Automatic data collection in the Israeli CPI: measuring the sharing economy in the sector of short term rentals and web scraping for flights and hotels

Merav Oren - Yiftach meravo@cbs.gov.il

הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية

INTRODUCTION

The focus of this project is to include the Sharing Economy (collaborative economy, access economy, peer to peer economy) in the framework of our CPI. This economy is based on the sharing of products, services and properties that are under-utilized, between private costumers. The sharing economy is the market that connects people who need a particular product or service to people who have unused products or a surplus of time or skills. This type of economy is beginning to generate interest in statistical offices, and initial steps are being made to estimate the extent of its activity in the official statistics.

This project examines two issues that improve the information produced by ICBS, in the CPI:

1. The possibility of collecting, automatically, quality statistics on prices from the Internet as an important data source for official statistics, using data tools, such as "web scrapers" to navigate websites to extract the content from them. The main advantage of this method is that it can be used to collect large amounts of data in a short period of time, which is the first attempt at the CBS to automatically collect data from Internet sites for the CPI.

2. To estimate the Sharing economy: A first step was made to evaluate the scope of the sharing economy by simulations of automatic data collection on the Internet. The sharing economy is believed to have been growing rapidly in the last decade and is yet to be estimated in a methodological process. Consumption of accommodation and short term rentals are some of the major services in the sharing economy therefore the focus is on Airbnb services, where people advertise and rent short-term accommodation.

Eventually, our ambition is to implement the methodology in the Israeli CPI.

AUTOMATIC DATA COLLECTION WEB SCRAPING

Automated data collection on the Internet can be integrated into existing activity in the statistical offices on five aspects:

1. Completion of data collected through surveys - may improve the data that is collected manually in terms of speed, frequency accuracy and quality.
2. Replacing collection of data from manually to online.
3. Replace collection by survey.
4. Collecting completely new information for the purpose of producing new statistical information that has not yet been produced like sharing economy.

The advantages of automatic web collection:

Efficiency, Speed, Frequency, Scope, Quality, Expand the information on each item,

Reduce the: workload, manpower costs and the burden on responders

Disadvantages and Challenges in Automated Collection:

1. Initial investment: the implementation requires appropriate skills and software.
2. Web Scraping is limited only to retailers with web sites.
3. There are websites that use blocking technologies to prevent Web Scraping.
4. Lack of manual checking.
5. Web site changes: A Web Scraping robot will not be able to function in a case of changes in the structure of web pages and it will be necessary to reprogram.
6. Site content may be specific to IP.
7. Automatic collection process includes all the products on the website, including products that are not so popular.
8. it is not possible to go back in time to recover the data.
9. Data processing maintenance and storage costs may increase

Experimental price collection on flights and hotels:

The purpose: to introduce a method of collecting prices for flights and hotels via web scraping. This method will allow:

- to represent the increasing number of online users.
- Extend significantly the amount of items for the CPI.
- Increase the frequency of reflecting price volatility during the month

Flights: In web scraping, tickets are booked up to a year in advance at 21 popular destinations for Israeli flights in three time periods: weekdays (4 days), full week (upcoming destinations) and weekly airfare (for distant destinations). The prices will be compared to prices we receive from the airlines themselves for the coming year. We will examine whether the trends of the two sources are similar throughout the year.

Hotels: we will collect all the prices from Israeli web sites, for different types and dates of reservations: weekends, weekdays and holidays. The prices that will be received and the trends of their changes will be equal to those obtained in the field.

We will examine what is the optimal frequency that should be conducted: on a daily basis or weekly one. Are the data of good quality and comparable? What is the method for replacement?

We will examine the stability of the web sites

WEB DATA COLLECTION PROCESS AIRBNB

The major steps of the price collection process are as follows:

1. Choosing a web scraper and learning how to use automatic data collection tools on the job. Two programs were selected simultaneously: Octaparse and ParseHub.
2. Locating the properties: locating all properties appearing on the Airbnb site within Israel. Then filtering out price levels and subtracting duplicate records.
3. Collecting the information: collect all the details from all pages of the properties collected in the previous stage.
4. Create an ID code for each property and an ID number for each host.
5. Optimizing the data: reclaiming the name of the locality and subtracting text from numeric fields.
6. Link to files with geographic information.
7. Collecting the total price - beyond the base price that always appears at the top of each rental property, the total price could be for a specific date, or including all other additions and discounts.
8. Update the file : add new properties and subtract those that do not exist from the site.
9. Gathering information about guests hosted on Airbnb properties in Israel. This is relevant for other statistics on tourism and not necessarily for the CPI.
10. Data processing.

