Data integration challenges ICT survey and web scraped data from enterprise websites

National Institute of Statistics, Italy

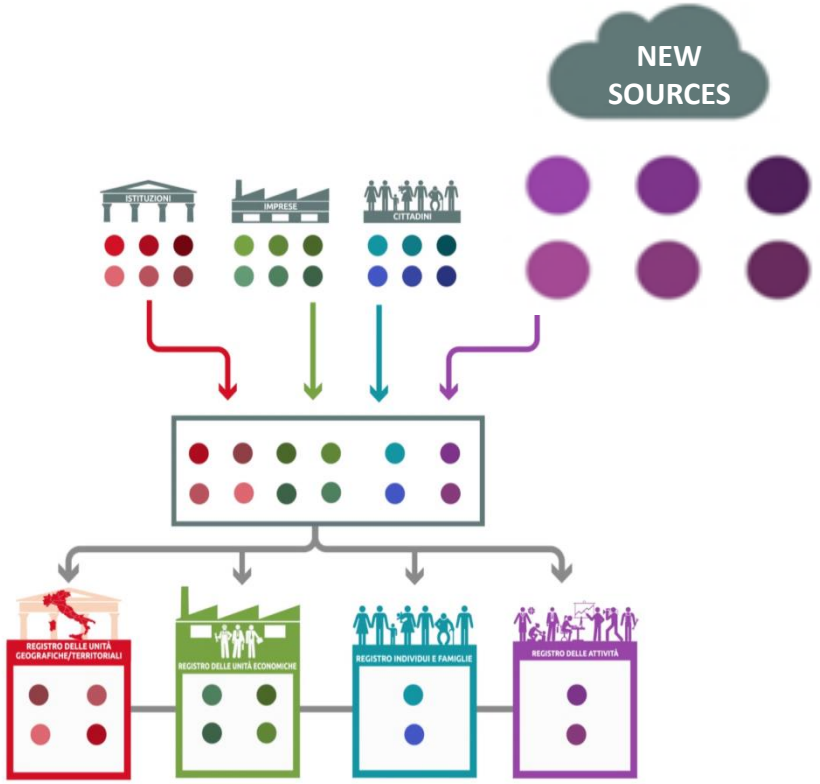
Integration and inference: a paradigm shift?

BIG DATA use is a part of a thorough change in the whole statistical production process.

VERACITY + VALIDITY
directly affect the accuracy of the estimates.

MULTISOURCE APPROACH
combined use of survey, administrative and BD sources

PARADIGM SHIFT
of statistical inference



A case study: the ICT survey

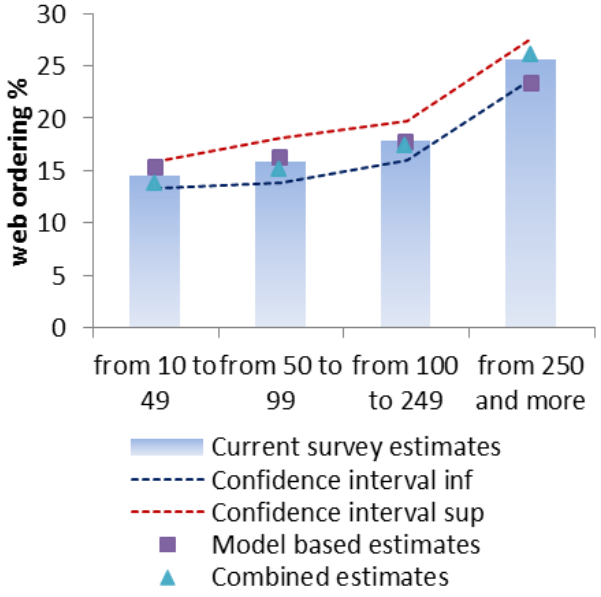
5 phases

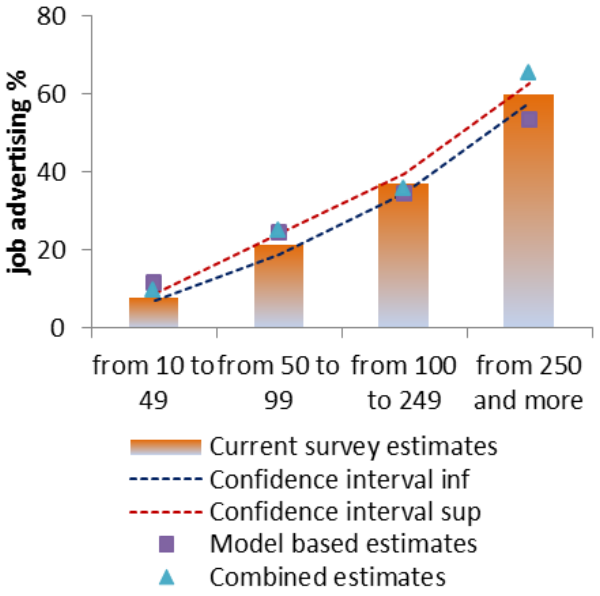
1 LIST OF URLs
2 WEB SCRAPING
3 TEXT MINING
4 MODEL FITTING

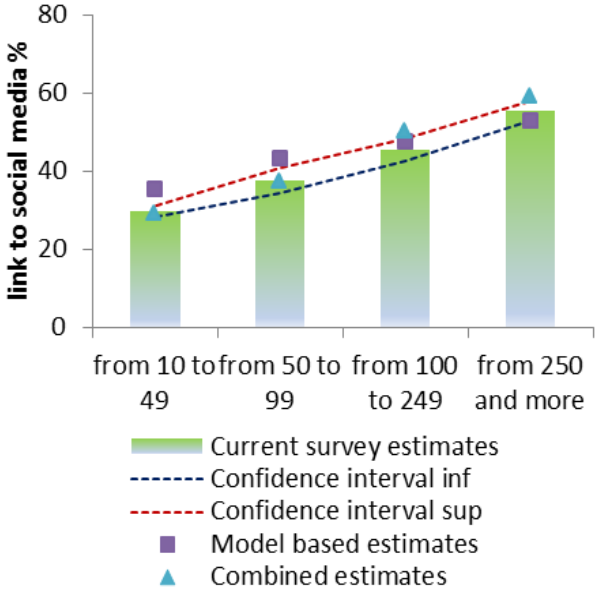
DATA COLLECTION AND PROCESSING

5 ESTIMATION

A case study: the ICT survey







Open issues and concluding remarks

- Quality is an open issue
- The simulation exercise shows that the combined estimation procedures are competitive with the traditional one
- We are still experimenting

Data integration challenges ICT survey and web scraped data from enterprise websites

National Institute of Statistics, Italy

DGINS
Bucarest, October 10th 2018

Istat | Istituto Nazionale di Statistica