Forecasting Changes in the Fruit & Vegetables Index Using Data from the Retail Price Database

Dor Goldenberg and Yonatan Rosen*

Summary

The Consumer Price Index (CPI) is released on a monthly basis by the Central Bureau of Statistics, at a lag of 15 days after the end of the indexed month. Toward the end of each month, the Bank of Israel publishes a forecast of the CPI for analysis purposes. Although its weight in the CPI is low, the fruit and vegetables component is extremely volatile and therefore potentially makes a significant contribution to errors in the CPI forecast. In recent years the Bank of Israel has begun collecting data on the prices of products sold in the outlets of Israel's major food retailers. This information is reported by law[1], is transmitted from the major retailers' special websites, and is entered into a unique database created by the Bank known as "the Retail Price Database." This study presents a forecast of the fruit and vegetables component of the CPI based on the prices that were entered into the Retail Price Database. A combination of this forecast and other forecasts that currently exist at the Bank of Israel improved predictive quality by 23 percent in the tested period. This project, which uses Big Data to improve existing forecasting methods, is the first of its kind at the Bank and constitutes a pilot for more extensive use of Big Data to support policy making.

* Information and Statistics Division of the Bank of Israel.

[1] The Promotion of Competition in the Food Industry Law.

1. Introduction

The main goal of the Bank of Israel's monetary policy, as of that of many other central banks in the world, is to maintain price stability. The Bank of Israel is committed to attain the inflation target set out by the government in consultation with the Bank of Israel Governor, and bases itself on the CPI that the Central Bureau of Statistics (CBS) publishes monthly. The CPI measures the change of the prices in a "fixed basket" of products and services over time. This basket includes several main components, secondary components, and subcomponents. The main components include food, fruit and vegetables, clothing and footwear, housing, and health. To support the management of its monetary policy, the Bank of Israel produces a monthly forecast of the CPI and its components 15 days in advance of the release of the CPI by the CBS.

The category of fresh fruit and vegetables accounts for approximately 2 percent of the CPI[2]. The importance of a precise forecast of this component of the CPI may seem negligible due to its low weight in the CPI, yet because fruit and vegetable prices tend to be extremely volatile and difficult to forecast, errors in the forecasts of this component are occasionally much larger than errors in the forecasts of other CPI components. As a result, the contribution of a forecasting error in the fruit and vegetables component to the forecasting error in the CPI significantly exceeds this category's share of the CPI and reaches 0.05 percentage points on average and 0.25-0.1 percentage points in extreme cases[3].

The use of new data sources, and specifically Big Data sources, in forecasting the official CPI released by the Central Bureau of Statistics has become common in recent years. A prominent example is the Billion Prices Project (Cavallo and Rigobon, 2016), a joint academic project of researchers from Harvard University and MIT. This project, which uses high-frequency Big Data to calculate official consumer price indices of a broad group of countries, focuses on the collection of prices of products sold online, for two main reasons: First, online prices are typically highly available data that can be collected relatively easily at high frequency and low cost. Second, the underlying assumption is that these prices constitute a good approximation of official price indices that are typically based on the prices in brick-and-mortar stores. In this study we effectively forgo the need to collect online price data by using the Retail Price Database, which provides data on fruit and vegetable prices in physical stores at a daily frequency[4].

This study joins the pilot project that was recently conducted by the Sveriges Riksbank (Central Bank of Sweden), which included an effort to forecast changes in fruit and vegetable prices in Sweden based on prices collected online (Hull et al., 2017). The results of the pilot program are encouraging and indicate that online prices may be useful in forecasting the fruit and vegetable component.
In this study we describe the Retail Price Database and its use in forecasting the rate of change in the fruit and vegetable component of the CPI. We begin with a description of the data and the challenges we faced in working with this database, and proceed to describe the calculations and results. We conclude with a discussion of directions for the path forward. Israel to expand and broaden its analyses of Israel's credit market, and to monitor the trends and emerging threats in this market.

