Consumer Price Index in the Czech Republic – New Sources and Data Processing

Jaroslav Sixta[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Petr Musil[2] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 1.8.2023, **Accepted** (reviewed) 3.11.2023, **Published** 15.3.2024

Abstract

Consumer price index has been in the centre of interest for many years, since being published in 1990s in the Czech Republic but recent price growth raised more questions on methodology and data sources used in price statistics. Users are interested not only in the figures itself but also in statistical issue influencing interpretation and the quality of consumer price index that is often used as an approximation of inflation rate. The paper introduces price statistics compiled by the Czech Statistical Office and it specifically focuses on data sources and in particular scanner data. The paper explains how advanced statistical methods such as machine learning are implemented in official statistical production. We think that the official statistics is being on the historical junction where modern methods are going to be implemented. Our paper shows the usage of machine learning procedures applied on scanner data within consumer price index. Used method is based on logistic regression and powerful Python solution and that provides fast and high quality results.

Keywords	***DOI***	***JEL code***
Consumer price index, scanner data, machine learning, logistic regression	*https://doi.org/10.54694/stat.2023.37*	*C 43, E 31*

INTRODUCTION

Price statistics produces various indicators describing either price development or price level in a given economy. A palette of available indicators differs among countries depending on a level of development of price statistics. A minimum set of indicators is laid down by the EU regulation for EU members. It can be said that the European data set exceeds standards in other countries in terms of coverage, timeliness and frequency. In addition, Eurozone countries are obliged to compile flash estimate of a consumer price index that is being published at the end of a month.

[1] Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Curchill Sq. 1938/4, 130 67 Prague, Czech Republic. Corresponding author: e-mail: sixta@vse.cz. The author is a Vice-President of the Czech Statistical Office, Na Padesátém 81, 100 82 Prague 10, Czech Republic.
[2] Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, W. Curchill Sq. 1938/4, 130 67 Prague, Czech Republic. E-mail: petr.musil@vse.cz. The author is also working at the Czech Statistical Office, Na Padesátém 81, 100 82 Prague 10, Czech Republic.

Price indices are used for two main purposes: to describe price fluctuation of products or a group of products and to deflate nominal indicators to volume terms (process of statistical deflation). Application of price indices in a deflation process is well described in respective manuals, e.g. Handbook on price and volume measures in national accounts (European Commission, 2016). To this end several types of prices indices are required: Producers price indices (PPIs) measuring price development of supplies at basic prices to a domestic market, prices of foreign trade and derived terms of trade, price indices of real estates and land and consumer price index (CPI). The latest is the most famous and well known by not only statisticians and economists but also general public. CPI enables a comparison of household income and expenditures constituting the main component of GDP in developed countries. We can say, CPI is an approximation of inflation rate and one has to consider its limitations. Last but not least consumer price index is an indicator used for inflation targeting.

Price indices are normally produced and published by National Statistical Institutes (NSIs). Users in the EU benefit from a high level of harmonization laid down in the EU legislation. Harmonised Index of Consumer Prices (HICP) is well-known but other types of price indices are also subject to harmonization. On one hand, price indices are still based on the same principles: Laspeyres formula, monthly frequency for most indices etc. On the other hand, most statistical institutes are creative and have been continuously improving quality of price statistics. For instance, revisions of a weighting scheme are more frequent to capture changes in consumption habits. In addition, new data sources have been acquired and used in the production process. The most valuable source is scanner data that allow us to make a jump in quality. Obviously, scanner data have been processed in different way than data collected within a field survey. Our paper describes a particular detail that is very innovative and significantly improved the quality and possibility of price comparisons, scanner data. The process of construction of CPI based on scanner data cannot be fully automatized up to now but machine learning processes (MLP) radically increased the scope and efficiency of such statistical process. Since 2019, when scanner data were incorporated into statistical production of CPI the amount of data has been continuously rising reaching about 500 thousands records per month. About 10 thousand products are new every month and need to be classified into the classification of consumption by purpose (COICOP). Machine-learning is connected with artificial intelligence (AI).[3] The situation usually stands as the computers are learning from training data and later derived algorithms and parameters that are used for predictions. It is supposed that such algorithms are still improving by both supervised and non-supervised learning. The third existing approach is reinforcement training nearly completely automatic. Up to now, the most prevailing in official statistics in supervised learning. It is obvious that at the beginning, human work is necessary but later on we can use supervised machine-learning processes. The incentives for changes and switchover to partly or fully automatized process grow both from internal or external environment. Most of European statistical office are facing cost reductions and the pressure on efficiency. Optimal allocation of scarce resources – qualified staff is very necessary. Statisticians working for state statistical agencies and offices are usually public clerks, conservative ones but not blind. Enormous spread of modern techniques couldn't stay be overlooked. External environment also determines newcomers from universities with excellent knowledge of modern IT tools such as R and Python programming languages. Time to time, some of these activities could be associated with the term Big Data but it is not our case. The Czech Statistical Office receives twice a month a batch of large amount of pre aggregated data by individual products (later described in a detail), quantities sold and sales from this product. This allows the Czech Statistical Office to compute average prices that completely respects the real demand on particular product, CPI.

