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

Björn Forssell

Statistics Sweden

facebook.com/statistiskacentralbyranscb @SCB_nyheter statistiska_centralbyran_scb www.linkedin.com/company/scb

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 usuage for 2017:

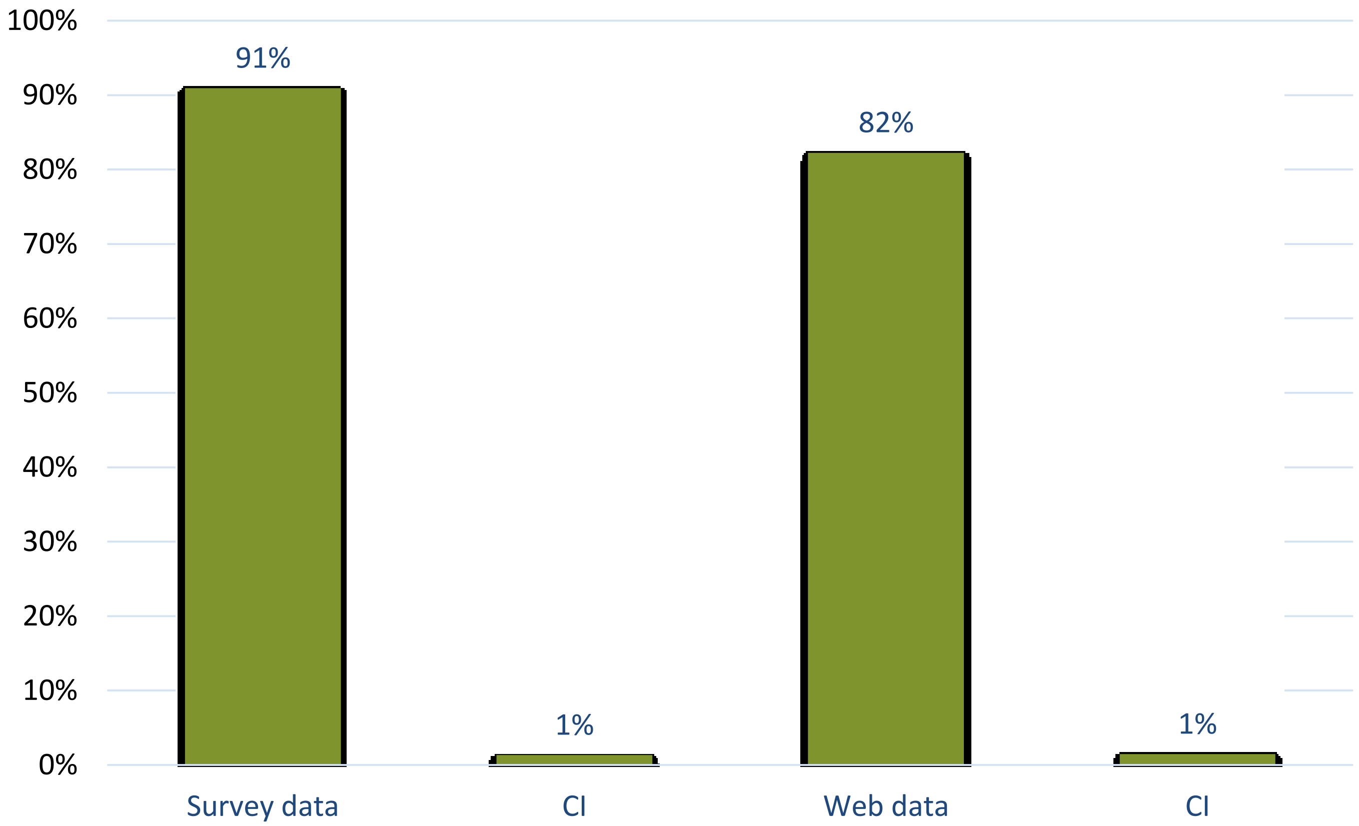
- Share of companies with website
- Use of social networks
- Use of blogs and microblogs

