Analysing Immigration Trends Using Facebook Advertising Data

1. Background

Motivation

Traditional data like censuses, registers, and surveys (e.g., Labour Force Survey) form the basis of immigration studies. They give official, standardized migrant estimates based on residence duration (Yildiz et al., 2024; Rampazzo et al., 2021). Yet, they suffer from delays of 1-2 years in publication. Accuracy varies by country due to infrastructure differences. Coverage is often incomplete, missing origins, emigrants, or short-term moves (Yildiz et al., 2024; Rampazzo et al., 2021). For example, censuses occur every decade, ignore real-time shifts, and skip details like migration age or returns (Rampazzo et al., 2021).

Social media data from Facebook's Advertising Platform provide a fresh, instant alternative. Users report hometowns, residences, languages, and demographics like age or sex. This yields aggregated estimates for advertisers, reusable as migrant proxies (Zagheni et al., 2017; Rampazzo et al., 2021). It allows quick tracking via metrics like Monthly Active Users, surpassing traditional speed (Rampazzo et al., 2021; Yildiz et al., 2024). Such data broadens migrant definitions to include brief or hidden movements, aiding theory refinement (Rampazzo et al., 2021; Spyratos et al., 2019). Biases exist, like uneven user representation (more young users in rich countries), reporting errors, and platform updates. Corrections use penetration rates to adjust estimates (Zagheni et al., 2017; Rampazzo et al., 2021).

Immigration research needs these digital sources for timely, detailed views in a fast-changing world. Traditional data alone lags behind. Merging with Facebook data fills gaps, boosts precision via fixes, and signals early migration changes for policies like EU flows (Yildiz et al., 2024; Zagheni et al., 2017). It's affordable, reusing ad data without new setups. Limits include private firm access and unclear algorithms (Spyratos et al., 2019; Rampazzo et al., 2021).

Literature Review

Recent works advance immigration analysis by blending Facebook data with traditional sources to estimate migrant numbers, tackling key flaws. Zagheni et al. (2017) first used Facebook's platform for real-time stock tracking. They relied on hometowns and residences as migration signs. A ratio correction addressed biases by scaling via penetration rates against official stats. They noted issues like sample skew, vague migrant labels, and platform shifts.

Spyratos et al. (2019) progressed by mapping global mobility with Facebook Network data. It offers aggregated tracking beyond surveys. They stressed access hurdles from corporate ownership, calling for ethical partnerships.

Rampazzo et al. (2021) created a Bayesian model for UK migrant stocks. It fused Labour Force Survey with Facebook metrics like active users, languages, and expat tags. This yielded precise, current European migrant counts, handling biases (e.g., 11% duplicates, 5% fakes). They widened migration scopes to short stays and proposed survey add-ons for elusive groups, improving methods and theories.

Yildiz et al. (2024) combined EU traditional data (censuses, registers, surveys) with Facebook for bilateral stock predictions. Their models deliver instant estimates and early alerts on mobility. They criticized traditional delays, errors, and gaps, affirming mixed methods for full European migration insights.

Together, these studies show digital data's growing value in enhancing migration research with efficient tools, despite bias and access challenges.

Our contribution

The United Nations Economic and Social Commission for Asia and the Pacific (ESCAP) serves as the UN's regional hub for fostering cooperation across 53 member states and nine associate members, spanning from Central Asia to the Pacific Islands. This diverse area, home to over 60% of the global population, grapples with profound socioeconomic disparities, rapid urbanization, and environmental pressures that drive complex migration dynamics. Studying immigration trends here is crucial. The Asia-Pacific region hosts more than 40% of the world's international migrants, fueling economic growth while exposing vulnerabilities like labor exploitation and climate-induced displacement. Accurate, timely data on these flows can inform policies to harness benefits—such as remittances and skills transfer—while mitigating risks, from inequality in origin countries to integration challenges in destinations like Australia and Singapore. Yet, traditional sources often lag in this vast region, where data collection varies widely across developing and developed economies.

Our study extends the innovative use of Facebook Advertising data to analyze immigration trends among ESCAP members and associates, addressing a key gap in the literature. Prior works, such as those by Zagheni et al. (2017) and Yildiz et al. (2024), centered on Europe and the UK, yielding valuable models for Western contexts but overlooking Asia-Pacific's unique patterns—like intra-regional labor migration and South-South flows. We adapt these methods to estimate bilateral migrant stocks in real-time for ESCAP countries, applying bias corrections tailored to the region's uneven digital penetration and cultural reporting norms.

This provides granular, near-instant insights where censuses and surveys falter, enabling early detection of shifts driven by events like economic booms in Gulf states or Pacific climate crises. By focusing on this underrepresented area, our approach supplements global efforts, promoting equitable migration governance and supporting ESCAP's goals for sustainable development. Ultimately, it demonstrates the scalability of digital traces for diverse geographies, paving the way for more inclusive demographic research.

