Data Science for Migration Studies
The SoBigData.eu Exploratory

Dino Pedreschi

BIG DATA AND ALTERNATIVE DATA SOURCES ON MIGRATION: FROM CASE STUDIES TO POLICY SUPPORT

European Commission
Joint Research Centre (JRC)
Ispra (Italy), 30 November 2017
A Multidisciplinary European Infrastructure for Big Data and Social Data Mining providing an integrated ecosystem for ethically sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”.

SoBigData is...
data “proxies” of social life

Shopping patterns & lifestyle

Desires, opinions, sentiments

Relationships & social ties

Movements
The Consortium

Italy
United Kingdom
Germany
Estonia
Finland
Switzerland
Nederlands

Consiglio Nazionale delle Ricerche
The University Of Sheffield
Università di Pisa
Fraunhofer FHR
TARTU ÜLIKHOOL
IMT
Leibniz Universität Hannover
KING'S College LONDON
SCUOLA NORMALE SUPERIORE
Aalto University
ETH Zürich
TU Delft

Existing national RI’s to be integrated
The pillars for reaching the goal

- a **distributed data ecosystem** for procurement, access and curation of big social data
- a **platform of interoperable, social data mining methods** and associated skills: tools, methodologies and services for mining, analysing, and visualising complex and massive datasets
- a **community of multidisciplinary scientists, innovators, public bodies, citizen organisations, SMEs**, as well as data science students at any level of education, brought together by extensive networking and innovation actions
Building the e-infra & boosting joint research: exploratories

Our first exploratories: i.e. the Research Environments tailored to specific multidisciplinary domains

- different resources will be available (and discoverable) by exploratories: data, methods and results/publications

- We have now four driving scenarios (more to come):
  - societal debates
  - societal well-being and economic performance
  - city of citizens
  - migration studies
Exploratory: Big Data for City of Citizens

Personal Mobility, Social + Mobility,
Big Data for **Well Being and Economic Performance**

Deprivation Index (in France) predicted with Mobile Phone traces

![Map of France with deprivation index](image)

![Graph showing mean entropy vs deciles](image)

Deprivation Index (in France) predicted with Mobile Phone traces

\[
\begin{align*}
    d_i^{(n)} &= \sum_{j=1}^{V} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i \\
    p_j^{(n)} &= \sum_{i=1}^{V} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j
\end{align*}
\]
Polarization, controversy and topic trends on societal debates through social media
Exploratory: Big Data for Migration Studies
Migration studies

• Use **unconventional big datasets** (social networks, mobile phones, publications, supermarkets) to study migration (flows, stocks, impact on countries of origin and destination)

• Use **official datasets** to validate results from unconventional datasets

• Use both types of data to evaluate impact of policies on migration

• We are conducting several projects
Big datasets

- **Social network and web data**
  - Twitter Streaming data: various twitter datasets from project partners, in various languages, with geolocation
  - GDELT Knowledge Graph database. a Big Data repertoire of online news articles.

- **Mobile phone data**
  - Orange dataset: mobile calls between Senegal and the rest of the world (country to country, 2012).

- **Highly educated migrants**
  - Company data (Estonia and Italy): members of the governing boards of companies (with place of birth).
  - Publication data: DBLP (computer science) and APS (physics)

- **Supermarket transaction data**
  - Itemized shopping transactions (over 10 years, with place of birth)
The story: Migration stages

• **GO: Understanding migration stocks and flows**
  • Nowcasting migration through the Twitter lens
  • Brain-drain and scientific migration
  • Policy and migration

• **STAY: Evaluating migrant integration**
  • Sentiment related to migration topics
  • Migration and language
  • Multi-culturality and sentiment
  • Migrant start-uppers

• **RETURN: Return of migrants**
  • Data journalism approach
Some examples

GO
Migration stocks

- **Ongoing** analysis
  - Large curated collection of geo-localised tweets
- Estimate user residence
  - monthly - country from which the user posted in most days
- Estimate user nationality
  - Language and network most used
- Compare with official data
Nationalities on twitter

Official statistics

GB

Official statistics

Nationalities on twitter
Brain drain

• **Question:**
  • what is the extent of migration of highly education migrants and what is the effect on the receiving community

• **Ongoing Analysis:**
  • quantify *scientific migration* in various scientific communities
  • evaluate *success* of immigrants in science
  • quantify *entrepreneurship* of immigrants (company data) and success in society
  • underline *achievements* of immigrants in an attempt to understand whether migration is beneficial both for individuals and receiving society
Scientific migration

- **Question:**
  - What features affect scientists migration?

