Forced Migration from Ukraine: Lessons Learned from Organic Data

Lisa Singh, Katharine Donato, Ali Arab, Nathan Wycoff, Elizabeth Jacobs
Georgetown University, USA
Broad Goal

What types of organic data can improve our understanding of emerging and/or prolonged forced migration?
What is Organic Data?

Non-design data generated as part of a person’s routine and/or a society’s normal functions
<table>
<thead>
<tr>
<th><strong>Strengths</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Generated in a more natural setting</td>
</tr>
<tr>
<td>Offers real-time data for analysis</td>
</tr>
<tr>
<td>Promising in difficult-to-access environments, where design data are hard to obtain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Weaknesses</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lots of it – Hard to process</td>
</tr>
<tr>
<td>Difficult to generate variables</td>
</tr>
<tr>
<td>Noisy, partial and biased</td>
</tr>
<tr>
<td>Possible ethical considerations</td>
</tr>
</tbody>
</table>
Organic (Big) Data Sources Used in Research

- Newspapers (MaDD)
- Twitter (MaDD)
- Google trends (MaDD)
- ACLED events
- GDELT
- Facebook advertising
- LinkedIn
- Reddit
- YouTube comments
- WhatsApp
- Social media public groups
Predicting International Migration Flow from Ukraine

First 6 Months of Conflict

<table>
<thead>
<tr>
<th>Country</th>
<th>Total Flow</th>
<th>Share(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poland</td>
<td>5,820,189</td>
<td>60.63</td>
</tr>
<tr>
<td>Hungary</td>
<td>1,300,034</td>
<td>13.54</td>
</tr>
<tr>
<td>Romania</td>
<td>1,147,112</td>
<td>11.95</td>
</tr>
<tr>
<td>Slovakia</td>
<td>739,635</td>
<td>7.70</td>
</tr>
<tr>
<td>Moldova</td>
<td>593,045</td>
<td>6.18</td>
</tr>
</tbody>
</table>

Based on UNHCR flow data
UNHCR Flow Data from Ukraine
Constructing Variables: Twitter Example

Ukrainian keywords used to construct conversation buzz variables

• Flee measures
  • Flee: I am leaving; going to; taking train to; arrived at

• Insecurity measures
  • Physical: Weapons; soldiers; rockets; bombs; explosion; attack; deaths
  • Food: Hunger; food shortage; rationing, drinking water
  • Health: COVID; corona; omicron; pandemic; hospitals; medical supplies

• Contextual measures
  • Political: Zelensky; Putin; negotiations; declaration; protests; war
  • Economic: Economy; exchange rate; gas; oil; sanctions; exports; money
Modeling: Using Organic Variable to Measure Flow

\[ \mathbb{E}[ \log(y_t) ] = z'_t \beta + \delta_{d(t)} \]

- \( \log(y_t) \) - Order of Magnitude of Outflow to Slovakia, Hungary and Poland.
- \( z_t = \frac{1}{2W} \sum_{\omega=-W}^{W} x_{t+t+\omega} \) - Lag and Aggregation (Laggregated) Organic Variable:
  - \( W \) – Window radius
  - \( \tau \) – Lag
- \( \beta \) – Regression coefficient vector
- \( \delta_{d(t)} \) - Day of week effect.

Gaussian, Poisson, Negative Binomial likelihood give qualitatively similar results
Comparing Data Sources Relationship to Flow

Counts of Individuals Leaving Ukraine

LEGEND

- UNHCR Flow Estimate
- Organic Data Estimate
Explainability vs Timeliness

Lagging vs Leading

Legend:
- Red: Google trends
- Black: GDELT
- Yellow: ACLED
- Green: Newspapers
- Light green: Twitter

Axes:
- Y: Fit (R squared)
- X: Lag (Days)

Topics:
- Travel
- Economic
- Political
- Food insecurity
- Health insecurity
- Event count
- Total positive tone
- Total negative tone
- Fatalities
- Physical insecurity
- Flee

Relationships:
- Fit vs Lag (Days)
- Correlation between explainability and timeliness

Insecurity dimensions:
- Economic
- Physical
- Political
- Health

Event types:
- Migration
- Conflict
- Health crises
- Economic downturns

Data sources:
- Google trends
- GDELT
- ACLED
- Newspapers
- Twitter
‘Prediction Error’ at Two Time Points

PHASE 1:
GDELT – negative tone News - health

PHASE 2:
Previous week’s mean Trends – Food insecurity
Lessons Learned

Because Ukraine has more granular flow data, using a simple model that considers the mean flow from the previous week is reasonable (after the initial hump) for international migration.

When a crisis emerges, public organic data sources are a viable option for modeling the changing dynamics of flow.

For this crisis, Google trends data (generally) is the best organic data signal for retrospective analysis and nowcasting. All organic sources captured the two phases of the crisis.

For longer term forecasting in countries with more sparse flow data, more variables are needed and models that quantify uncertainty and allow for more variation in temporal and spatial resolution are important (hierarchical Bayesian).
### MaDD Core Team

#### Social Scientists
- **Faculty**
  - Katharine Donato
  - Elizabeth Ferris
  - Susan Martin
- **MDI/ISIM Fellow**
  - Nathan Wycoff
  - Elizabeth Jacobs*

#### Computational Scientists
- **Faculty**
  - Lisa Singh
  - Ali Arab
  - Ameeta Agrawal
- **MDI Technical Team**
  - Colton Padden*
  - Yiqing Ren
  - Maanasa Vatsavayi

#### Interdisciplinary Students
- Yanchen Wang (PhD)
- Rob Churchill (PhD)*
- Yaguang Liu (PhD)
- Ken Kawintiranon (PhD)*
- Didier Akilimali (MS)
- Qihang Wang (MS)
- Sonali Rathinam (MS)
- Aidan Pizzo (undergrad)
- Jenny Park (undergrad) * Recent alumni

---

*Thanks to McCourt Institute, Massive Data Institute, & Institute for the Study of International Migration for their funding.*