Alternative data collection methods -

focus on online data

Presentation prepared by

Ragnhild Nygaard, Statistics Norway

for the UNECE/ILO Meeting on CPIs, Geneva, 2.-4. May 2016

Statistisk sentralbyrå
Statistics Norway

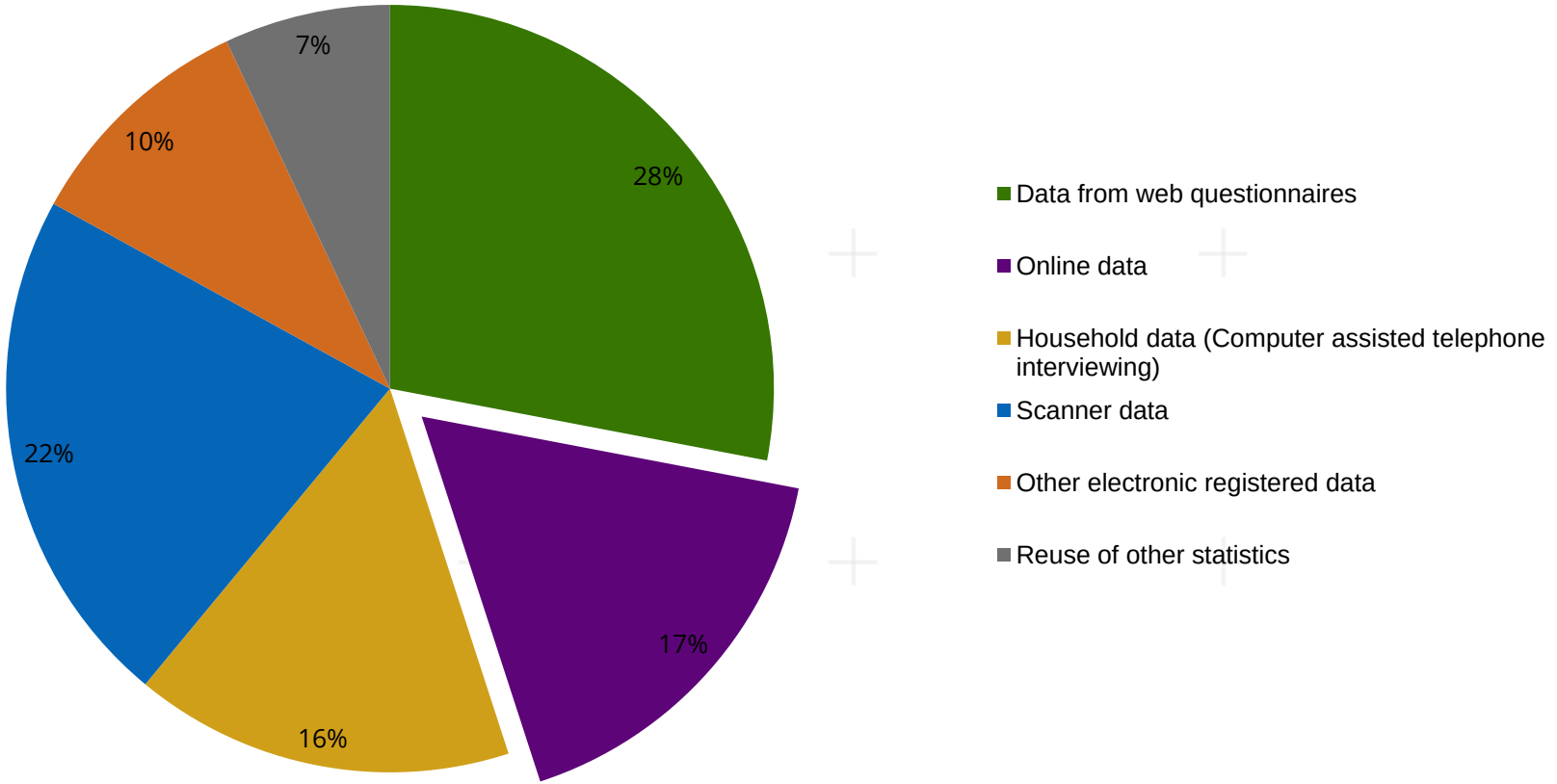
Contents

- Data sources and data collection methods in the Norwegian CPI
  - Major changes over time
- Online data project
- Building a production system for automatic extraction of online data
- Web scraped data for home electronics
- Big data classification and calculation