The Bank of Israel's Information and Statistics Department, which is responsible for collecting, producing, and making available the economy's financial statistics, manages the anonymized database ("the Statistical Credit Database") that the Bank of Israel uses to perform its functions. This study describes the scope of the Statistical Credit Database, the information anonymization process, and its diverse uses, notes the challenges and issues that emerged in the process of developing the database, and reviews similar databases around the world.

[2] The fruit and vegetables component includes frozen, pickled, and preserved vegetables, and dried and preserved fruits—components that constitute less than 1 percent of the CPI.
[3] The contribution of the error is defined as the forecasting error of a component multiplied by the component's weight in the total CPI.
[4] In effect, the Database also includes information on "online" outlets.

2. The Data

A. The target variable: the CBS Fruit and Vegetables Index

Every month the CBS collects the prices of fruits and vegetables according to the household consumption basket (which is sampled in the Household Expenditure Survey). On the basis of these prices, the CBS constructs the Fruit and Vegetable Index that is used to calculate the rate of change of the fruit and vegetables component of the general CPI[5]. The prices are sampled by surveyors that the CBS sends to various" layers" of vendors: greengrocers, retail chains, and the markets. The Fruit and Vegetable Index is constructed from the prices that are sampled in the various layers. In this Index, each fruit and vegetable is assigned a weight that corresponds to their weight in the Household Expenditure Survey. The Index also takes into account various aspects of seasonality.

B. The Retail Price Database

The Kedmi Committee was established in the summer of 2011 following the "Cottage Cheese" Protest with the aim of examining the degree of competition in the food market. In response to the Committee's findings, in 2014 the Knesset passed the Promotion of Competition in the Food Industry Law ("the Food law"), requiring the major retail chains to display the prices of all the products they sell in all their outlets every day. Consequently, the Bank of Israel decided to establish a database, known as the Retail Price Database ("the Database"), to hold the data published by the chains in a single repository that would facilitate data processing and analysis.

3. Challenges

A. Missing data

The Database includes only the prices of the products that are sold in the retail chains. As such, the Database does not include all the prices that are used to construct the Index, which also include the prices of fruits and vegetables in stores that do not appear in the Database, such as markets, neighborhood grocery stores, and greengrocers. Therefore, if the other layers behave similarly to retail chains, a change in Database prices will reflect the total change that is reflected in the Index. However, if prices in the markets decline while retail chains raise their prices, the change in the Database will only reflect the rising prices in the retail chains but not the reduction in prices in the markets. Since there is some degree of competition between the layers, it is reasonable to assume that their activities are positively correlated. However, the diverging behavior of prices in the other layers constitutes a source of error in the forecast.

B. Historical depth

The CPI is published monthly. Because the Database only contains four years of data, the Bank has data on both the Fruit and Vegetable Index and the prices of the products that comprise the Index for only about 54 observations[6]. Due to the limited amount of data, it is not possible to validate assumptions in a parametric model using out-of-sample (test) data, and overfitting becomes a genuine concern. As a result, at this stage it is not possible to practically estimate a statistical model that requires calibration of parameters. To overcome this problem we used a method that is based exclusively on information from the outset.

[5] See Statistical Bulletin 153, Consumer Price Index, October 2016, Central Bureau of Statistics.
[6] In effect we use only 40 observations due to the poor quality of the Database before 2017.

C. Big Data

This project, which is based on Big Data, is the first of its kind at the Bank, and as such it poses challenges that are not encountered in working with ordinary data.

This project, which is based on Big Data, is the first of its kind at the Bank, and as such it poses challenges that are not encountered in working with ordinary data.

The Database contains an enormous amount of data (approximately 10 million observations), which precludes the use of ordinary tools. Cloud-based tools make it possible to store such amounts of data and support the distributed retrieval and processing capabilities that are required to effectively work with such an amount of data.

The data are received in various formats, and it is a difficult task to transform the existing data into data amenable to processing. Missing data, transposed columns, and problematic texts all increase the challenge of working with the existing data. Although all the data are received in the same structure, there are key fields in which there is no uniformity across retail chains or even within a single chain. For example, each retail chain has a different catalog number for each fruit and vegetable. Moreover, a single retail chain may have multiple catalog numbers assigned to a single product, and occasionally multiple products share a single catalog number[7]. As a result, identifying a product within a chain and across chains is a difficult task[8].

