Using Big Data to sample minorities in Western European Countries: Muslims

2018 CSDI Workshop, Limerick, Ireland
Overview

1. Introduction
   Background topic & purpose

2. Literature Review
   Exploration and analysis of previous studies

3. Methodology
   Sampling setup

4. Fieldwork & Sampling Results
   Lessons learned from sample generation & fieldwork results

5. Conclusions
   Future designs and possibilities
1. Introduction
“We would like to conduct a NatRep study of the Muslim Population in Countries X, Y & Z”
What creates interest?

- Significant Part Western Society
- Refugee Crisis
- Volatility in Middle East
- Right wing movements in Europe
What creates interest?

Muslims make up 4.9% of Europe’s population in 2016

Muslims in the EU, Norway and Switzerland in 2050: medium migration scenario

PEW RESEARCH CENTER
Sampling Challenge

1. Target Population
   Registered, self-reported?

2. Multi-Country
   Comparable results in multiple countries

3. “Representative”
   Is there data on Muslim populations? If so, on which level of detail? Weighting-Targets?
2. Literature Review
Previous Approaches

September 2017

Muslims living in the EU face discrimination in a broad range of settings – and particularly when looking for work, on the job, and when trying to access public or private services. The report examines how characteristics – such as an individual's first and last name, skin colour and the wearing of visible religious symbols like a headscarf, for example – may trigger discriminatory treatment and harassment.

Downloads:

Second European Union Minorities and Discrimination Survey (EU-MIDIS II): Muslims - Selected findings

1.53 MB
Previous Approaches

- White Pages
- Targeted RDD Sample
- Specific Providers
- Unknown Methodology

Landline Based
Full Probability Sample

Dual-Frame

Muslim Population

Drawn Sample

~3-5%

Incidence Rate
Targeted RDD Sample

13.3%  14.1%  11.3%
Rotterdam  The Hague  Amsterdam
List-Assisted RDD Sample

Name Database

White Pages

Sample
Specific Provider

Dutch Providers
Sampling Approach Evaluation

- White Pages
- Targeted RDD Sample
- Social Media

- Landline Based
- Mobile Based
3. Methodology
“How do we compose our sample?”
Sample Composition

N = 1000

50% Social Media & Messenger Sample
25% Targeted RDD Landline Sample
25% List-Assisted Sample
Onomastic Approach: Name Database

Country Identification → First / Surname → Exception Names → Filtering

Country Specific Iteration
Majority Muslim Countries
Generation of Social Media Sampling

Big Data

Cell Frame

+31 6 XXXX XXXX

Sample
Which Public Sources are used?
4. Fieldwork & Sampling Results
SAMPLE GENERATION

- 100% Raw Sample
- 45% Screened Sample
- 13% Matched Sample
- 1.2% Suitable Respondent
- 1 Suitable Respondent
Example Netherlands:

11.1 Profiled records needed to get 1 suitable record

Respondent Identification

53% to 68%: Not much differences between countries

→ Self-Reporting
### Fieldwork Results

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>TYPE</th>
<th>CENUS DATA</th>
<th>SOCIAL MEDIA</th>
<th>LISTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>GENDER</td>
<td>Male</td>
<td>49.1%</td>
<td>63.8%</td>
<td>41.8%</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>50.9%</td>
<td>36.2%</td>
<td>58.2%</td>
</tr>
<tr>
<td>AGE</td>
<td>18-29</td>
<td>10.4%</td>
<td>44.7%</td>
<td>5.1%</td>
</tr>
<tr>
<td></td>
<td>30-49</td>
<td>30.5%</td>
<td>43.1%</td>
<td>16.8%</td>
</tr>
<tr>
<td></td>
<td>50+</td>
<td>59.1%</td>
<td>12.2%</td>
<td>78.1%</td>
</tr>
</tbody>
</table>
# Fieldwork Results

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>TYPE</th>
<th>CENUS DATA</th>
<th>SOCIAL MEDIA</th>
<th>LISTED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>Lower Sec or Less</td>
<td>31.3%</td>
<td>24.6%</td>
<td>12.7%</td>
</tr>
<tr>
<td></td>
<td>Upper Sec + Non-Tert.</td>
<td>40.6%</td>
<td>43.3%</td>
<td>39.1%</td>
</tr>
<tr>
<td></td>
<td>Tertiary</td>
<td>28.1%</td>
<td>32.1%</td>
<td>48.2%</td>
</tr>
</tbody>
</table>
Fieldwork Results

Zuid-Holland
Factor 1.7

Noord-Holland
Factor 1.2

Language Problems Social Media
23.6%

Language Problems Listed
3.8%
5. Conclusion & Moving Ahead
“Why can’t we just use an online panel?”
## ADVANTAGES AND DISADVANTAGES

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Reach of first generation</td>
<td>• Age Max: ~70</td>
</tr>
<tr>
<td>• Larger coverage</td>
<td>• Bias towards online</td>
</tr>
<tr>
<td>• Younger Age groups well</td>
<td>• Dependency on name database</td>
</tr>
<tr>
<td>represented</td>
<td>• Only access to public profiles</td>
</tr>
<tr>
<td>• Geographical coverage</td>
<td></td>
</tr>
</tbody>
</table>
Future: Full Probability Frame

Increased Matching Rate
Addition of more data sources to increasing matching rates

Onomastic Approach
More detail to local population data. Not just full name but also consideration of first / surname.

Removing non-qualifier
Reverse Logic: Screen-out those that do not qualify.

Full Probability Frame
Heading towards a full single frame.
Thank you!