Ways of Reducing coverage and sampling error as part of the Total Survey Error Framework for Establishment Surveys in Europe: Recent Developments

Nikola Jovanovski - Sample Solutions BV
Carsten Broich - Sample Solutions BV
Introduction
Company Background
European Establishment Surveys
Big Data as a crucial part in sampling

Research
Sample challenges and Research Problems
ESRA 2019 Future considerations

Establishment Sample sources
What are the different sources of establishment sample data for the EU27

Enrichment Methodology
How can we use a Big Data approach to decrease sampling error in European Establishment Surveys

Pros and Cons
Benefits and limitations from using the enrichment approach

Conclusion
Research Problem address
Q&A
1. INTRODUCTION
About Sample Solutions

Background
Founded in the Netherlands, back in 2009 with the focus on Business & Consumer Telephone Sample

Sample Survey Platform B2B module
Specialized B2B Database designed Survey Research - instant counts and sample ordering

Multi country B2B sample Projects
Eurobarometer, London Economics, PwC, ABB, World Bank, American Express
WHAT IS BIG DATA?
“Big data” is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.
European Establishments Surveys:

- Flash Eurobarometer (business survey part)
- European Companies Survey (ECS)
- European Survey of Enterprises on New and Emerging Risks (ESENER)
Sampling frame sources per country:

- B2B data vendors (finance, marketing, sales purposes)
- National Business Register
- Chamber of Commerce Directory
2. RESEARCH
Issues among the Sampling sources

- Decentralized Sample data vending (1 vendor per country)
  - Under representativeness among certain:
    - Industries
    - Company Sizes
    - Area typologies
    - Subsidiaries (Branches)
  - Variable sample costs among countries
  - Limited coverage of Contact details of decision makers and officers
  - No or limited coverage of email addresses

Source: Business Count that includes entire frame with and without phone number records
Sample comparison: Phone v Email

Establishment sample

➔ Offline - phone
  ◆ Key breakdown values (Industry code, Size)
  ◆ No direct contact point with decision maker
  ◆ Overall better response rates

➔ Online - email
  ◆ Key breakdown values (Industry code, Size)
  ◆ Direct email of Decision Maker
  ◆ Targeting specific job roles based on research

Phone
  • Response rates
  • Coverage among Sole Traders
  • Wide availability
  • Rapid interview turnaround time

Email
  • Direct approach
  • Desired role
  • Convenient time of interview
  • Low Cost
  • Non opt-in sampling
Research questions

What are the ways of reducing coverage and sampling error for Establishment Surveys in EU27:

1. How can Big Data be used for increasing the telephone coverage in the Sampling Frame?
2. How can the Sampling frame go online, for a mixed mode sampling approach?
3. How can we ensure sample data accuracy and validity in a international and longitudinal study?
Future considerations:

From ESRA 2019 we have posted these future action steps:

➔ Use Big Data to Identify Company Size (no. of Employees) on businesses with unknown size
➔ Include a validation step in the enrichment process to lower data inaccuracy
➔ Include a Area Typology stratification criteria in order to make a more representable sample to frame
➔ Design sample of all countries from a centralized source
➔ Source Contact person data and email
3. ESTABLISHMENT SAMPLE SOURCES
### ESTABLISHMENT SAMPLE ELEMENTS:

| 01 | LOCATION | • Country, City, Postcode, Region, HQ Location, Branches Location |
| 02 | INDUSTRY | • SIC code, Keywords (from LinkedIn), NAICS code, NACE code |
| 03 | COMPANY SIZE | • Employee Range • Revenue Range |
| 04 | COMPANY STATUS | • Founded date • Operating status, Entity Type |
| 05 | CONTACT OPTIONS | • Email available, Phone available, Website URL, Decision maker (C-level) |
| 06 | TECHNOLOGY | • Technologies divided into categories |
| 07 | CONTACTS INFO | • Departments, Job title • C-level, VP level, Director, Manager, Non-Manager |
Three sources for Establishment sample sources:

**TRADITIONAL**
- Available lists
- Basic Business information

**ENRICHMENT**
- Online lookups
- Big Data

**GENERATED**
- Email patterns
- Verification of email activity
Traditional sources

National Registers
State or Privately run
Categorized with NACE and SME sizes

Chamber of Commerce
Rich in business information
Can be Member only

Phone book directories
Open and free access for general public

Commercial Data vendors
Paid access
For Marketing, Credit report and Sales purposes
Rich in detail
**CHARACTERISTICS OF TRADITIONAL SOURCES:**