2. Data

Data Source

In this study, we utilized audience estimates from the Facebook Advertising Platform as the primary data source for analyzing immigration trends. Initial attempts to access this data involved applying for direct access through the Facebook Marketing API. However, these efforts were unsuccessful due to account restrictions imposed by Meta, which prevented the creation or management of new Business Managers and Ad Accounts necessary for API integration.

Subsequently, we explored open-source tools to facilitate data retrieval. We tested the fbRads package (available at https://github.com/daroczig/fbRads), an R client library designed to interact with the Facebook Marketing API for creating, reading, updating, and deleting ad-related content such as custom audiences, campaigns, and ads. Developed starting in 2015, fbRads was one of the earliest third-party SDKs for the API and has been used to manage substantial ad spends across multiple accounts. It requires setting up a Facebook App, obtaining an OAuth token with ads management permissions, and initializing the package with an Ad Account ID. Functions like `fbad_get_my_ad_accounts`, `fbad_init`, and `fbad_list_ad` enable querying ad data, including audience estimates based on user demographics, locations, and behaviors. Despite its historical utility, the package is now outdated and archived as of March 2024, primarily because the maintainer faced permanent account restrictions from Meta in August 2023, limiting further development and testing. When implemented, it failed to extract current data, likely due to evolving API changes, privacy policies, and authentication requirements.

Finally, we successfully obtained the required data through the Social Media Audience API hosted by the Leverhulme Centre for Demographic Science (LCDS) at the University of Oxford (accessible at http://digitrace-datahub.ndph.ox.ac.uk/api/v1/). This API provides access to a database of aggregated social media user counts from platforms like Facebook and

Instagram, segmented by locations, time periods, and demographics. It includes endpoints such as `/query` for raw data retrieval, `/query_clean` for processed estimates (optionally with geojson geometries), `/list_collections` for viewing accessible data collections, `/monitor_collections` for recent data summaries, `/data_overview` for country-specific overviews, and `/write` for contributing new data (with authentication). Data are queried using parameters like platform, country (ISO-2 codes), gender, age ranges, dates, and geographic levels (e.g., countries, regions, cities), with audience metrics including Daily Active Users (DAU) and Monthly Active Users (MAU). Access requires an approved token for read/write permissions, ensuring compliance with Meta's Platform Terms. After contacting the Oxford researchers responsible for the platform, we obtained credentials and extracted Facebook audience data via API calls in R or Python, focusing on migrant proxies derived from user-reported locations and demographics. This approach yielded reliable, anonymized estimates suitable for our analysis, overcoming the limitations of direct API access and outdated tools.

Data classification

We extracted the data using the API's query endpoints, such as /query for raw outputs or /query_clean for refined formats, applying various optional arguments to tailor the results. These arguments, illustrated in the accompanying figure, include filters for country (using ISO-2 codes), gender (coded as 0 for all, 1 for male, or 2 for female), age ranges (with minimum and maximum bounds), date intervals (start and end), specific data collections, contributor IDs, and validation status (defaulting to true for validated entries only).

The temporal scope of the data differs across countries, extending from 2022 to 2025, with refresh rates varying from every 2-3 days in some cases to every 2-3 weeks in others, depending on the underlying collection processes.

Key response variables from the API, which form the basis of our migrant stock estimates, encompass:

- **dau**: Daily Active Users, estimating predicted daily user engagement.
- **mau**: Monthly Active Users, estimating predicted monthly engagement.
- **mau_lower** and **mau_upper**: The lower and upper bounds of MAU estimates, accounting for uncertainty in activity predictions.
- **geo_location**: JSON-structured data on predicted original residences of users. For instance, in Hong Kong datasets, this differentiates local residents by sub-regions (e.g., North District, Sha Tin District, Kowloon City District) from expatriates by

their source countries (e.g., India, Kenya), facilitating the analysis of bilateral migration patterns and demographic shifts.

These elements enable the creation of detailed profiles for trend analysis, with potential biases—such as varying platform adoption rates—mitigated through modeling techniques outlined in subsequent sections.

Optional Arguments

Argument	Description
country	A single country code using ISO-2 format (see https://www.iban.com/country-codes). If omitted, all countries will be returned, but please note that this may result in longer processing time.
gender	Gender of the population represented by data. Acceptable values: 0, 1, 2, where 0=all, 1=male, 2=female.
age_min	Lower bound of age of the audience size reported. Default = 0.
age_max	Upper bound of age of the audience size reported. Default = 999.
date_start	Earliest date to include in the query result
date_end	Latest date to include in the query result
collection	The name of a collection from which you would like to return data. The names of all collections can be obtained from the collections endpoint.
contributor_id	ID number of contributor whose data you would like to return
valid	Return only data marked as valid by contributors? Default = true. Acceptable values: true, t, yes, y, on, 1. All other values will be treated as: false.

```
import requests
import pandas as pd
import json
# query arguments
args = {
  "token": "[redacted]",
  "platform": "facebook",
  "country": "JP",
  "gender": "0",
  "date_start": "2020-01-10",
  }

response = requests.get(url='http://digitrace-datahub.ndph.ox.ac.uk/api/v1/query', params=args)

# format response as dictionary
response = response.json()

# check status
print(response.get('status'))
print(response.get('message'))

# extract data as pandas dataframe
if response.get('status') == 200:
  data = pd.DataFrame(json.loads(response.get('data')))
  data.to_csv('JAPAN.csv', index=False)
```