4. Estimation

This section explains the estimation method. The method weights the individual prices in the chain outlets into a single Fruit and Vegetable Index as described below. The forecast is constructed on the basis of the fruit and vegetable prices of the major retail chains[9].

The first step in this process is to identify the specific fruits and vegetables in each chain, or in other words, to find the catalog numbers of interest that can be used to construct a single price series for each chain. This step is performed by ordering all the products by name and selecting the catalog numbers whose product descriptions match the fruit or vegetable of interest[10]. In the second step, daily prices are consolidated into a monthly series for each catalog number by calculating the mean weights across the days of the week, where weekends are given greater weight than the days at the beginning of the week. The monthly series of catalog numbers are consolidated to a single series for each fruit or vegetable by calculating the average price across all the outlets of a single chain, and calculating the average price across all the catalog numbers used by the chain whose description corresponds to the product in question. The result of this stage is a mean monthly price series for each fruit and vegetable in each chain.

[7] The product type is determined according to the text that describes the product, but we cannot rule out the possibility that the text was entered in error and the catalogue number does not truly represent the product described in the text. Furthermore, the name of a fruit or vegetable may appear in hundreds of different products.

[8] Especially since we cannot rule out the possibility that a single catalogue number is used for one product in a specific month and for a second product in another month.

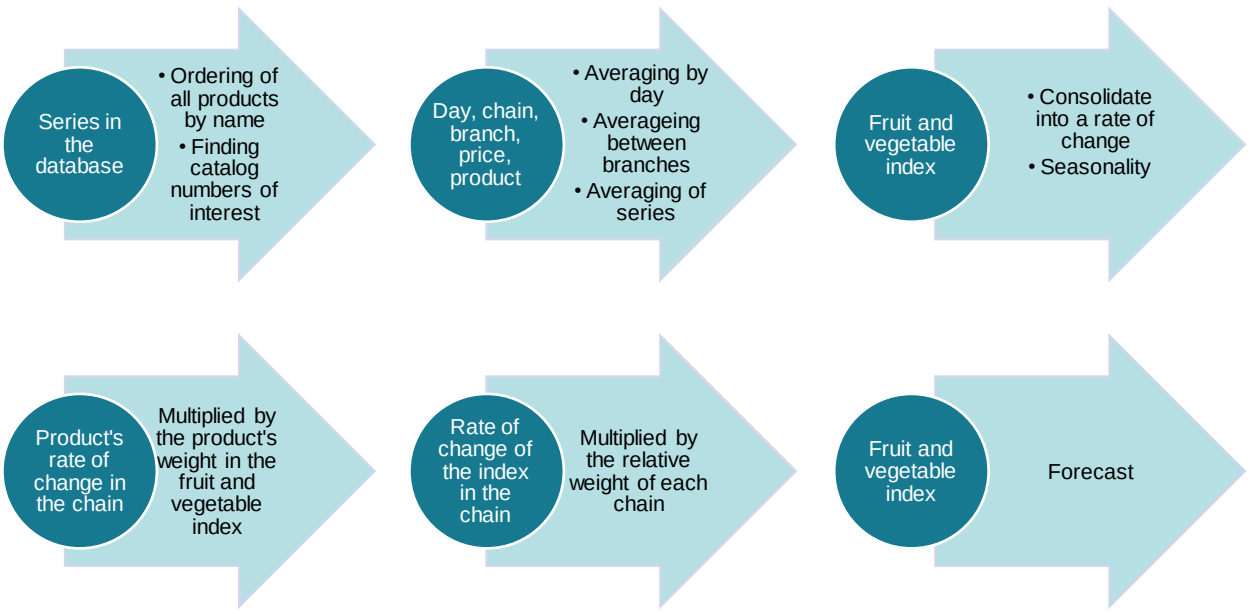
[9] As noted earlier, the products do not have an identical name or catalogue number across chains or within chains, and it is therefore challenging to identify a specific fruit or vegetable. As a result, the products are not selected automatically and the forecast is therefore based on information from the major retail chains under the assumption that a change in the prices in the major chains is a good approximation of the price changes in all retail chains.