<table>
<thead>
<tr>
<th>BRAND</th>
<th>MANDATORY</th>
<th>FREE ACCESS</th>
<th>EMAIL ADDRESSES</th>
<th>CONTACTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Business Register</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Phone book</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Commercial Data Vendor</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Chamber of Commerce</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

*Applies to majority of cases in the EU27*
BIG DATA - ENRICHMENT SOURCES

01. Search Engines
02. Directory lookup
03. Social Media
04. Review sites
Enrichment sources and tools:

**Search**
- Google Snippet
- Company name and address search
- Email address search
- Geographic lookup

**Social Media**
- Facebook
- Phone lookup
- Linkedin
- Contact and Emp. size lookup

**Directory lookups**
- Reverse phone lookups
- Company Lookups( Company Name + Address)
- Contact person lookup

**Review Sites**
- Company lookup
- Geographic lookup
SAMPLE DATA ENRICHMENT

Contact personnel
- Decision makers
- Officers
- Employees

Sizes
- From employee count
- And Reports

Telephone number
- For missing records
- Incorrect/Outdated
- Of branches

Technology Enrichment
- From a variety of
- 300 Web
- technologies
- incorporated on the website

Misc details
- VAT number,
- BusinessDescription,
- Listed email addresses
- Entity form
## Which source can produce which data

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>Employee size</th>
<th>Phone number</th>
<th>Contact person</th>
<th>Email Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkedIn</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td>Google</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Facebook</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Directory lookup</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Yelp &amp; Tripadvisor</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
</tr>
</tbody>
</table>
Overall Enrichment rates among elements

- **Phone number**: 75%
- **Employee size**: 60%
- **Contacts**: 40%
- **Email address**: 20%
Challenges in Big Data enrichment

Native encoding
Recognition and acceptance of letters such as Greek, Cyrillic, Nordic etc.

Registration Name
Legal Name ≠ Merchant Name

Incorrect Location
Wrong or no address information reduces enrichment productivity

Inactive businesses
Input frames includes potential dead records
GENERATED SAMPLE SOURCE:
EMAIL ADDRESS SAMPLING METHODOLOGY

Input Data
Sourced Company URL to create email domain
Contact person Name

Generating Patterns
Apply input data to the pattern Combinations eg. firstname@company.com
f.lastname@company.com
firstname.lastname@company.com

Validation of email activity
Screen all combinations using a SMTP service to see whether the email server recognizes the email addresses
Email Categorization

- **Personal**
  - Carsten@samplesolutions.eu
  - info@samplesolutions.eu
- **General**
  - Nikola@gmail.com
  - Sample.Solutions@gmail.com

Categorization: Lookup table for general
Name database for personal
Freemail Provider list
SAMPLE DATA VERIFICATION

DATA VERIFICATION

EMAILS
- VERIFIED
- NOT VERIFIED
- NO STATUS

URLS
- WORKING
- NOT WORKING
- REDIRECTING
- NO STATUS
5. ENRICHMENT METHODOLOGY
Outline

European enterprises and establishments sampling featured in both online and offline sampling modes

Full Sampling method

- Mixed Mode Sample (Online and Offline): defined by:
  - Direct and/or Company Email
  - Company phone
- Merging and blending multiple business sample sources for a country into one frame ensuring maximum coverage among the EU27
- Collected email address, contact person data and phone number information from publicly available sources (URLs, Social Media etc)
- Generate Email addresses where not available
- Ensure Generated vs. Collected email data (compliance)
Full Sample Enrichment

- **Business Register**: Merge and blend sample frame among multiple sources
- **Online lookup**: Enriching otherwise unavailable data for business frame
- **Validation**: Ensuring generated output is valid
- **Evaluate variables available**: Data availability differs from source to source
- **Generation**: Inhouse generating patterns for missing data
Enriching B2B records with phone numbers

Compile list
Blank records and incorrect phone numbers are submitted

Web Mining
Bulk web query to find listing of the company

Enriched list
The enriched list is delivered to the FW

Input information
Company Name +
Address Information

Phone number
The phone number from the result is extracted
ESENER-3 Results: enrichment of wrong telephone data

![Bar Chart]