[3] Google very nicely explains the connection points and differences, see: <*https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning*>.

The main purpose of the paper is to describe and explain modern statistical methods that have recently been implemented in processing of scanner data. We believe that the methods are inspirational for official statisticians across statistical domains. The first part introduces price statistics with the focus on consumer price index. Next, data sources and data processing are described. Last section is devoted to machine learning based on logistic regression that is applied for data processing.

1 METHODOLOGY

A number of products that are produced and consumed in national economy is finite but hardly countable. For each single product, it is possible to estimate a price index either describing price development over time or comparing prices between regions. However, products are quite heterogeneous, consequently an average price cannot be calculated (e.g. average price of 1 kilogram of vegetable). In order to estimate an 'overall' aggregated index, it is necessary to introduce weights and averaging individual price indices. Laspeyres' formula is applied for the most of price indices. All of published price indices are estimates of 'theoretical' price indices, which are not observable. For example, Cost of living index (CLI) is a theoretical price index for household expenditure defined as 'A ratio that measures impact of price change on consumer well-being' (Fixler, 1993). Laspeyres index using weighting scheme from previous period is upper bound of CLI while Paasche index using weighting scheme from current period is lower bound of CLI (Schultz, 1939). It should be noted that several assumptions need to be held, in particular a typical (decreasing) demand curve that is depicted for instance in Samuelson, Nordhaus (2009). The relationship between Laspeyres and Paasche indices is also described by so-called Bortkiewicz formula (Lippe, 2012). Superlative indices overcome caveats of Laspeyres and Paasche indices by various types of averages of weights (Diewert, 1976).The most famous ones are Fisher, Törnqvist.

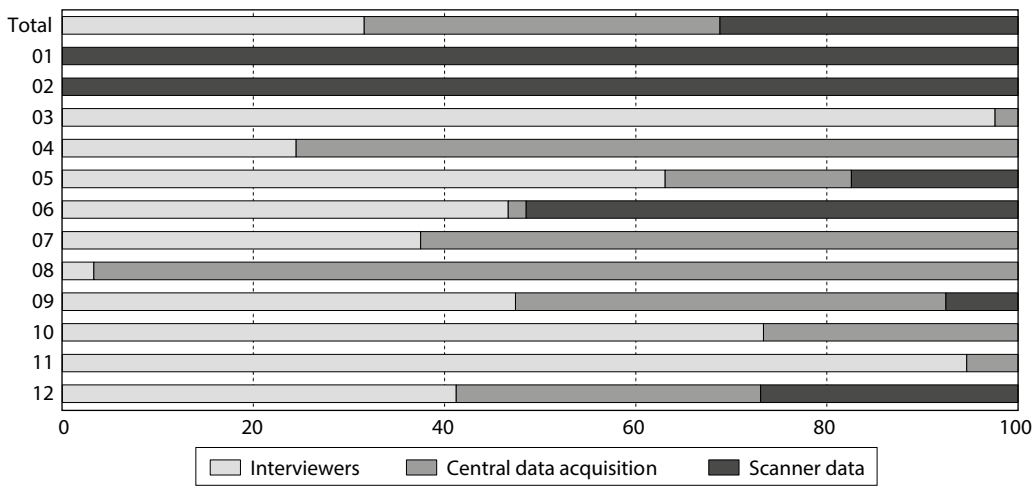
It is well-known that almost all price indices are Laspeyres types. In other words superlative price indices, which are closer to CLI, are not produced and published by National Statistical Institutes. Most price indices including consumer price index are published monthly while weights are based on annual structures. Price statistics is considered as short-term statistics showing emerging trends in national economy. For instance, consumer price index is usually released in the middle of the following month; moreover, flash estimate of HICP is published at the end of a given month. At that time no data source that provides information on products being transacted (produced, consumed) is available. In addition, weights are derived from annual structures in order to eliminate seasonal effect, short-run fluctuation. It does not mean that weights are outdated as regular update is carried out, e.g. HICP weights of ECOICOP categories are updated annually, CPI weights once in 2 or 3 years including detailed breakdown by price representatives. Similarly, weights of other price indices are regularly updated. Consequently, neither Paasche formula nor superlative price indices can be applied.

In practice, the weights are estimated using several data sources. Among them, the most relevant ones are national accounts and household budget survey. National accounts depict comprehensively national economy including the household sector nevertheless the level of detail is not sufficient to derive weights for representatives. As consequence national accounts data adjusted to methodological differences such as non-monetary transactions for instance agriculture self-supply are deployed to estimate weights of COICOP groups. Household budget survey data serve as a supplementary source to derive detailed weights of price representatives within a given COICOP group. It should be noted that household budget survey is completely replaced by scanner data for relevant product groups. In order to keep the weighting scheme relevant they are updated once every two years while the HICP regulation (2016/792) requires annual update. Due to the Covid-19 outbreak and related containment measures that substantially affect a consumer basket the current weights applied from January 2022 are based on household expenditure average in 2019–2021. It is believed that a change in the structure of household expenditure caused by the Covid-19 outbreak is temporary and therefore should not be fully reflected in the weighting scheme.