- 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

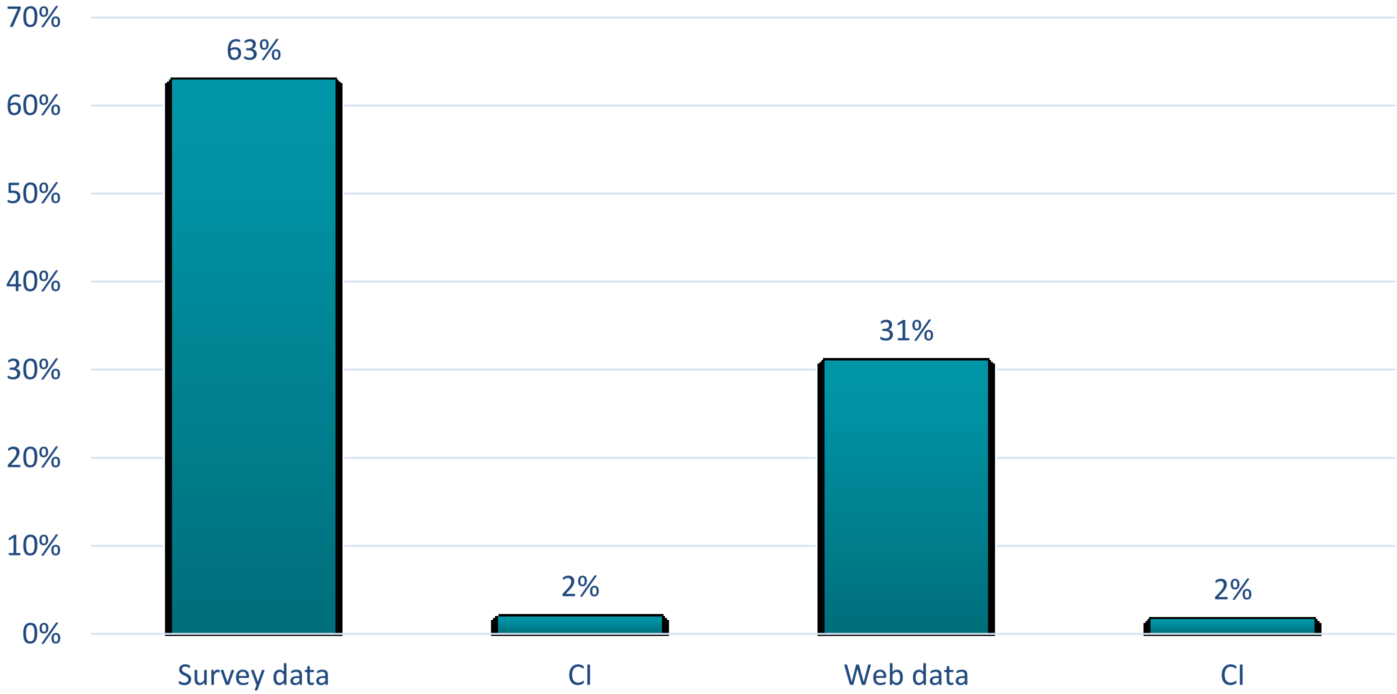


Web page after industry

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

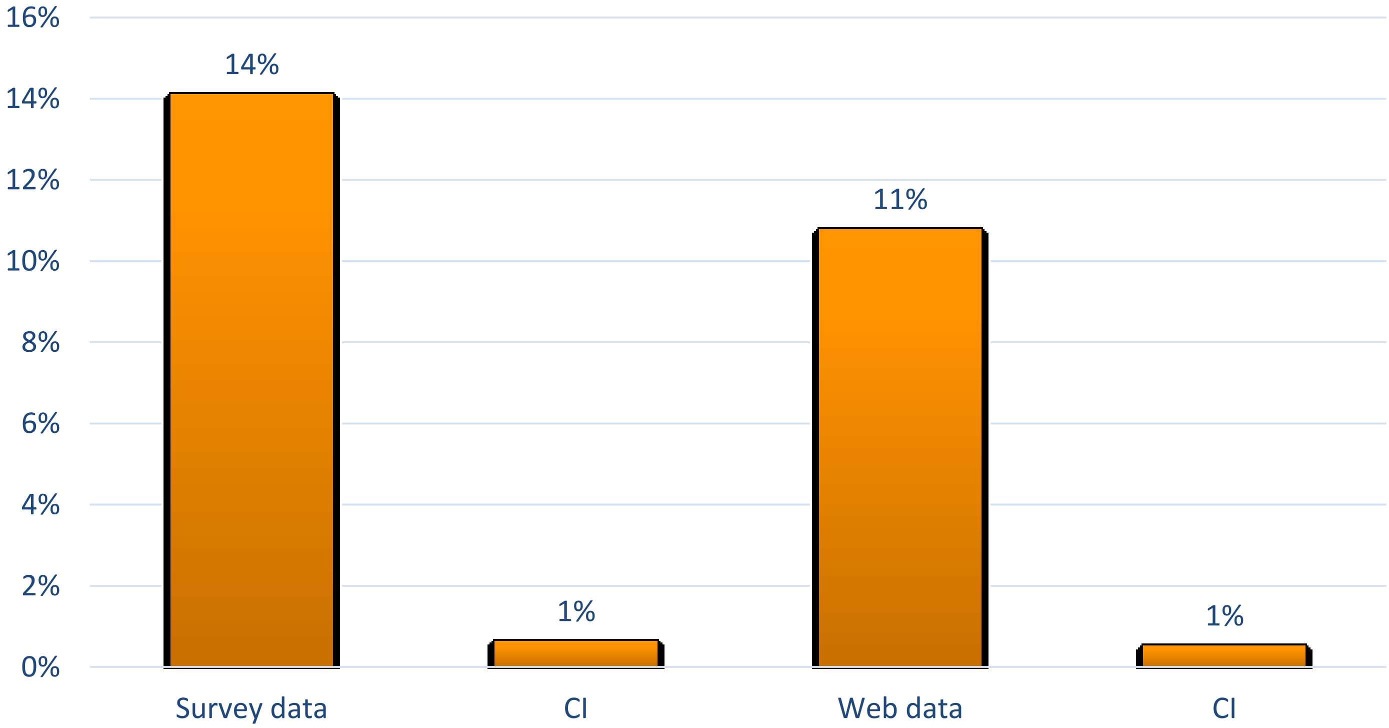
Results - social networks

- 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

facebook.com/statistiskacentralbyranscb @SCB_nyheter statistiska_centralbyran_scb www.linkedin.com/company/scb