Time series modelling and index decomposition as efficient tools for CPI validation

Keywords: *Consumer Price Index, COICOP, time series models, index decomposition*

1 Introduction

The Consumer Price Index (CPI) measures the rate of price inflation as experienced by consumers and is widely used as a general measure of inflation. The CPI is compiled and published by National Statistical Institutes (NSI) each month (or each quarter in some countries). The compilation process is embedded in a hierarchical structure of consumer goods and services, known as COICOP, in which individual items have to be classified by NSIs.

The current COICOP classification system (ECOICOP) consists of 4 so-called "structure levels", with 13 "divisions" at the 2-digit top level, which are subdivided into 303 "subclasses" at the most detailed 5-digit level. NSIs publish price indices at each of these four levels each month, so that the 5-digit level, als called COICOP-5, is the lowest publication level. For example, division 03 Clothing and Footwear is distinguished into three subclasses for garments at the most detailed level (men, women and infants). These subclasses are still very broad, so that NSIs usually refine such subclasses by introducing more detailed product categories below COICOP-5, which are sometimes referred to as the COICOP-6 level. On the other hand, we also find product categories at COICOP-5 level that are rather tightly defined. A COICOP-5 like Chocolate in division 01 may not need to be refined into more detailed subgroups.

The CPI compilation process starts by calculating prices for products that are homogeneous in terms of quality characteristics, which are combined in order to calculate price indices at COICOP-6 or COICOP-5 level. These COICOP indices are weighted and combined to produce indices for each higher COICOP level and eventually for the overall CPI. Index compilation above COICOP-5 level is tied to European regulations to which all EU member states must comply. Within COICOP-5 level, member states have a certain amount of freedom to decide on an index method and related aspects, such as defining homogeneous products.

Price collection for consumer goods and services has rapidly evolved over the past 20 years. NSIs are focusing on acquiring large electronic data sets, like transaction data from retailers or by scraping prices and product information from websites. These developments have motivated research on index methods that exploit the higher degrees of dynamics in these data sets in a better way than traditional bilateral index methods. This had led to a broad range of studies on multilateral index methods. These methods are not only tailored to the dynamics of product assortments, but they also allow to compile index series that are transitive, that is, do not suffer from drift that results from month on month index chaining that is typical for weighted bilateral methods like Laspeyres, Paasche and Fisher indices.

Statistics Netherlands was the first NSI in Europe to introduce a multilateral method in its CPI (the Geary-Khamis method, see [1]). Multilateral methods have a higher degree of technical complexity than bilateral methods and are capable of processing large data sets. The increased complexity brought about by these developments calls for a higher need of sophisticated methods that support CPI personnel in their task of interpreting the results. This is an important step between index compilation and publication. It is desirable to understand why an index changes at a certain rate from month to month or from year to year, in the first place to gain insight and confidence in the results and secondly to answer questions from the media.

This paper suggests a two-step approach to this so-called "index validation" stage, which is summarised in Section 2. A time series modelling approach is suggested for detecting anomalous index movements at COICOP levels. The results are used in the second stage to compute contributions from each individual product to COICOP index changes, which identifies the main drivers behind large index changes found at the first stage. A prototype tool is briefly illustrated in Section 3. Section 4 concludes.

2 A two-stage approach to index validation

2.1 Stage 1: Time series modelling of COICOP indices

CPI personnel are involved with checking and interpreting price indices at different levels of aggregation, which is a very time consuming activity each month. A set of rules are defined, which serve as criteria for deciding whether to analyse an index in more depth or not. Commonly used bandwidths for index change increase the efficiency of this index validation stage, but also have drawbacks. For example, indices that do not change between two successive months are not likely to be analysed, since the rate of change will never fall outside a predefined bandwidth. Time series modelling can be used to overcome such drawbacks, while it can also further improve the efficiency of the index validation stage.

Monthly compilation and publication of price indices results in a monthly updated non-negative series $\{P_{0,t} : t \in \{0, 1, ..., T\}\}$ of index values $P_{0,t}$, where 0 denotes the initial month of the series and T the current month. Price indices are ratio numbers in the sense that an index expresses a value with respect to a reference period, in this case month 0. This means that $P_{0,0} = 1$, as there can be no change between the same pair of months.

We suggest to use time series modelling for indices at COICOP-5 and higher levels in a top-down fashion, starting from the 2-digit division level. Within a time series framework, price indices can be considered as realisations of a stochastic process. Autoregressive Moving Average (ARMA) models are a classical type of time series models, which we consider in this study. The index $P_{0,t}$ is then considered a random variable for the price index of some COICOP in month t with respect to month 0. Price indices according to ARMA models can be formalised as follows:

$$P_{0,t} - \alpha_1 P_{0,t-1} - \cdots - \alpha_p P_{0,t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{1}$$

where the α_i are the parameters of the autoregressive part of the model, the θ_i are the parameters of the moving average part and the ε_i are error terms. The error terms are generally assumed to be independent, identically distributed variables sampled from a normal distribution with zero mean. Extensions of the ARMA model are ARIMA and SARIMA, where the first also includes differences between price indices in successive months, while SARIMA also includes a seasonal component. These extensions are considered in this study as well.