2 DATA COLLECTION

As mentioned above, all price indices are based on sampling i.e. prices of selected products are surveyed only. Traditionally, prices had been collected in outlets by interviewers. At later stage central data acquisition was introduced especially for products that are supplied centrally e.g. energy, gasoline, communication service. The most modern and sophisticated method is a collection of electronic records ('scanner data'). Scanner data were firstly applied by Dutch Statistical Office in 2002, some other countries started to use this data source in 2010 (Bialek, 2020). The method is triggered by the EU Regulation (2016/792) that lays down obligation of statistical units to provide scanner data to NSIs. The Czech Statistical Office launched voluntary collection of scanner data a couple of years before the Regulation came into force. Since then the collection became binding for retailers whose revenues exceeded a given threshold. The first product group was foodstuff next ones were drug products and drugstore products. Currently, hobby markets have been asked to provide scanner data. It is planned to extend a coverage substantially (Bookstores, Fashion & Clothing retailers) in the following years. The Czech Statistical Office plans to replace a price collection in outlets by scanner for all products for which benefits surpasses costs. Figure 1 depicts methods of data collection used in consumer price index by main COICOP categories.

Figure 1 Data collection methods by COICOP categories



Note: 01 – Food and non-alcoholic beverages, 02 – Alcoholic beverages and tobacco, 03 – Clothing and footwear, 04 – Housing, water, gas electricity and other fuels, 05 – Furnishings, household equipment and routine maintenance of the house, 06 – Health, 07 – Transport, 08 – Communications, 10 – Education, 11 – Restaurants and hotels, 12 – Miscellaneous goods and services.
Source: Own elaboration based on Rojíček and Sixta (2022)

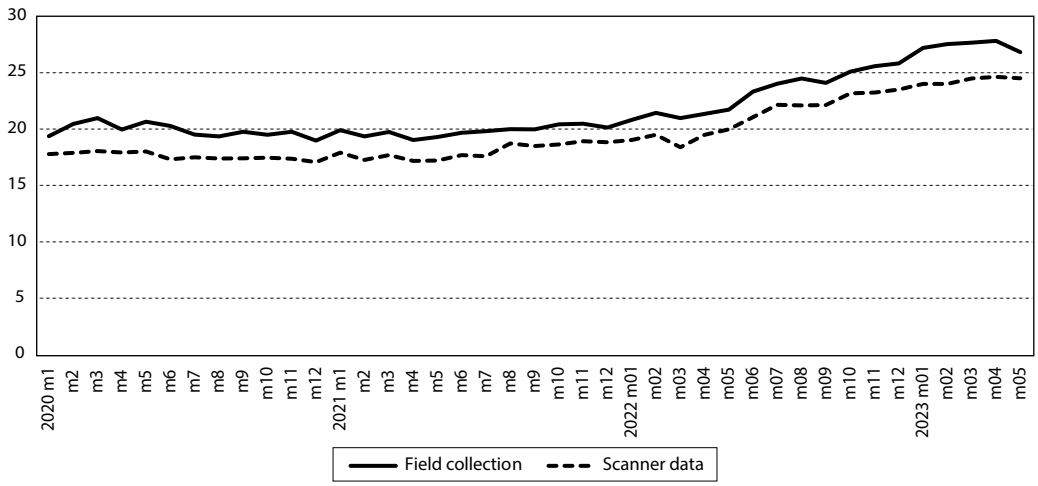
Scanner data represent a substantial improvement in the quality of consumer price index. Firstly, number of products whose prices are surveyed have jumped from hundreds to approximately 100 000. Secondly, actually realized prices instead of spot prices are used. Last but not least, comprehensive data on household expenditure are available and can be deployed in statistics.

Prices of about 700 products of which approx. 150 food products were surveyed before the implementation of scanner data. It should be noted that interviewers collected prices at the moment of collection, i.e. spot prices. Indeed, Nielsen reports that more than 50%[4] of products, in particular

[4] *<https://www.seznamzpravy.cz/clanek/ekonomika-byznys-trendy-analyzy-cesi-jdou-jeste-vic-po-slevach-v-akci-kupuji-uz-60-procent-zbozi-216642>.*