Data sources and methods in the Norwegian CPI

- Traditional data collection methods
  - Web questionnaires, e-mails, telephone interviews, «copy paste» of prices online
- Data sources and methods gradually changed over the last decades;
  - Technological progress
    - Today much easier access to large amount of electronic data in different formats
  - Globalisation
    - Changed consumer pattern, establishments operate across borders
  - Structural changes
    - Increased market concentration (fewer and bigger units)
  - Reduction of the response burden
  - Aiming at increased efficiency and quality

Data sources in the Norwegian CPI



Online data project

- Increased focus on scanner data and online data in many European countries and in Eurostat
- Statistics Norway just finalized the first online data project (2 year project)
- Online price data - a second best option to scanner data
  - Large amount of data
  - Unstructured data
  - Lack of quantity information
  - Product-offers vs transaction prices
- Online data project has looked into;
  - The general e-commerce
  - Price setting strategies
  - Set up costs, regular maintenance costs and legal issues related to web scraping
  - Classification and calculation based on big data

Automated extraction of online price data - web scraping

- Statistics Norway is building up experience within web scraping in the Price Statistics unit
- Automated extraction is partly used for collection of airline fares and for the prices of dental services
- For goods - an experimental production system for automated extraction of data
    - Not implemented into the CPI
- Important to make proper cost-benefit analyses before starting to build a web scraping system for CPI production

Automated extraction of online price data - web scraping II

- Can be quite effective for extracting price information from different consumer portals;
  - Many prices per web site
    - In our case for instance; prices of dental services
- Less effective if few prices per web site
  - In many cases national prices for specific services easy available online
  - Can be easier to just collect the prices manually

Experimental production system

- Gradually built a robust production system of automated extraction of data
- More than 1 year of comparable online data
- System based on free software which can easily be downloaded from the internet (import.io)
- Make daily download of prices and price related information from 4 online retailers registered in Norway with highest expenditure figures
  - Most sold products
  - Home electronics and personal care products

Experimental production system II

- Made gradual improvements to reduce the maintenance costs
  - Built a separate Java program outside the import.io environment that communicates with the robots
  - Important to automatically control all data files
  - Data will be missing due to e.g. server failure or changed URLs and this must be discovered at once – too late afterwards

Experimental production system III

- Build-in checks
  - Automatic retries if import.io server fails
  - Sizes of the data files are compared
  - Look for missing variables
- Data stored into csv-files
  - Information added (predefined codes, date etc.)
  - Data cleaned (e.g. remove unnecessary signs)
- For a trained import.io user setting up or train a new robot is not very time consuming (no programming skills needed)
  - But the maintenance costs of the system is of course related to the number of web sites scraped or robots used

Can we use the daily collected data of home electronics?

- Extensive information of characteristics available online
  - ..but not structured
- One way of using the data is to imitate traditional data collection methods
  - In our case that would be to act as a respondent and price variants of certain representative items based on traditional methodology
- Another way of using the data is to use *all* the data scraped
- How to classify all the product-offers?
  - We started by classifying the products into product groups identified through the URLs (àla COICOP6)
    - "freezers", "TVs", "laptops", "refrigerators" etc.

Can we use daily collected data of home electronics? II

- Further stratification into more detailed homogenous product groups or segments must be made
  - Make use of the product codes - splitting product codes
    - For instance: “KSV|29|NW|30” or “SONXPZ|2|BK”
  - Make use of the product texts - unsupervised machine learning
    - Try to find some hidden structure in the unstructured data
    - Data is clustered into different detailed segments based on similarities in the product text
  - Also tested further breakdown into price segments to represent simple, medium and advanced models
    - In order to come closer to “similar” products
    - Based on the idea that models with different price level are available in the market at the same time, price differentials will often reflect quality differentials

Can we use daily collected data of home electronics? III

- Monthly indices – arithmetic mean of the daily extracted data
- Made some test calculations
  - Monthly chained matched model approach (match of identical product codes) with no replacements
    - Product groups showed, as expected, a clear downward bias
    - In many cases we see an high introductory price and a low price on it way out of the market
    - Products with short product life span – on average the same product code is extracted for a period of 6 months
  - Stratification methods - geometric mean of the prices based on detailed homogenous product groups
    - Direct comparisons - imply that we assume minor quality differences within the same segments

Can we use daily collected data of home electronics? IV

- Use of hedonic method is challenging which would require structured data of price-determining characteristics
- Structured scanner data of home electronics could open up for other methods ahead
- Work in progress

Can we use daily collected data of home electronics? V