Alternative way of data collection – using web scraping for ICT-statistics

Björn Forssell
Statistics Sweden
ICT usage in enterprises

- The purpose of the survey is to highlight the availability and use of ICT-technology among enterprises in Sweden. Regulated through Eurostat.
- Examples of topics included in the survey is use of computers, ICT specialists, internet use, social media and website use, cloud services, e-commerce, software development and IT and environment.
- The Swedish survey for 2017 contains about 112 questions.
Purpose and background

- Funds from the Innovation lab at Statistics Sweden
- Statistics Sweden have purchased a license from the company Vainus database for six months. Vainu provides a data-driven company database using open data from the Internet and data from our Statistical Business Register to match enterprises.
- The purpose of this evaluation have been to try to find alternative ways to reduce the burden of reporting for enterprises while increasing the quality of the statistics we collect.
Scope of this project

- The project has investigated 3 main variables related to social media and website usage for 2017:
  - Share of companies with website
  - Use of social networks
  - Use of blogs and microblogs

- These were chosen as test variables because they focus on external use via the Internet and most comparable to our survey
Methodology and assumptions

- Same sample as original survey for 10+ employees enterprises.
- Treating the web data as survey data
- Variables do not have the same definition in full extent → careful conclusions from data.
- Quality check of web data before estimates. Took away around 300 enterprises with wrong match of data.
- The estimates are made on the population comprised of enterprises that are both included in the data set from Vainu and who have responded to the regular survey.
Results – Use of web page

Survey data: 91%
Cl: 1%
Web data: 82%
Cl: 1%
## Web page after industry

<table>
<thead>
<tr>
<th>Industry</th>
<th>Web data</th>
<th>Survey</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>250 or more employees</td>
<td>99%</td>
<td>98%</td>
<td>1%</td>
</tr>
<tr>
<td>ICT sector</td>
<td>94%</td>
<td>97%</td>
<td>-3%</td>
</tr>
<tr>
<td>50-249 employees</td>
<td>94%</td>
<td>97%</td>
<td>-3%</td>
</tr>
<tr>
<td>Information and communication enterprises</td>
<td>93%</td>
<td>96%</td>
<td>-3%</td>
</tr>
<tr>
<td>Energy and recycling</td>
<td>95%</td>
<td>99%</td>
<td>-4%</td>
</tr>
<tr>
<td>Transport and storage enterprises</td>
<td>80%</td>
<td>76%</td>
<td>4%</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>89%</td>
<td>93%</td>
<td>-4%</td>
</tr>
<tr>
<td>Other service companies</td>
<td>86%</td>
<td>91%</td>
<td>-5%</td>
</tr>
<tr>
<td>10+ employees (total)</td>
<td>83%</td>
<td>91%</td>
<td>-8%</td>
</tr>
<tr>
<td>Real estate companies and managers</td>
<td>87%</td>
<td>95%</td>
<td>-8%</td>
</tr>
<tr>
<td>10-49 employees</td>
<td>81%</td>
<td>90%</td>
<td>-9%</td>
</tr>
<tr>
<td>Trade</td>
<td>86%</td>
<td>95%</td>
<td>-9%</td>
</tr>
<tr>
<td>Construction</td>
<td>77%</td>
<td>88%</td>
<td>-10%</td>
</tr>
<tr>
<td>Accommodation and food services</td>
<td>63%</td>
<td>89%</td>
<td>-26%</td>
</tr>
</tbody>
</table>
Results - social networks

- In survey: Social networks e.g. Facebook, LinkedIn.
- In web data: Facebook or LinkedIn.
Results - use of blogs or microblogs

- In survey: Enterprise's blog or microblogs e.g. Twitter
- In web data: Twitter only
General conclusions

- A significant difference in the results even if high correlation. Better match in some industrial classification and worse in some.
- Expensive service for us to buy.
- Quality check of the data still needed in high extent.
- Legal aspect of extracting open data?
- Still interesting if the services improve in coming years or if we could do it ourself. Could be just used as a help when doing quality check of normal survey data!
- Possible data to get in our Statistical Business Register when collecting normal business data?
Thank you for listening!
Questions?

bjorn.forssell@scb.se
Statistics Sweden