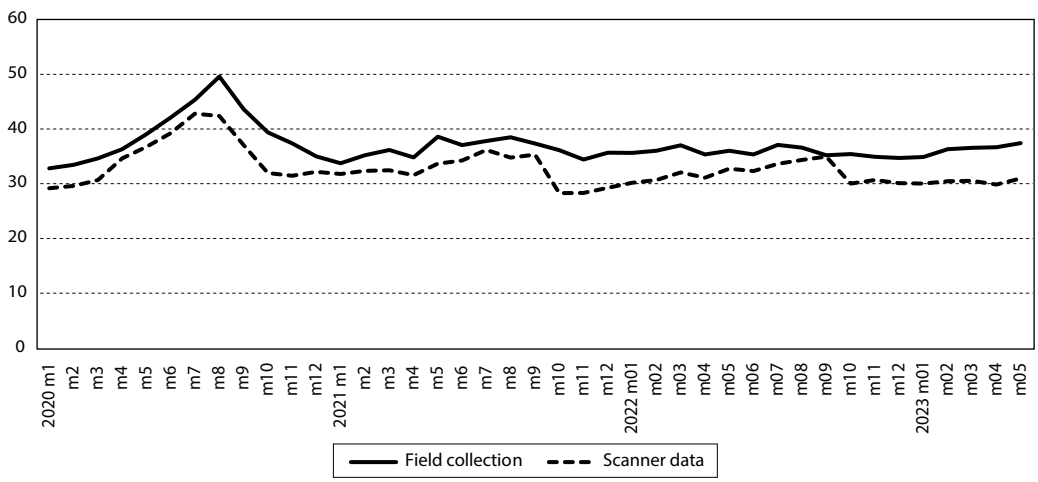
food stuff, are sold in promotion. Spot prices do not necessarily reflect actual (realized) prices affected by promotion. It does not necessarily distort price indices as it can be assumed that price evolution of spot prices and realized prices is the same. On the other hand it may affect international comparison as price level in a country with a higher share of promotion can be overestimated. The Czech Statistical Office stopped field collection of prices once scanner data of products became available. However, there is an exception of weekly survey of food stuff (about 10 items). Users have requested weekly prices for basic food stuff that cannot be gathered from scanner data. Actual prices from scanner data are higher than prices collected in shops by interviewers for all items but prices evolution does not necessarily vary. A difference in average price of semi skimmed milk can be seen stable over time while apple prices indicate also dissimilar evolution in certain time period.

Figure 2 Price of semi skimmed milk (CZK/liter)



Source: Own elaboration

Figure 3 Price of apples (CZK/kg)



Source: Own elaboration

On one hand, implementation of scanner data cuts of costs on a price collection in outlets. On the other hand, data protection and processing require additional ICT costs and highly qualified staff. Processing of scanner data poses a challenge to National Statistical Institutes. Statistical units (retailers) provide data sets that are not fully standardized due to different accounting software and internal classifications that proved to be very useful for further data processing. Even though EAN/GTIN codes have been introduced and are in fact useful for data processing those codes themselves are not sufficient for classifying items (products) to statistical classifications mainly the product classification (CPA) and classification of individual consumption by purpose (COICOP). There is no bridge table between EAN/GTIN codes and statistical classifications. Above that, EAN/GTIN codes are not entirely unique and stable during the time.

Scanner data also allow estimation of superlative and multilateral indices. Bialek (2021) presented comparison of indices for dairy products. He came to the conclusion that data filtering is extremely important with substantial impact on the results.

3 DATA PROCESSING

Utilization of scanner data poses a challenge on National Statistical Institutes. Scanner data represent enormously rich data source for official statistics, namely price, retail trade and national accounts statistics. At the same time scanner data need to be processed into information. Even though scanner data received by the Czech Statistical Office are not real big data as individual records (transactions) are not submitted to the CZSO. In other words, data are pre-processed by reporting statistical units. Only monthly aggregates in breakdown by products defined by EAN/GTIN or internal codes are sent. It means that data on individual transactions or higher frequency data are not available to the CZSO that limit amount of transferred data. In addition, monthly frequency suffices as no more frequent statistics are produced. However, respondents are requested to send data twice a month. The first transmission includes partial (incomplete) data for the first three weeks of a given month, which are used for consumer price index. The second (complete) data transmission is utilized in other statistical domains.

It should be noted that processing of scanner data is challenging and advances statistical methods need to be deployed. Data structure is not fully standardized as each respondent provides the data from its database. While certain variables are the same (number of products sold, revenue with/without VAT) other may differ (internal classification, description). In addition, product variety gradually changes as about 5 000 new products are identified every month. Those items need to be classified into statistical classifications such as ECOICOP, CPA. Obviously, manual data processing would be very resource consuming. Internal classifications are very helpful for data processing nevertheless statistical classifications, especially ECOICOP, are very detailed.

4 MACHINE LEARNING

The crucial point of all the effort leading to automatic coding or classification is to find suitable base that will be used for supervised training. Further on, selection of appropriate statistical method is necessary. With respect to that, process started detailed studying of the products sold at selected retail chain. Experts used up all the information hidden in shopkeepers' information system ranging from typical words to the position of the goods in the shop. It was soon recognised that this demanding work can be used only for a short time in testing period. During the regular production, it had to be found something else.

After several different attempts for implementation of some type of partially or fully automatized procedures, it was decided to select logistic regression model inbuilt into Python big data and scientific environment Pandas and Scikit-learn library. Logistic regression provides easy and reliable solution with

very limited set of assumptions[5] and it is applicable on categorical variables (characters). It was found that universal GTIN codes (bar codes) are not easily translatable into statistical classification (COICOP) and the description of the product by both letters, numbers and words would have to be used. Even though that all data providers use their own specific internal system and coding system for sold products that was deeply used up during preparatory phase.

Finally, the process of classification of newly observed products that are classified by logistic regression and the results lead to continuously spreading knowledge database that is used for further classification. Obviously, the quality of classification can be tested only ex-post which we allow in this paper. For illustration purposes, we present the process of computation on reduced example counting 100 000 records and test classification quality on the randomly selected 5% sample. Further on, we also present actual information based on complete set of data.