<table>
<thead>
<tr>
<th>Country</th>
<th>List</th>
<th>Enriched</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>BE</td>
<td>227</td>
<td>229</td>
</tr>
<tr>
<td>BG</td>
<td>240</td>
<td>257</td>
</tr>
<tr>
<td>HR</td>
<td>44</td>
<td>5</td>
</tr>
<tr>
<td>CY</td>
<td>143</td>
<td>123</td>
</tr>
<tr>
<td>CZ</td>
<td>557</td>
<td>550</td>
</tr>
<tr>
<td>DK</td>
<td>571</td>
<td>561</td>
</tr>
<tr>
<td>EE</td>
<td>64</td>
<td>42</td>
</tr>
<tr>
<td>FI</td>
<td>118</td>
<td>166</td>
</tr>
<tr>
<td>FR</td>
<td>253</td>
<td>3089</td>
</tr>
<tr>
<td>DE</td>
<td>3251</td>
<td>0</td>
</tr>
<tr>
<td>GR</td>
<td>12</td>
<td>452</td>
</tr>
<tr>
<td>HU</td>
<td>1121</td>
<td>19</td>
</tr>
<tr>
<td>IS</td>
<td>109</td>
<td>173</td>
</tr>
<tr>
<td>IE</td>
<td>252</td>
<td>535</td>
</tr>
<tr>
<td>IT</td>
<td>1006</td>
<td>16</td>
</tr>
<tr>
<td>LV</td>
<td>78</td>
<td>18</td>
</tr>
<tr>
<td>LT</td>
<td>75</td>
<td>116</td>
</tr>
<tr>
<td>LX</td>
<td>197</td>
<td>2</td>
</tr>
<tr>
<td>MK</td>
<td>87</td>
<td>25</td>
</tr>
<tr>
<td>MT</td>
<td>156</td>
<td>352</td>
</tr>
<tr>
<td>NL</td>
<td>552</td>
<td>241</td>
</tr>
<tr>
<td>NO</td>
<td>350</td>
<td>29</td>
</tr>
<tr>
<td>PT</td>
<td>79</td>
<td>14</td>
</tr>
<tr>
<td>RO</td>
<td>57</td>
<td>25</td>
</tr>
<tr>
<td>RS</td>
<td>136</td>
<td>44</td>
</tr>
<tr>
<td>SK</td>
<td>186</td>
<td>29</td>
</tr>
<tr>
<td>SI</td>
<td>342</td>
<td>140</td>
</tr>
<tr>
<td>ES</td>
<td>193</td>
<td>775</td>
</tr>
<tr>
<td>SE</td>
<td>216</td>
<td>75</td>
</tr>
<tr>
<td>CH</td>
<td>339</td>
<td>216</td>
</tr>
<tr>
<td>GB</td>
<td>817</td>
<td>354</td>
</tr>
</tbody>
</table>
ESENER-3 Results: Enrichment of without phone
Strategies for limiting sample Coverage & Bias

Sources
- Source validation

Cross-validation
- Referencing one data fragment across multiple sources

Cross-enrichment
- Using multiple sources

Sustainability
- Keeping the data up to date

*Weighting is possible but might get too statistical (adapting selection probability)*
5. PROS AND CONS:
BENEFITS

DETAILED RICH SAMPLE DATA
Data contains more data elements than any other individual source

Verification
Data accuracy is validated by multiple sources

Sample going Online
Combining phone and online in survey fieldwork

Centralized source for EU27 data
One stop shop for comprehensive data in all 27 countries

Increased coverage
More sample data can be accessed via phone or email
B. Limitations/Possible Issues

- Lack of data availability among Sole Traders
- Sources limiting available information
- Underrepresentation among certain industries that are fully offline
- Unintentional sourcing of inactive businesses
- What are the ethical aspects that should be considered when sampling email addresses?
Research questions

What are the ways of reducing coverage and sampling error for Establishment Surveys in EU27:

1. How can Big Data be used for increasing the telephone coverage in the Sampling Frame
2. How can the Sampling frame go online, for a mixed mode sampling approach?
3. How can we ensure sample data accuracy and validity in a international and longitudinal study?
Future considerations:

From ESRA 2019 we have posted these future action steps:

➔ Use Big Data to Identify Company Size (no. of Employees) on businesses with unknown size (COMPLETED)
➔ Include a validation step in the enrichment process to lower data inaccuracy (COMPLETED)
➔ Include a Area Typology stratification criteria in order to make a more representable sample to frame (PENDING)
➔ Design sample of all countries from a centralized source (PENDING)
➔ Source Contact person data and email (COMPLETED)