[10] To reduce the search space, the products are also ordered by unit size. In contrast to other products, most fresh fruit and vegetables are sold by the kg.

The monthly price series makes it possible to calculate the rate of change in the price of each fruit and vegetable in each chain. The rates of change in all products are consolidated into the rate of change of the Fruit and Vegetable Index by calculating a weighted average that is based on the weight of each fruit and vegetable in the CPI. To prevent skewness caused by products that are out of season[11], the rate of change in the Index is determined only on the basis of products that are in season[12].

In the final step, the separately calculated fruit and vegetable indices are now consolidated into a single forecast, after each chain receives a different weight[13]. The entire process is described in the flowchart below. Finally, this forecast is averaged with the forecasts that currently exist at the Bank to create a final forecast of the Fruit and Vegetable Index. The reason for this step is that statistical forecasting is imprecise by nature and in the best case offers an adequate approximation of the target variable. Typically, no single predictor is consistently accurate. Research offers empirical evidence that averaging across forecasts partially resolves these concerns as the average forecast is resilient to the errors of any single constituent forecasting model (see e.g., Altavilla and de Grauwe, 2006).

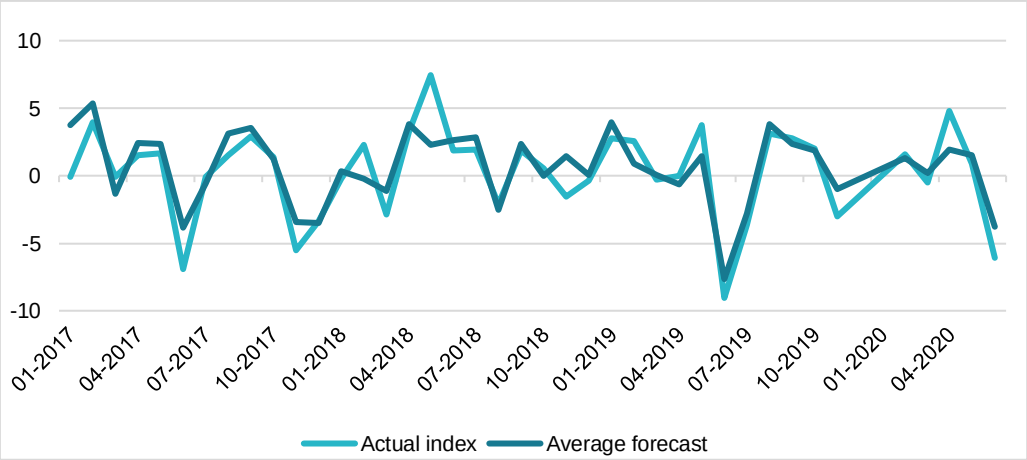
Chart 1: Description of forecast



[11] The season of a fruit or vegetable is determined according to the number of outlets in which it is sold throughout the chain.
[12] In other words, the rate of change calculated for products that are not in season is the rate of change of all other products that are in season, such that products that are not in season do not skew the index.
[13] According to relative size.

5. Findings

Figure 1: Fruit and Vegetable Index and Average Forecast (Retailers and Ministry of Agriculture)