Standard model of logistic regression for binomial variable (belongs to the group 1 and does not belong to group 0) can be described by Formula (1). Dependent variable (y) expresses 0 or 1 depending on the group belonging. It is rewritten as share of probability belonging to the group (P(y_i=1) or π) to opposite case, the share represents the chance – odd. Right side of the equation express exponent of linear regression with unknown vector of parameters ($\boldsymbol{\beta}$) and explanatory variables (x_1 to x_k):

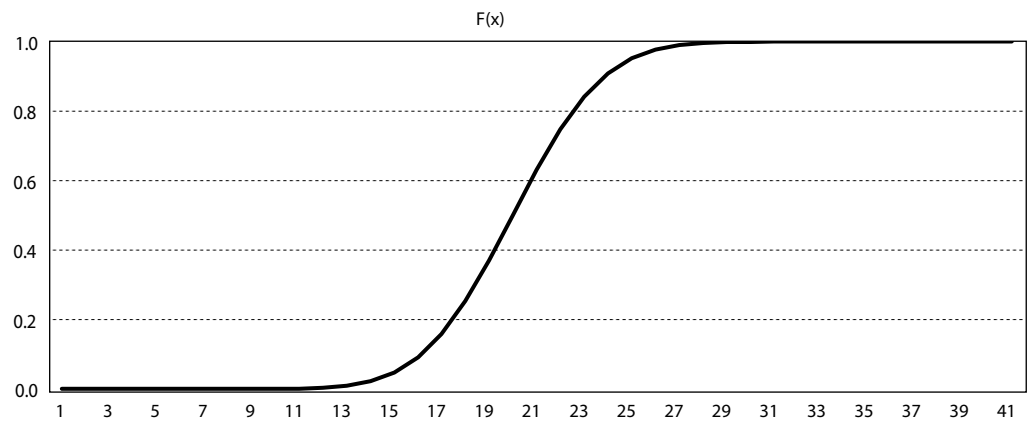
$$\text{odd } y_i = \frac{P(y_i = 1)}{1 - P(y_i = 1)} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}. \tag{1}$$

For transformation purposes, the model is rewritten for logit (left side of equation) as:

$$\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \tag{2}$$

Generally, the probability of classification in logistic regression has very suitable form and in optimal case leads to relatively strict separation between probability of belonging and non-belonging to group, e.g. sample logistic function see Figure 4.

Figure 4 Probability of classification in logistic regression



Source: Own elaboration

[5] Despite logistic regression does not require the same set of assumptions as linear regression, some assumptions are necessary, mainly dependent variable to be binary, independent variables should not be too highly correlated, independent variables are linearly related to the log odds and usually large sample size is recommended.

The description was provided on the basis of binomial dependent variable but in practice we use 157 COICOP groups, we need 156 equations. This multinomial comparison is based on the one vs all method, where for each product is estimated probability of belonging to tested group versus all other. It means that a product is being tested where it belongs to foodstuff, beverages, tobacco, …domestic appliances, …sport equipment, etc. The product is classified according to the highest probability in corresponding group.

The set of dependent variables correspond to the probability of belonging to 157 categories (given by specific COICOP group). All explanatory variables are derived from the product description used by the shopkeepers. This is represented by letters, numbers and words in the description and therefore all explanatory variable are categorical. For better explanation, see following Table 1 representing very short part of the explanatory variable, called dictionary (contains full or shortened description of the products). Due to computational issues, the Czech Statistical Office limits the number of explanatory variables to 150 000.

Table 1 Sample of the dictionary

Number	Text
1	0
…	…
24	00ml
…	…
203	100x200cm
…	…
17 778	Irsai
…	…
41 905	Zweigeltrebe
…	…

Source: Own elaboration

The explanatory variables cannot be used in the presented form, they have to be translated into a Boolean type of variables, representing the occurrence of concrete variant of the text. Processing of the text is very modern and Python environment allows many possibilities such as text recognising. The way that is used for logistic regression is based on transformation of explanatory variables into Boolean type, 0 or 1 when exact character/letter/word presents. A Python tool Vectorizer is used. In this way, variables are processed in a form of unit-zero matrix, see Table 2.

Table 2 Boolean representation of the matrix of explanatory variables

	Explanatory variables alphabetically ordered									
	1	**…**	**24**	**…**	**203**	**…**	**17 778**	**…**	**41 905**	**…**
Observation	0	…	00ml	…	100x200cm	…	Irsai	…	Zweigeltrebe	…
1	0	…	0	…	0	…	1	…	…	…
2	0	…	0	…	0	…	0	…	1	…
…	…	…	…	…	…	…	…	…	…	…
55 000	1	…	1	…	0	…	0	…	0	…
…	…	…	…	…	…	…	…	…	…	…
195 000	0	…	0	…	1	…	0	…	0	…
…	…	…	…	…	…	…	…	…	…	…

Source: Own elaboration

The estimates of parameters are completely done in Python. Since we have 150 equation with more than 150 000 parameters, it is not possible to simply present all coefficients and their p-values. Only synthetic quality information can be easily interpreted. Table 3 brings the sample of estimated parameters,

$\hat{\beta}_0, \hat{\beta}_1, .. \hat{\beta}_{150000}$.

