francisco, CA USA.

Tovey, A. (2017). Why STEM needs fresh ideas to bring about the fourth Industrial Revolution.

Tranos, E., Gheasi, M., and Nijkamp, P. (2015). International Migration: A Global Complex Network. *Environment and Planning B: Planning and Design*, 42(1):4–22. _eprint: https://doi.org/10.1068/b39042.

Tsoi, K. K., Chan, N. B., Chan, F. C., Zhang, L., Lee, A. C., and Meng, H. M. (2018). How can we better use Twitter to find a person who got lost due to dementia? *NPJ Digital Medicine*, 1(1):1–9.

Tung, R. L. and Verbeke, A. (2010). Beyond Hofstede and GLOBE: Improving the quality of cross-cultural research.

UNESCO (2017). Cracking the Code: Girls' and Womens' Education in Science, Technology, Engineering and Mathematics (STEM).

UNESCO (2020). Global Education Monitoring Report: Gender report, A new generation: 25 years of efforts for gender equality in education.

UNESCO (2021). Women a minority in Industry 4.0 fields. https://www.unesco.org/en/articles/women-minority-industry-40-fields.

United Nations Office on Drugs and Crime (2021). Global Report on Trafficking in Persons 2020. Technical report, United Nations Office on Drugs and Crime, Vienna, Austria.

Vieira, C. C., Alburez-Gutierrez, D., R. Nepomuceno, M., and Theile, T. (2022a). Desaparecidxs: Characterizing the Population of Missing Children Using Twitter. In *Proceedings of the ACM Web Science Conference*, pages 185–190. ACM.

Vieira, C. C., Fatehkia, M., Garimella, K., Weber, I., and Zagheni, E. (2022b). *Using Facebook and LinkedIn Data to Study International Mobility*, pages 141–158. Data Science for Migration and Mobility. Oxford University Press.

Vieira, C. C., Lohmann, S., and Zagheni, E. (2024). The Value of Cultural Similarity for Predicting Migration: Evidence from Food and Drink Interests in Digital Trace Data. *Population and Development Review*, 50(1):149–176.

Vieira, C. C., Lohmann, S., Zagheni, E., Vaz de Melo, P. O. S., Benevenuto, F., and Ribeiro, F. N. (2022c). The Interplay of Migration and Cultural Similarity between Countries: Evidence from Facebook Data on Food and Drink Interests. *PLOS ONE*, 17(2):1–21.

Vieira, C. C., Mousavi, S., Ayalon, O., Dash, A., Gummadi, K. P., and Zannettou, S. (2026a). Exploring the Limits of Predicting User Watching Behavior with Short-Form Videos on TikTok. In *Companion Proceedings of the ACM Web Science Conference*. ACM.

Vieira, C. C., Ribeiro, F. N., Melo, P. O. S. V. d., Benevenuto, F., and Zagheni, E. (2020). Using Facebook Data to Measure Cultural Distance Between Countries: the Case of Brazilian Cuisine. In *Proceedings of the Web Conference*, pages 3091–3097. ACM.

Vieira, C. C., Sanlitürk, E., and Zagheni, E. (2026b). Forced Migration and Information Seeking Behavior on Wikipedia: Insights from the Ukrainian Refugee Crisis. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Vieira, C. C. and Vasconcelos, M. (2021). Using Facebook Ads Data to Assess Gender Balance in STEM: Evidence from Brazil. In *Companion Proceedings of the Web Conference*, pages 145–153. ACM.

Vieira, C. C. and Vasconcelos, M. (2025). Characterizing Global Gender Gaps in STEM Using Facebook Data. In *Proceedings of the International Society of Scientometrics and Informetrics Conference*, volume 1, pages 386–400.

Violot, C., Elmas, T., Bilogrevic, I., and Humbert, M. (2024). Shorts vs. Regular Videos on YouTube: A Comparative Analysis of User Engagement and Content Creation Trends. In *Proceedings of the ACM Web Science Conference*, pages 213–223.

Vombatkere, K., Mousavi, S., Zannettou, S., Roesner, F., and Gummadi, K. P. (2024). TikTok and the Art of Personalization: Investigating Exploration and Exploitation on Social Media Feeds. In *Proceedings of the ACM Web Conference 2024*, pages 3789–3797.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151. American Association for the Advancement of Science.

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., and Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, 595(7866):197–204.

Weber, I., Kashyap, R., and Zagheni, E. (2018). Using advertising audience estimates to improve global development statistics. *ITU Journal: ICT Discoveries*, 1(2). International Telecommunication Union.

WEF (2021). Global Gender Gap Report.

White, R. (2013). Is cultural distance a determinant of international migration flows? Evidence from Denmark, Germany, and the Netherlands. *Economics Bulletin*, 33(3).

Wladyka, D. K. (2017). Queries to Google Search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies*, 25(4):312.

World Economic Forum (2016). Women and Work in the Fourth Industrial Revolution.

Wright, J. D., Marsden, P. V., et al. (2010). Survey research and social science: History, current practice, and future prospects. In *Handbook of Survey Research*, pages 3–26. Emerald Group Publishing Limited.

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*, pages 1–2, Oxford, United Kingdom. ACM.

You, Q., García-García, D., Paluri, M., Luo, J., and Joo, J. (2017). Cultural diffusion and trends in facebook photographs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 347–356. Issue: 1.

Zagheni, E. and Weber, I. (2012). You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the ACM Web Science Conference*, pages 348–351.

Zagheni, E. and Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25. Emerald Group Publishing Limited.

Zagheni, E., Weber, I., Gummadi, K., and others (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 43(4):721–734. The Population Council, Inc.

Zannettou, S. (2021). "I Won the Election!": an empirical analysis of soft moderation interventions on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 865–876.

Zannettou, S., Nemes-Nemeth, O., Ayalon, O., Goetzen, A., Gummadi, K. P., Redmiles, E. M., and Roesner, F. (2024). Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Zeileis, A., Kleiber, C., Krämer, W., and Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(1–2):109–123. Elsevier.

Zimmer, F. and Scheibe, K. (2020). Age-and Gender-dependent Differences of Asylum Seekers' Information Behavior and Online Media Usage. In *HICSS*, pages 1–10.

Zipf, G. K. (1946). The P1 P2/D hypothesis: on the intercity movement of persons. *American Sociological Review*. JSTOR.

Zou, L., Xia, L., Ding, Z., Song, J., Liu, W., and Yin, D. (2019). Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2810–2818.