2.2 Stage 2: Product contributions at COICOP-5 level

The first stage typically results in a number of product categories at COICOP-5 level which require further investigation. The second stage is concerned with deepening the understanding behind the question why an index at COICOP-5 level falls outside a bandwidth as observed at the first stage. Product contributions to month on month (mom) or year on year (yoy) indices at COICOP-5 level can be calculated, which reveal the main drivers of mom or yoy indices.

Methods for deriving product contributions obviously depend on the index methods used to compile a CPI. Statistics Netherlands makes use of two methods for most of its CPI: the Geary-Khamis (GK) method for transaction data and a weighted Dutot type method for web scraped data and traditional forms of price collection (sample-based approaches).

Calculating product contributions for a weighted Dutot index is rather straightforward. Such an index is a weighted sum of indices of average prices of homogeneous products in the current period and the comparison period. The index formula therefore automatically yields a contribution of each product to an index, which combines a product weight with the change in product prices between two periods.

Suppose we have a COICOP that consists of a set K of homogeneous products. Each product may contain one or more items with a unique product ID (e.g. a barcode), but which are of the same quality. Let w_k denote the weight of product $k \in K$ and let $\overline{p}_{k,t}$ be an (unweighted) average price of product k in month t. The weighted Dutot index for the set of products K in month t with respect to month 0 can be written as follows:

$$P_{0,t} = \sum_{k \in K} w_k \frac{\overline{p}_{k,t}}{\overline{p}_{k,0}} \tag{2}$$

If we are interested in the product contributions for the index between months 0 and t, then the contribution of product k is simply equal to $w_k \overline{p}_{k,t} / \overline{p}_{k,0}$. Calculating product contributions for multilateral methods is more difficult because of their more complex nature. We are developing methods for the GK method, which yield contributions to the change in expenditure, quantities sold and price. Product contributions to price change are calculated as a normalised difference between the contributions to expenditure and quantity change. Expenditure and quantity are both additive measures, so that product contributions to expenditure and quantity indices can be written as additive expressions as well.

3 Prototype tool

A prototype tool has been developed, which contains modules for the two stages described in Section 2. The module for stage 1 includes different types of time series models, amongst which ARIMA and SARIMA. The tool, named Valinda, produces a 'forecasted' index in the current month, together with other output like measures of model fits to COICOP indices for the best fitting model and confidence intervals. The published index value can be compared with a confidence interval to decide whether to analyse a COICOP further or not. An example of the output of stage 1 is shown in Figure 1. The user interface uses an R-shiny dashboard.

A result of stage 1 may be that further analysis of an index at COICOP-5 level is needed. The published index at this level is a weighted sum of the indices for different

Figure 1: Example output of stage 1, which shows the results of two time series models at overall CPI level. Red = CPI, black = trend line.

retail chains. These indices constitute the starting point of stage 2, in which product contributions to indices of retail chains are calculated. Figure 2 shows typical output for indices calculated with the GK method. The left graph shows mom per cent changes in overall expenditure, quantity and price, while the graph on the right shows product contributions to price change in a specific month. The products with the largest positive and negative contributions can be singled out and inspected further.

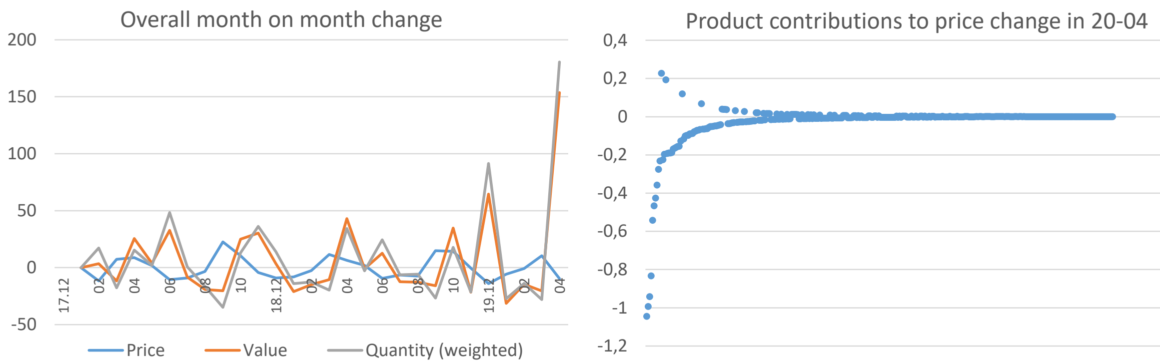

Figure 2: Output of stage 2, showing mom percent changes to overall expenditure, quantity and price (left) and product contributions (in pp) to price change (right).

4 Future work

First experiences with the tool look promising, as the tool performs quickly and has the potential to automate the index validation process to a high degree. Future work will include further development and testing of time series models and index decomposition methods for mom and yoy indices for all COICOPs.

References

[1] Antonio Gavino Chessa. A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 2016(1):49–69, 2016. https://ec.europa.eu/eurostat/cros/content/new-methodology-processing-scanner-data-dutch-cpi-antonio-g-chessa_en.