Table 3 Illustration of fitted model

	b0	b1(500)														
		10	100	100g	100ml	12	140	15	150	150g	150ml	16	1l	20	200	200g
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	–0,0	0,3	–0,5	–1,2	–0,2	–0,0	–0,0	–0,1	–0,2	–0,2	2,7	–0,0	–0,4	–0,0	–0,4	–0,8
2	–0,0	–0,1	–2,0	–1,7	–0,0	–0,1	–0,0	–0,6	–0,8	–1,0	–0,2	–0,0	–1,9	–0,1		
3	–0,0	–0,2	–0,3	–0,4	–0,0	0,1	–0,0	–0,0	–0,8	0,1	–0,0	–0,0	–0,1	0		
4	–0,0	–0,2	0,9	0,5	–0,1	–0,8	0,1	–0,4	0,2	0,5	–0,0	–0,1	–1,1	0,8		
5	–0,0	–0,1	–0,1	–0,3	–0,1	–0,9	–0,4	–0,0	0,8	–1,2	–0,0	–0,0	–0,1	–0,0		
6	–0,0	–0,3	0,4	–0,2	–0,2	–0,1	–0,1	–0,2	1,3	–0,8	–0,3	–0,0	–0,7	–0,0		
7	–0,0	–0,8	0,8	0,8	–0,0	–0,1	–0,6	–0,5	1,1	0,0	–0,0	–0,0	–1,6	–1,9		
8	0,0	–1,0	0,8	0,5	–0,2	–0,4	1,1	0,5	–0,7	–1,1	–0,2	–0,1	–1,6	–2,1		
9	–0,0	1,2	1,7	2,4	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,5	–0,0	–0,0	1,4	–0,7	2,3
10	–0,0	0,8	–0,4	–0,2	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,7	–0,0	–0,0	–0,7	–0,8	–1,2
11	–0,0	–0,0	–0,1	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,0	–0,1	–0,0	–0,0			
12	0,0	–0,6	–0,6	–0,2	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,9	–0,0	–0,0	–		
13	–0,0	–0,1	–0,1	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,0	–0,1	–0,0	–0,0			
14	–0,0	–0,6	–0,2	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,0	–0,4	–0,0	–0,0			
15	0,0	0,8	0,1	1,5	–0,0	–0,1	–0,3	–0,7	0,3	–1,4	–0,2	–0,0	–0,1	–0,1	–0,0	–1,1
16	0,0	–1,2	–0,4	0,7	–0,0	1,2	–0,0	0,7	1,2	0,9	–1,8	–0,0	–0,2	1,5		
17	0,0	–0,1	–0,6	–0,2	–0,0	–0,0	–0,0	–0,0	–0,1	–0,3	–0,3	–0,0	–0,0	–		
18	–0,0	–0,0	–0,3	–0,5	–0,0	–0,4	–0,0	–0,0	–0,1	–0,3	–0,0	–0,0	–0,0	–		
19	0,0	–0,0	–0,2	–0,1	–0,0	–0,0	–0,0	–0,0	–0,1	–0,1	–0,1	–0,0	–0,0			
20	–0,0	–0,0	–0,1	–0,1	–0,0	–0,1	–0,0	–0,0	–0,1	–0,1	–0,1	–0,0	–0,0			
21	0,0	–0,0	–0,2	2,0	–0,0	–0,2	–0,0	–0,0	–0,1	1,4	–0,1	–0,0	–0,0	0		
22	0,0	–0,1	0,2	–0,6	–0,1	–0,4	–0,0	–0,1	0,6	1,5	–0,5	–0,0	–0,1	–0,0		

Source: Own elaboration

The interpretation of regression coefficients is not very easy since they are reflecting transformed variables. For the reflecting quality of the model, the CZSO estimates the rate of correctly classified products, sensitivity and specificity and ROC. These tests are done ex-post by splitting the data randomly into testing (95% of the sample) and prediction group (5% random sample). The results can be classified according to the belonging to a particular COICOP group (belongs – 1 positive, does not belong – 0 negative), see Table 4.

Table 4 Possible classification of outcomes

		Actual	
		False (0)	**True (1)**
Predicted	False (0)	True negative (TN)	False negative (FN)
	True (1)	False positive (FP)	True positive (TP)