To refine the dataset for analyzing immigration trends, we selectively retained audience estimates from two age ranges: 13-999 and 18-999. The 13-999 range aims to capture all users, but it sometimes shows lower active user counts than 18-999, which seems illogical. This discrepancy may stem from strict laws or social pressures limiting ad targeting for teenagers. As a result, Facebook's estimates might be overly cautious to comply with child protection rules. Thus, the 18-999 data helps optimize predictions by focusing on adults, who

often dominate migrant populations and exhibit higher engagement.

For local residents, we kept only country-level aggregates, discarding district details. The API developers at the University of Oxford noted that district data suffers from greater inaccuracies due to inconsistent geolocation quality. Since our study focuses on country-based immigration, this simplifies the dataset and reduces noise.

We also split the data into female and male cohorts for analysis. This approach reveals gender-specific migration patterns, such as female flows in domestic work or male trends in construction. It allows us to assess policy impacts that vary by gender, like asylum criteria. Plus, it enhances demographic projections with nuanced insights into societal factors, such as gender inequality, aligning with modern demographic research.

```python
import pandas as pd

# Read CSV file
df = pd.read_csv("/Users/machine/Desktop/research/GUY ABEL/51_US/US_F.csv")

filtered_13 = df[
    (df['age_min'] == 13) &
    (df['age_max'] == 999) &
    (df['geo_locations'].str.contains('countries', case=False, na=False)) &
    (df['geo_locations'].str.contains('recent', case=False, na=False))
]

filtered_18 = df[
    (df['age_min'] == 18) &
    (df['age_max'] == 999) &
    (df['geo_locations'].str.contains('countries', case=False, na=False)) &
    (df['geo_locations'].str.contains('recent', case=False, na=False))
]


filtered_expats_13 = df[
    (df['age_min'] == 13) &
    (df['age_max'] == 999) &
    (df['geo_locations'].str.contains('countries', case=False, na=False)) &
    (~df['geo_locations'].str.contains('recent', case=False, na=False))
]

filtered_expats_18 = df[
    (df['age_min'] == 18) &
    (df['age_max'] == 999) &
    (df['geo_locations'].str.contains('countries', case=False, na=False)) &
    (~df['geo_locations'].str.contains('recent', case=False, na=False))
]

with pd.ExcelWriter("US_F_filtered.xlsx") as writer:
    filtered_13.to_excel(writer, sheet_name="male_local_13", index=False)
    filtered_18.to_excel(writer, sheet_name="male_local_18", index=False)
    filtered_expats_13.to_excel(writer, sheet_name="male_expats_13", index=False)
    filtered_expats_18.to_excel(writer, sheet_name="male_expats_18", index=False)
```

3. Future Directions

The next phase of this research will focus on deepening the analysis of Facebook Advertising data to explore immigration trends specifically within ESCAP (Economic and Social Commission for Asia and the Pacific) countries. This will involve refining our models to

account for regional variations in migration patterns, such as labor mobility in Southeast Asia or refugee movements in South Asia. We also aim to develop real-time tracking tools using automated API queries to monitor these trends dynamically. Additionally, integrating machine learning techniques could improve predictions of future migration flows, while partnerships with regional organizations might expand the dataset to include understudied ESCAP nations.

4. Limitations

This study faces several limitations tied to the use of Facebook Advertising data, including biases from an unrepresentative user base—favoring younger, urban populations—which may overlook older or rural migrant groups. Dependence on third-party APIs poses risks, as evidenced by account restrictions and the obsolescence of tools like fbRads, potentially disrupting data access. Varying update frequencies across countries hinder temporal consistency, and the exclusion of sub-national data limits local policy insights. Lastly, privacy regulations and ethical concerns about repurposing user data could constrain the study's broader applicability.

References

Alexander, M., Polimis, K., & Zagheni, E. (2022). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*, *41*(1), 1-28.

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., & 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.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PloS one*, *14*(10), e0224134.

Yildiz, D., Wiśniowski, A., Abel, G. J., Weber, I., Zagheni, E., Gendronneau, C., & Hoorens, S. (2024). Integrating Traditional and Social Media Data to Predict Bilateral Migrant Stocks in the European Union. *International Migration Review*, *59*(1), 90-118. https://doi.org/10.1177/01979183241249969 (Original work published 2025)

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform

to monitor stocks of migrants. *Population and Development Review*, 721-734.