24,241 properties were found for short-term rental in 673 localities According to the automatic collection method.43.6% of the properties are in Tel-Aviv

Each property has a base price per night for the guest. This price doesn't includes additional service fees, like: cleaning fee, specific dates and discount for lengthy periods.

The price is affected by the property's type, size and location

Average night base price per property is $172.2. Average base price in Jerusalem is higher than the price in Tel Aviv $175.9 compared to $160 respectively. According to the six main cities comparison, we can see differences in the average total price. The highest price is in Herzeliya and the lowest is in Haifa. The Prices in Jerusalem and Netanya are influenced by September and April Holidays. In Eilat the prices are high in August and December and The prices in Tel Aviv in May 2019 are influenced by the Eurovision Song Contest.

FIGURE 1: Number of properties per city/area

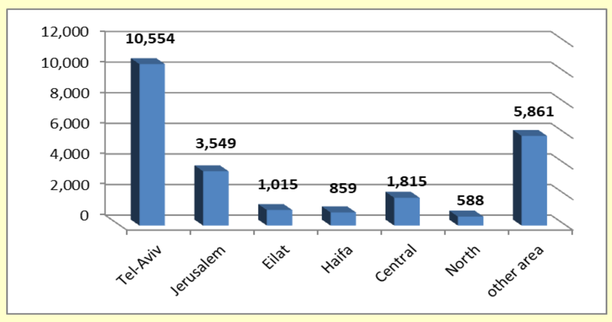


FIGURE 2: Guests by continent

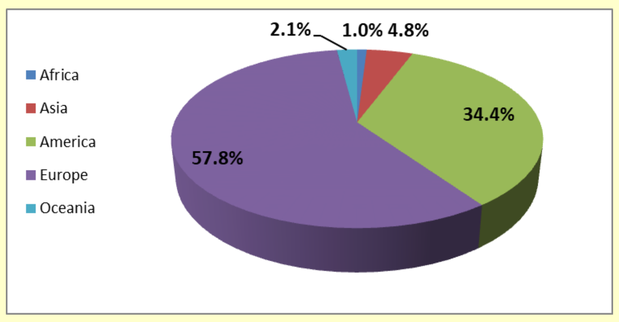


FIGURE 3: Price trend for several cities

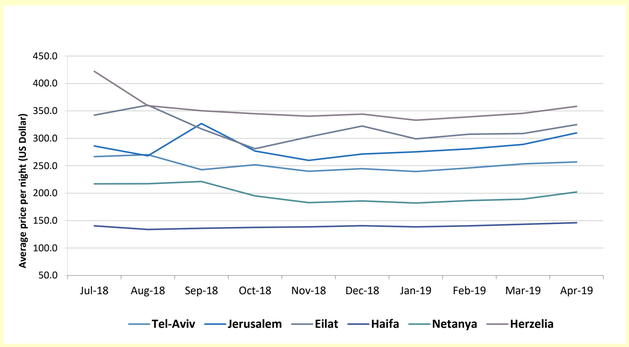
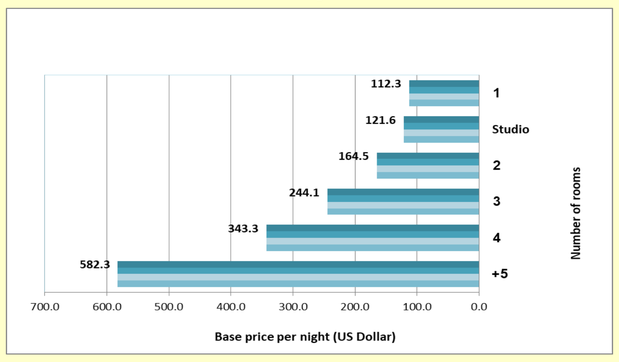


FIGURE 4: Price level by number of rooms



SUMMARY AND CONCLUSION

The experiment on measurement of sharing (collaborative) economy illustrates the possibility of producing reliable estimates based on data from the internet on short-term rental of properties. In addition, the experience accumulated in this work will be used in the future to implement automatic collection in the consumer price index for services that are consumed over the web like flights and hotels.

The number of dynamics transactions within new economies are growing daily, therefore the importance of utilizing new sources of data collection for statistical offices is acute. It will be necessary to examine various methodological considerations, such as defining the investigated population, defining the characteristics or indicators to be examined, and more. All aspects must be tested: Technical, organizational, budgetary, methodological, data quality control, coverage, legal aspects, and the ethical dimension. In addition, how to properly use data from companies and commercial sites.

Utilizing Big Data sources to produce official statistics will require expanding the existing knowledge of practitioners, using modern technologies of artificial intelligence, machine learning, text mining, natural language processing techniques, and more.