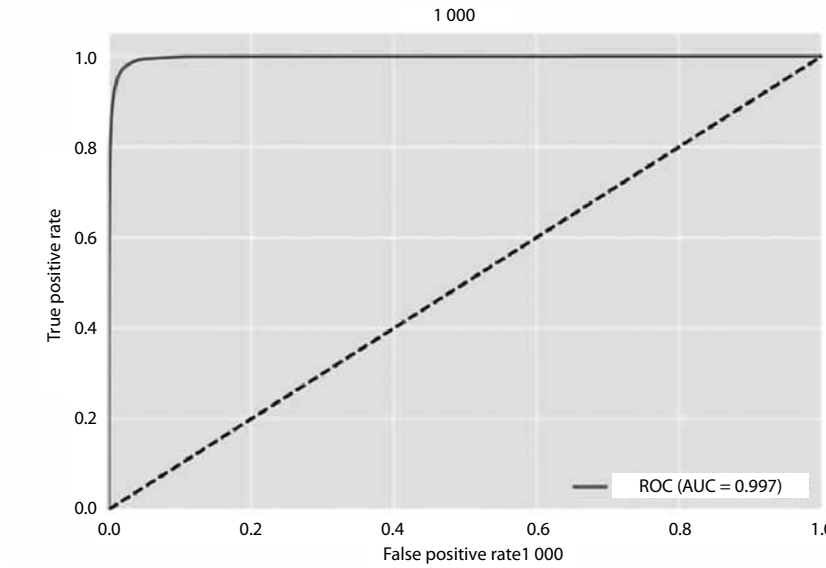
Source: Own elaboration

Standard measures in logistic regression such sensitivity and specificity are used as well as ROC curves. Since one to all approach is applied, the data for different COICOP groups had to be finally aggregated for overall assessment. Python environment and libraries Pandas and Scikit-learn are used. The overall success rate exceeds 95% when classified new product. ROC[6] is perfect for illustration of binomial classification,

[6] ROC is receiver operating characteristic curve and it is used for assessing the quality of classification with respect to sensitivity and specificity.

in our case it is obtained as an average of classification in detailed COICOP level. Since, the sub-groups are not well balanced, ROC provides just indicative picture of the quality of classification. For ROC of food and non-alcoholic beverages, see Figure 5. Despite the limitation of ROC for aggregated data, the quality of classification by logistic regression is enormous. The development of the true positive rate is very good and the Area Under the Curve (AUC) reached 0.997 when 1 is perfect fit.

Figure 5 Estimation of ROC for food and non-alcoholic beverages, COICOP 10



Source: Own elaboration based on the estimates of M. Král (CZSO)

The quality of classification of products (rate of truly classified) reaches very high numbers, exceeding 0.9, 90% of products are correctly classified. When studied for seven possible COICOP groups where these products belong, results are higher than expected. The probability of correct classification is presented in the Table 5. It is influenced by extreme observation and non-weighted average (i.e. all products without

Table 5 Probability of classification into COICOP group

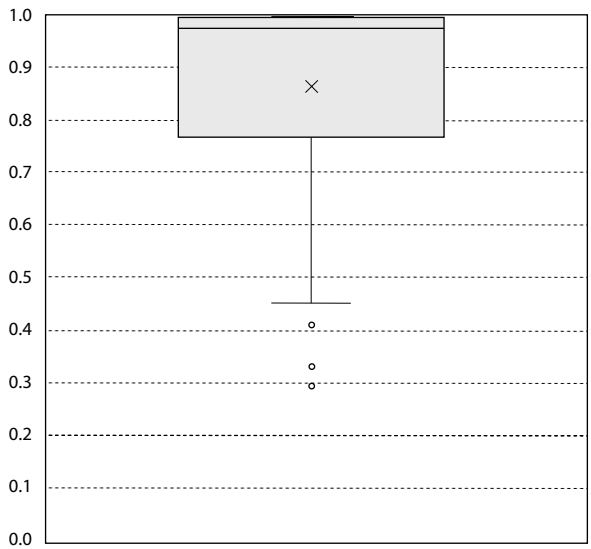
True classification		Predicted classification								
		01	02	03	04	05	07	09	12	Mean
1	Food and non-alcoholic beverages	0.68								0.68
2	Alcoholic beverages, tobacco		0.80							0.80
3	Clothing and footwear			0.95				0.41		0.92
5	Furnishings, household equipment and routine household maintenance				0.70	0.84				0.82
7	Transport						0.98			0.98
9	Recreation and culture							0.92		0.92
12	Miscellaneous goods and services								0.98	0.98

Source: Own computation on selected subset of data

respecting the amount of sales) does not fall below 0.68.[7] The figures in the tables come from aggregation of maximal probabilities belonging to the group. It must be taken into account that these aggregations do not play role in practice since products are classified on the lowest possible detail. The only problematic group seems to be clothing and footwear where significant misclassification was found. It means that for some products, the highest probabilities pointed at incorrect group, from clothing to recreation and culture. The distinction between sport tools and clothing is not easy in practical statistics and moreover this group is not currently fully covered by specialised shops. It means that these products come from retailers with high variety of products.

Similar view on the quality is presented on the Figure 6, this box-plot provides a distribution of probabilities of correct classification within our sample. It is clear that lower quartile reaching nearly 0.8 and median is close to 0.97. In the sample of 100 thousands of products, just three extremes were observed, from 0.3 to 0.5.

Figure 6 Distribution of probabilities in a sample



Source: Own elaboration

With respect to finds presented above, the overall classification of products is very successful and it allows very fast large data processing. Currently, the efficiency reaches upper limit of CZSO's possibilities. The knowledge database that being developed is continuously enlarging.

