Thesis proposal

Lynette Joosten

February 28, 2017

Abstract

In this document I describe my thesis proposal for the master thesis project of the master Information Studies: track Data Science. My thesis will be written during an internship at CBS, the Central Bureau of Statistics.

1 Personal details

My email `mailto:lynette.joosten@student.uva.nl`

My supervisors email `mailto:e.kanoulas@uva.nl`

The wiki on my github account `https://github.com/LynetteJoosten/Thesis/wiki`

2 Research question

Back in the days CBS computed the Consumer Price Index (from now on: CPI) by sending people to supermarkets with a grocery bag which they filled with random items. Of these grocery bags the average price was calculated, which became the CPI. Based on this first approach price models were devised which were used to calculate the CPI. Nowadays CBS gets a lot of data from supermarkets, including their weekly sales per product. The old models are still used to compute the CPI, while there might be easier methods. My research question will thus be focused on new methods to measure the CPI:
What is the best method to approximate the consumer price index of the CBS?
To answer this research question, I have defined the following subquestions:

- How is the CPI calculated at the moment?
- ?

3 Related Literature

One important source of information is *The Billion Prices Project* by MIT [7]. They have written a lot of research papers about experiments with the offline price index.
Using Google Scholar, the following articles seem relevant at a first glance: [8], [4], [1], [9], [3], [6], [5], and especially [2].

4 Methodology

4.1 Resources

This research will be focused on a data set provided by the Albert Heijn supermarket chain. Weekly the CBS gets a new data set from Albert Heijn consisting of all products, their price and the amount of products sold. This data set consists normally of .. rows and .. columns. It is already processed, so I will not have to do any preprocessing or cleaning of the data set myself. Python and R will be used to test the different approaches mentioned in the section about the research question. Scala or PySpark will be used if the data needs to be processed on a cluster.

4.2 Methods

The following methods will be used to approximate the CPI:

- Taking the average of all prices of all products.
- Taking a sample of 10, 100, 1000 and 10.000 products.
- Create a weighted model of the products and their prices.

4.3 Evaluation

The results will be evaluated in two ways:

- Can the new consumer price index show the same differences over time as the old consumer price index?
- Is the absolute number of the new consumer price index the same or relatively close to the old consumer price index?

5 Risk assessment

6 Project plan

Week	Deliverables
1	Literature research, answer subquestion 1
2	Set up experiment environment
3	Method 1: average of total prices
4	Method 2a: sample of 10 and 100 products
5	Method 2b: sample of 1000 and 10.000 products
6	Mid-term results
7	Method 3: weighted model
8	Method 3: weighted model
9	Evaluation of results
10	Conclusion and discussion
11	Hand in thesis
12	Defense

References

[1] Katharine G Abraham, John S Greenlees, and Brent R Moulton. “Working to improve the consumer price index”. In: *The Journal of Economic Perspectives* 12.1 (1998), pp. 27–36.

[2] Jan De Haan. “Generalised fisher price indexes and the use of scanner data in the Consumer Price Index (CPI)”. In: *Journal of Official Statistics* 18.1 (2002), p. 61.

[3] Robert C Feenstra and Matthew D Shapiro. “Introduction to” Scanner Data and Price Indexes””. In: *Scanner Data and Price Indexes.* University of Chicago Press, 2003, pp. 1–14.

[4] Jerry Hausman. “Sources of bias and solutions to bias in the consumer price index”. In: *the Journal of Economic perspectives* 17.1 (2003), pp. 23–44.

[5] William J Hawkes. “Reconciliation of consumer price index trends with corresponding trends in average prices for quasi-homogeneous goods using scanner data”. In: *International Conference on Price Indices, Voorburg.* 1997, pp. 16–18.

[6] Lorraine Ivancic, W Erwin Diewert, and Kevin J Fox. “Scanner data, time aggregation and the construction of price indexes”. In: *Journal of Econometrics* 161.1 (2011), pp. 24–35.

[7] MIT Sloan School of Management. *The Billion Prices Project.* 2017. URL: `http://bpp.mit.edu` (visited on 02/28/2017).

[8] Mick Silver. “ELEMENTARY AGGREGATES, MICRO-INDICES AND SCANNER DATA: SOME ISSUES IN THE COMPILATION OF CONSUMER PRICE INDICES”. In: *Review of Income and Wealth* 41.4 (1995), pp. 427–438.

[9] Jack E Triplett. “Using Scanner Data in Consumer Price Indexes. Some Neglected Conceptual Considerations”. In: *Scanner data and price indexes.* University of Chicago Press, 2003, pp. 151–162.