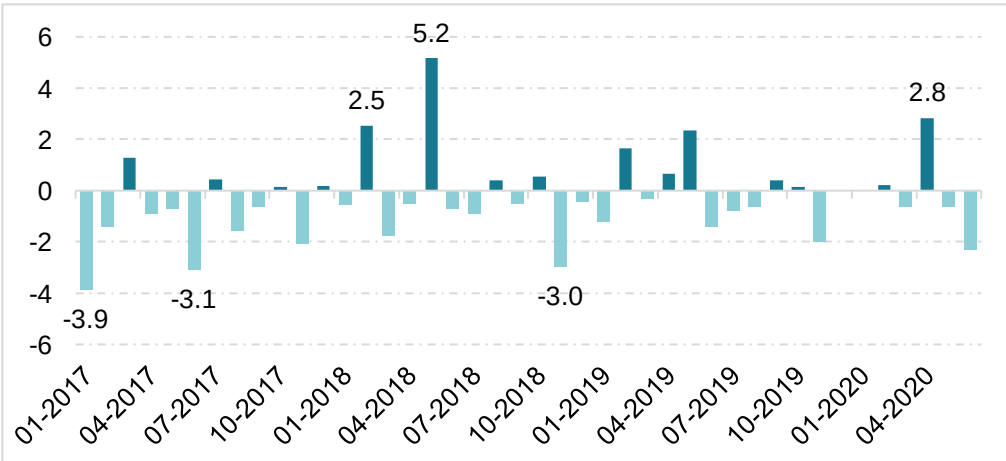
SOURCE: Based on Retail Price Database

Figure 1 shows the forecast (the average of the Bank's current forecast and the forecast calculated on the basis of the Database) compared to the Fruit and Vegetable Index released by the CBS between January 2017 and June 2020. The quality of the forecast was tested using the root mean square error (RMSE), which is calculated as follows:

$$\sqrt{\sum_{t=1:T} (y_t - \hat{y}_t)^2}$$

Where y_t is the real value of the Fruit and Vegetable Index at time t, and y_t is the forecast value (average of the forecasts) at time t. This calculation shows that averaging the Bank's forecasts with the forecast produced from the Database improved forecast quality by approximately 23 percent in the tested period. If we examine the RMSE of the Bank's forecast (2.2) compared to the forecast based on the Database (2.3), we see no improvement[14].

Figure 2: Errors in the Proposed Forecast Compared with the CBS Fruit and Vegetable Index

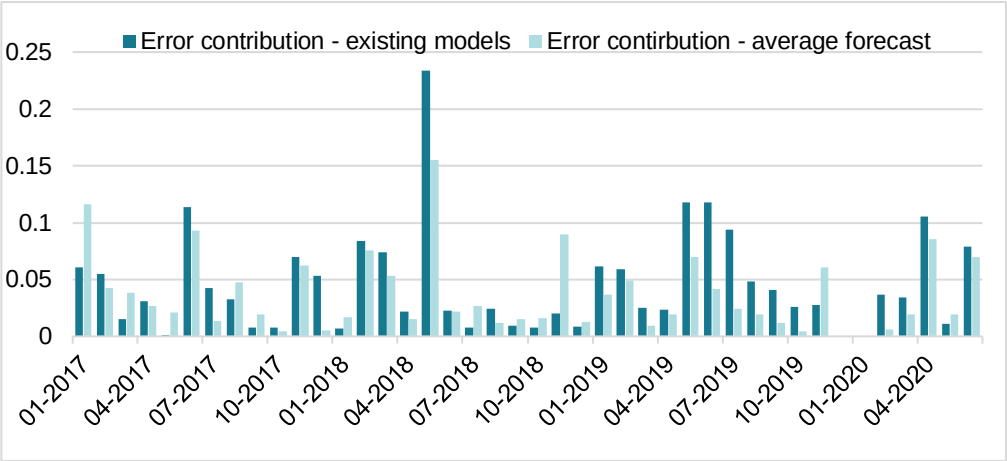


SOURCE: Based on Retail Price Database

[14] These indices were calculated on the basis of data from January 2017 to June 2020 (excluding December 2019 and January 2020). The RMSE was selected because it "punishes" large errors.