5 FURTHER DEVELOPMENT

The issue of machine-learning processing of scanner data will be supplemented by web-scrapped data in the near future. Some countries, such as Austria has been deployed scrapped data in price statistics (European Commission, 2020). Web-scrapping has lots of advantages but also limitations. Web scraping does not require cooperation with respondents that are just notified or discussed at the beginning of such process that is usually continuous and repeats with regular frequency and lasts long time.

[7] Due to computational issues, these results were estimated on basis of subset of all scanner data (100 000 observation) for illustration purposes.

The most important limitation is the sole existence of a spot price of product without quantity sold. Firstly, it is not possible to calculate true average price in a given period. Secondly, weights for aggregation into consumer basket are missing. However, practical solutions may be adopted to overcome missing quantities (European Commission, 2020). For example, it is recommended to Jevons index or more dynamic approaches (e.g. multilateral methods such as GEKS Jevons).

The Czech Statistical Office has experience with web-scraping in tourism statistics while price statistics in particular consumer price index rely on scanner data, central data acquisition and field collection. We believe that scanner data are more comprehensive data source than web-scraping especially for products whose price are volatile or their price elasticity of demand is high. Nevertheless, web-scraping may be applied in the future for companies that do not have electronic records at disposal.

CONCLUSION

The paper summarizes data sources and statistical methods used in price statistics notably consumer price index and brings readers overview of the current state of art of this new high quality statistics. Ongoing digitalization is a unique opportunity for modernization of (not only price) statistics using new data sources and data processing. The Czech Statistical Office started exploiting scanner data once respective legal act was adopted. Undoubtedly, scanner data have considerably improved quality of statistics. At the same time, new automatized procedures have to be applied to process huge data datasets within couple of days. Machine learning based on logistic regression fits very well.

Machine learning applied within the CZSO belongs to set of changes that should increase the quality and efficiency of statistical production. The possibilities of modern tools such as R and Python are huge but the most important for a successful application of ML procedures are high qualified statisticians and data analysts. The overall success rate exceeding 95% was not expected at the beginning since optimistic estimates were around 80%. This was a very big step into the new field and its possibilities. The success with scanner data motivated statisticians in other fields and now there can be found other projects at the CZSO being developed suited well for automatized and machine-learning classification. At the same time, machine learning itself it not a tool for quality adjustments or substitution of products. Highly qualified statisticians need to be employed by NSIs to develop and correctly apply modern statistical methods.

We are of the opinion that Laspeyres formula works well and the results are easily interpretable to the general public. Obviously, weights should be regularly updated to reflect changes in consumption habits. Nevertheless, scanner data allows to calculate also superlative price indices that will be subject of our future research.

ACKNOWLEDGEMENT

The support of the Technology Agency of the Czech Republic within the Project No SS04030013 (Center for Socio-Economic Research on Environmental Policy Impact Assessment) is gladly acknowledged.

Authors would like express special thanks to Michal Kral, expert from the Czech Statistical Office, who estimated ROC curves for the purposes of this paper.

References

BIALEK, J. (2020). Remarks on Price Index Methods for CPI Measurement Using Scanner Data [online]. *Statistika: Statistics and Economy Journal*, 100(1): 54–69. <https://www.czso.cz/documents/10180/125507867/32019720q1_54_bialek.pdf/f4ee19a0-75fd-41bf-b1fd-b192d177e125?version=1.2>.

BIALEK, J. (2021). Price Indices – a New R Package for Bilateral and Multilateral Price Index Calculations [online]. *Statistika: Statistics and Economy Journal*, 101(2): 122–141. <https://www.czso.cz/documents/10180/143550797/32019721q2_bialek.pdf/3cd5bf11-22f4-4ee5-b294-1d7d5909e4b4?version=1.2>.

DIEWERT, W. E. (1976). Exact and Superlative Index Numbers. *Journal of Econometrics*. 4(2): 115–145.
EUROPEAN COMMISSION. (2016). *Handbook on Price and Volume Measures in National Accounts*. Eurostat.
EUROPEAN COMMISSION. (2020). *Practical guidelines on web scraping for the HICP*. Eurostat.
FIXLER, D. (1993). The Consumer Price Index: Underlying Concepts and Caveats. *Monthly Labor Review*, 116.
LIPPE, P. (2012). Covariances and relationships between price indices: Notes on a theorem of Ladislaus von Bortkiewicz on linear index functions [online]. *MPRA Paper No. 38566*. <https://mpra.ub.uni-muenchen.de/38566>.
ROJÍČEK, M., SIXTA, J. (2022). *Machine Learning Process in the Production of Statistics*. [online]. <https://lsvdv-my.sharepoint.com/:w:/g/personal/ingabal_stat_gov_lt/EZZ09VOQmelOoaKsHPst3DUBpGVtY0d_emsGpmTmfBPQAw?e=L4WKjx>.
SAMUELSON, P., NORDHAUS, W. (2009) *Economics*. 19th Ed. NY: McGraw-Hill.
SCHULTZ, H. (1939). A Misunderstanding in Index-Number Theory: The True Konüs Condition on Cost-of-Living Index Numbers and Its Limitations. *Econometrica*, 7(1): 1–9.