- **Ongoing Analysis:**
  - The features that most distinguish migrant from non-migrant scientists are the number of collaborators (positive influence) and the ratio between collaborators internal and external to the current institution (negative influence).
Scientific migration

- **Question:**
  - When do scientists migrate, and where do they go?

- **Ongoing Analysis:**
  - Using machine learning we can predict whether a scientist will move or not in the next 5 years
  
<table>
<thead>
<tr>
<th>model</th>
<th>ACC</th>
<th>recall</th>
<th>precision</th>
<th>F1</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>tree</td>
<td>0.87</td>
<td>0.84</td>
<td>0.86</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>logit</td>
<td>0.83</td>
<td>0.89</td>
<td>0.65</td>
<td>0.75</td>
<td>0.85</td>
</tr>
</tbody>
</table>

- We can also predict whether the scientist will move to the institution with which they collaborated most or not (F1=0.72)
Policy and migration

• **Question:**
  • how does policy affect immigration, particularly the ratio between legal and illegal migrants?

• **Planned Analysis:**
  • quantify legal and illegal migration using official data but also alternative datasets (roaming data, tweets around hot areas, such as the “jungle of Calais”)
  • identify trend changes
  • **Challenge 1:** identify policy changes and other (economical, historical, legal) factors that could have cause observed trends
  • **Challenge 2:** identify possible policies to curb illegal migration
Some examples

STAY
Sentiment on migration topics:  
Perception of the Mediterranean Refugee Crisis

• What is the evolution of the discussions about refugees migration in Twitter?
• What is the sentiment of users across Europe in relation to the refugee crisis?
• What is the evolution of the perception in countries affected by the phenomenon?
• Are users more polarised in countries most impacted by the migration flow?
Sentiment on migration topics: Perception of the Mediterranean Refugee Crisis

- European country mentions
- Africa & Middle East country mentions
Sentiment on migration topics: Perception of the Mediterranean Refugee Crisis

- Internal and external perception by country
  - Index $\rho$ - the ratio between pro refugees users and against refugees users
  - Red means a higher predominance of positive sentiment, higher $\rho$
  - Yellow means a higher predominance of negative sentiment, lower $\rho$
Multiculturality and sentiment

- **Question:**
  - how does migration affect overall *sentiment* and *language* of a community?

- **Ongoing Analysis:**
  - quantify sentiment in tweets coming from different countries (geolocalised) and in different languages
  - compare sentiment of *various languages* in the same location
  - compare sentiment of *the same language* in various locations across the world.
  - compare sentiment across areas with different levels of immigration
Integration from sentiment

• **Question:**
  • Can we extract an integration index based on sentiment and language usage on twitter?

• **Ongoing Analysis:**
  • Sentiment spreading – annotate a large dictionary of words starting from a small annotated seed (ANEW, Bradley&Lang dictionary)
  • The correlation between original and modelled valence of words is an indicator of how language is used in multicultural regions of the world – a **superdiversity index**
  • By analysing superdiversity in subpopulations on Twitter we aim to underline the effects of the presence of migrants on language and sentiment at the same time.
Food and migration

• Data
  – customer, country of birth, association date,
  – shopping baskets, products details,
  – time and location of shopping, amount spent

• Objective
  – While the time passes is the shopping profile of foreign customers becoming more similar to those of Italian customers?
  – Can we efficiently quantify integration of various groups of immigrants?

• Collaboration with Paris School of Economics
Some examples
Return of migrants

- **Demal te niew (Go and Come Back)**
- Documentary - interviews with migrants returning to Senegal from Italy
  - Featured in Espresso, El Pais
  - Presented at Ethnographic Film Festival, Amsterdam and International Day of Migrants, Dakar.
Conclusions

• A research direction
  – Building an integration index based on various Big Data types:
    • Evaluation of Local Shopping Habits Adoption
    • Ethnical/Local Products Adoption
    • Analysis of Possible Recipes Consumption
    • Evaluation of Economic Condition through the Observation of Shopping Offers Adoption
    • Superdiversity index and sentiment variability across communities
• Collaboration through SoBigData Transnational Access
Ph.D. in Data Science

- Started: academic year 2017-2018
- Call for 2018-2019 opens in December 2017
- http://phd.sns.it/data-science/