Summary of the doctoral thesis

Adam Juszczak

"Application of Scraped and Scanned Data in Inflation Measurement"

The thesis was written at the Department of Statistical Methods, Faculty of Economics and Sociology, University of Lodz, under the supervision of Professor Jacek Białek.

The main objective of the doctoral thesis is to discuss the advantages and limitations associated with the use of new data sources in measuring inflation. The author focuses on a thorough comparison of various index formulas and examines their potential application within individual alternative data sources, specifically evaluating a series of methods to reduce the measurement bias of inflation based on scraped and scanned data, such as methods for expanding the time window to determine multilateral indices, data filtering, elimination of atypical observations, and data aggregation by homogeneous subgroups of products and outlets of a given retail network. The calculations were performed in R using packages such as *PriceIndices* and *microbenchmark* and involved operations on three types of data sets: scraped data obtained from one of the most popular online stores selling clothing and footwear; two sets of scanned data in the *PriceIndices* package and the data set attached to the International Monetary Fund (CPI Manual) publication; and seven sets of artificially generated data by the author.

The thesis consists of six chapters, with chapters 1-3 being theoretical, chapters 4-5 empirical, and chapter 6 being simulation oriented. The first chapter briefly overviews inflation research and presents the history and characteristics of CPI (Consumer Price Index) and HICP (Harmonized Indices of Consumer Prices) indicators. The sources of price and consumption data are also discussed, and various types of inflation measurement bias are examined, including how the sudden change in consumption structure caused by the COVID-19 pandemic has affected measurement accuracy.

The second chapter introduces the idea of using scraped data as an alternative data source for measuring inflation. In addition to the definition of scraped data, it provides information on

Adam Juszczak

their origin, acquisition methodology, and advantages of use. The practical limitations and methodological challenges associated with implementing scraped data in practice are also discussed. Finally, index formulas used to calculate the dynamics of product prices using scraped data are presented.

The third chapter discusses the idea of using scanned data to measure inflation, following a similar format to chapter two. The definition, origin, acquisition methodology, processing, classification, and matching of this type of data are presented. The biggest methodological and application challenges faced by statistical offices aiming to use scanned data in inflation estimates are also described. Finally, index formulas used to calculate the dynamics of scanned product prices, with a particular emphasis on multilateral formulas, are presented.

The fourth chapter presents an empirical study conducted on scraped data obtained from one of the largest online clothing and footwear stores. In addition to the results of calculations for individual index formulas, the impact of calculation window extension methods and filtering methods on the results were also presented. The analysis of the data set obtained from web scraping regarding the prices of clothing and footwear confirms the problem of high product turnover, which was described in studies by the British statistical office in relation to this segment. In none of the subcategories of the analyzed set of clothing and footwear, the percentage of repeated ID codes in the first and last month of the analysis exceeded 10%, and on average it was just under 3%.

The fifth chapter discusses the results of analyses conducted on scanned data obtained from the CPI Manual publication and two datasets included in *PriceIndices* package. Differences between the results of individual index formulas for three analyzed data sets are discussed. The impact of individual filtering and time window extension methods on price index results is also presented. Methods of aggregating data by product subgroups and outlets were also examined for selected index formulas. The calculations were also supported by an analysis of the calculation time required to use individual data aggregation methods.

In the final, sixth chapter, a simulation of the impact of price and quantity variability (consumption level) on the values of individual index formulas was conducted, taking into account seven artificially generated data sets. The individual index formulas were also evaluated based on the time required for their calculation depending on the data set size. This chapter also presents a comparison of the resistance of index formulas to changes in the sample structure using the jackknife method.

Adam [illegible]

The research results indicate low differences between results for bilateral indices, especially between Jevons and Dutot indices. Very high differences are observed when compared to chain indices and multilateral indices - UTPD and GEKS-J. Differences exceeding 25 percentage points indicate a very large measurement bias associated with calculations on a group of non-homogeneous products.

Analysis of bilateral index formulas based on three scanned data sets showed that the Törnqvist index formula provides the smallest differences compared to the Fisher formula. The Walsha and Sato-Vartia indices also showed slight differences compared to the Fisher index. Similar trends can also be observed for the chain versions of these formulas. The chain Laspeyres and Paasche indices, on the other hand, show a very high tendency towards chain drift.

In the case of multilateral formulas, very similar values of the TPD and Geary-Khamis indices can be observed. Indices based on the GEKS formula, with the exception of the GEKS-J index, also show small differences relative to each other in the case of two analyzed data sets (this applies especially to the GEKS-AQU and GEKS-AQI indices). For the third data set with high seasonal fluctuations, differences between the formulas of the indices are significantly higher, reaching several percentage points, indicating the susceptibility of some multilateral formulas to the high variability and/or seasonality of prices. This conclusion is confirmed in the simulation chapter, which shows that high price volatility affects the diversity of multilateral index results, even if the quantity volatility is very low.

For both data sources, both scraped and scanned, verified filtering methods indicated good index results with a price filter applied (1st and 99th percentiles). It eliminated a very small number of observations from the studied product groups while reducing high price fluctuations for some indices (e.g. GEKS-J) resulting from the influence of atypical values. Positive effects can also be observed in the case of indices with a price filter eliminating products whose price has risen more than twice or fallen below 50% of the previous month's price value. Filters based on quantity or sales value volatility should be evaluated less positively in the context of this analysis, mainly due to significantly greater interference with the sample. After their application, the percentage of unique products in the analysis decreased by up to 50%.

Analysis of methods for expanding multilateral formulas for scanned and scraped data showed a significant difference between the FBEW method and other methods for expanding the calculation window. Splice-type methods showed moderately similar values in most of the

Adam Sagan

analyzed cases. The FBMW method gave results more similar to splice methods than the FBEW method.

Evaluation of index aggregation methods showed significant differences between the double aggregation method (by subgroup and by an outlet with the Fisher formula) and aggregation only by subgroup (with the same aggregating formula). Much lower differences can be observed in comparison to aggregation only by outlet. Taking into account the several times lower calculation time for price indices aggregation only by outlet compared to the double filtration method based on the analyzed data, it can be concluded that for larger data sets from a relatively small number of stores, this aggregation formula may be the best compromise between improving calculation accuracy and time needed for finalization.

Calculation time is also an important factor in choosing index formulas themselves. As simulations conducted on generated data sets show, even in the case of 1000 products, significant differences in the calculation time of considered price indexes can be observed. Among multilateral formulas, GEKS formulas, especially CCDI index and GEKS-L, GEKS-GL, and GEKS-J indexes, are characterized by low calculation time. Calculations using the Geary-Khamis formula took about 4 times longer, and calculations using the SPQ index took 12 times longer (this index also had the lowest resistance to changes within the sample). The TPD index performed relatively poorly, requiring on average almost 400 times more calculation time than, for example, the CCDI index, due to its formula based on calculations performed on inverse matrices.

26.04.2023

Adam Jagoda