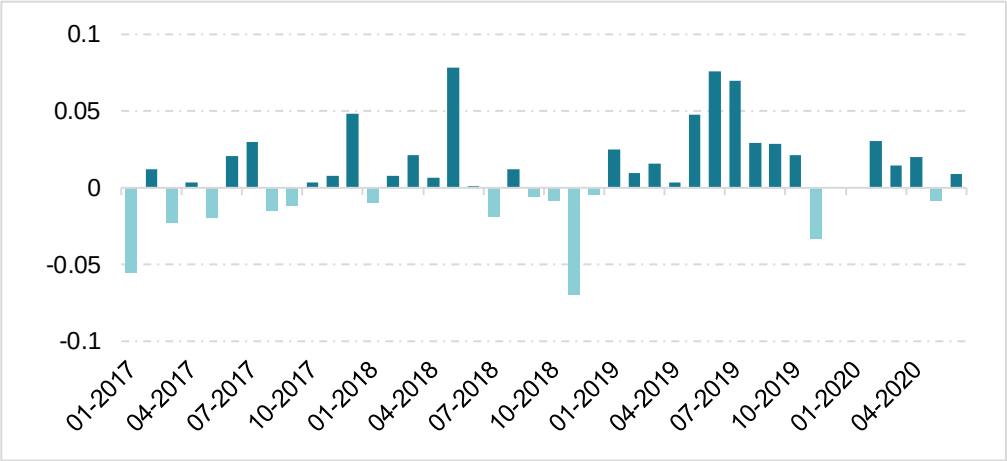
However, when an average of both forecasts is calculated, we obtain an RMSE of 1.7 percent, which represents a 23 percent improvement relative to the existing forecast in the test period. Figure 2 shows that almost all the errors of the proposed forecast are within a range of less than 1 standard deviation from the Fruit and Vegetable Index (3.3), and most errors are much closer. A summary of the performance of the two forecasts, including mean error and median error, are presented below in Table 1.

Figure 3: Contribution to the Consumer Price Index Forecasting Error



SOURCE: Based on Retail Price Database

Figure 4: Error in the Existing Forecast and Error in the Proposed Forecast - Difference in Absolute Value between Forecasts



SOURCE: Based on Retail Price Database

This improvement is also reflected in the fruit and vegetable component's contribution to the total forecasting error of the CPI. Figure 3 shows that the proposed forecast's contribution to the forecasting error is smaller (in absolute value) than the contribution of the existing forecast, and the difference in contribution may sometimes be large. For example, in July 2019, the existing forecast contributed 0.09 percentage points to the forecast error while the proposed forecast contributed 0.02 percentage points in that period, reflecting a 77 percent improvement. The proposed forecast's contribution to the error does not exceed 0.15 percentage points, and is 0.038 percentage points on average. Both parameters show an advantage over the existing forecast (maximum contribution of 0.23 percentage points and mean contribution of 0.048 percentage points). Figure 4 shows the difference between of the contributions of the existing forecast and the proposed forecast to the forecasting error in absolute terms[15]. It is evident that in most cases, the proposed forecast is superior, while the size of the error is more significant in the existing forecast.

Table 1: Performance of the various forecasts

Model	RMSE	Average error	Median error
Ministry of Agriculture Price Model (currently in the research)	2.2	1.6	1.0
Retailers database	2.3	1.9	1.6
Average of the two (Proposed forecast)	1.7	1.3	0.8

SOURCE: Based on Retail Price Database

Similarly, we can compare the forecasts by their contribution to the forecasting error of the overall CPI. Using the proposed forecast improves the CPI forecast by 5 percent in terms of RMSE. Specifically, in July 2019, the forecasting error of the existing forecast was 0.23 percentage points, compared to only 0.16 percentage points for the proposed forecast, reflecting a 30 percent improvement.

6. Reflections on the way forward

We have shown that the forecast based on an average of the data from the Database and the Bank of Israel's current forecast improved the Bank's forecasting ability of the Fruits and Vegetables component of the CPI in the reviewed period. Moreover, it is possible to improve the forecast that is calculated from the Database in several ways by using optimized data, prices net of discounts, a combination of forecasts from other databases, and other methods.

In conclusion, we see that it is possible to improve the forecast of the Fruit and Vegetable Index based on continuous itemized information. This is the beginning of a growing trend of making extensive use of Big Data in order to discover the answers to emerging questions and to improve the accuracy of current forecasts.

[15] To explain this figure in greater detail: A positive value in the graph implies that the proposed forecast is superior, and vice-versa. The column size reflects the size of the error. For example, a high positive value implies that the Bank's forecast has a larger error than the proposed forecast, and a high negative value implies that the proposed forecast has a larger error than the Bank's forecast.