INTERNATIONAL MONETARY FUND

IMF Country Report No. 22/141

REPUBLIC OF KAZAKHSTAN

TECHNICAL ASSISTANCE REPORT—SCANNER DATA FOR CPI MISSION

May 2022

This Technical Assistance report on Republic of Kazakhstan was prepared by a staff team of the International Monetary Fund. It is based on the information available at the time it was completed on November 2021.

Copies of this report are available to the public from

International Monetary Fund • Publication Services
PO Box 92780 • Washington, D.C. 20090
Telephone: (202) 623-7430 • Fax: (202) 623-7201
E-mail: publications@imf.org Web: http://www.imf.org
Price: $18.00 per printed copy

International Monetary Fund
Washington, D.C.

© 2022 International Monetary Fund

TECHNICAL REPORT

REPUBLIC OF KAZAKHSTAN

Scanner Data for CPI Mission

November 8–12, 2021

NOVEMBER 2021

PREPARED BY

Vanda Guerreiro

CCAMTAC

MEMBERS

Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyz Republic, Mongolia, Tajikistan, Turkmenistan, Uzbekistan

PARTNERS

Asian Development Bank, China, Korea, Poland, Switzerland

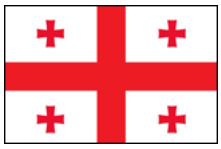
DISCLAIMER

The contents of this report constitute technical advice provided by the staff of the International Monetary Fund (IMF) to the authorities of Kazakhstan (the "TA recipient") in response to their request for technical assistance. This report (in whole or in part) or summaries thereof may be disclosed by the IMF to IMF Executive Directors and members of their staff, as well as to other agencies or instrumentalities of the TA recipient, and upon their request, to World Bank staff, and other technical assistance providers and donors with legitimate interest, including members of the Steering Committee of CCAMTAC, unless the TA recipient specifically objects to such disclosure (see Operational Guidelines for the Dissemination of Technical Assistance Information). Publication or Disclosure of this report (in whole or in part) or summaries thereof to parties outside the IMF other than agencies or instrumentalities of the TA recipient, World Bank staff, other technical assistance providers and donors with legitimate interest, including members of the Steering Committee of CCAMTAC, shall require the explicit consent of the TA recipient and the IMF's Statistics Department.

MEMBERS

Armenia

Azerbaijan



Georgia

Kazakhstan

Kyrgyz Republic

Mongolia

Tajikistan

Turkmenistan

Uzbekistan

PARTNERS

ADB

Ministry of Economy and Finance

Ministry
of Foreign Affairs
Republic of Poland

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Economic Affairs SECO

CONTENTS

Glossary 4

SUMMARY OF MISSION OUTCOMES AND PRIORITY RECOMMENDATIONS 5

DETAILED TECHNICAL ASSESSMENT AND RECOMMENDATIONS 6
A. Data Specifications 6
B. Data Analysis and COICOP Classification 8
C. Methodology 8
D. Officials Met During the Mission 11

FIGURE
1. Combining Different Data Sources 11

TABLES
1. Priority Recommendations 6
2. Key Variables of a Scanner Data File 7
3. Splicing Method 10

Glossary

BNS	Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan
CCAMTAC	Caucasus, Central Asia, and Mongolia Regional Capacity Development Center
COICOP	Classification of Individual Consumption According to Purpose
CPI	Consumer Price Index
EA	Elementary Aggregate
EAN	European Article Number
HBS	Household Budget Survey
MoU	Memorandum of Understanding
RPPI	Residential Property Price Index
SD	Scanner Data
SRC	State Revenue Committee
TA	Technical Assistance
WTPD	Weighted Time Product Dummy

SUMMARY OF MISSION OUTCOMES AND PRIORITY RECOMMENDATIONS

1. The purpose of the mission was to assist the Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan (BNS) with continuing its modernization of the Consumer Price Index (CPI). This was the first technical assistance (TA) mission to Kazakhstan on scanner data (SD). The mission was delivered remotely.

2. BNS aims at introducing scanner data in the CPI as a new data source by January 2024. Experimental indices should be tested during 2023 and 2024. The mission provided an overall view of the data collection process and requirements, the mapping of products to a national breakdown of the Classification of Individual Consumption According to Purpose (COICOP) 1999, and the main methods for compilation.

3. An increase in the quality of the CPI, and a reduction in CPI production costs and respondent burden are to be expected. In addition to the expected enhancement of the quality of the CPI, resource savings can be realized in the long run by reducing in-store price collection and data processing, since data are transmitted automatically, and all processing is further automatized. Nevertheless, the research and implementation phases require an investment namely in human resources skills and IT infrastructure.

4. BNS is expecting to have access to scanner data thru an online platform made available by the State Revenue Committee (SRC). This system is currently a work in progress and should be implemented by the first quarter of 2022. The SRC is currently completing the data received from the retailers with introducing product codes.

5. Meetings with the SRC are needed to secure data transmission, to agree on certain terms, and clarify data contents. Sample data should be obtained as soon as possible to guarantee that the data structure fits the needs of the CPI compilation. It is crucial to confirm whether the product codes being introduced by SRC are aligned to tight product specifications

6. The mission recommended that scanner data be introduced in a stepwise approach beginning with food products and a small sample of the larger retailers. To speed up the process of having access to data, BNS should recommend to the SRC to focus on coding the food products.

7. Once a decision is taken on the methodology to use, and a sufficient amount of data have been received, a follow up mission will be needed to provide practical guidance. The mission focused on the two methods mostly used in other countries: the dynamic basket approach and the multilateral method, weighted time product dummy.

To support progress in the above work areas, the mission recommended a detailed one-year action plan with the following priority recommendations carrying particular weight to make headway in improving the CPI.

Table 1. Priority Recommendations

Target Date	Priority Recommendation	Responsible Institutions
March 2022	Obtain data with the required characteristics.	**BNS and SRC**
July 2022	Using three months of data to classify Food products to the COICOP.	**BNS**

Further details on the priority recommendations and the related actions/milestones can be found in the action plan under *Detailed Technical Assessment and Recommendations*.

DETAILED TECHNICAL ASSESSMENT AND RECOMMENDATIONS

Priority	Action/Milestone	Target Completion Date
Outcome: Compile experimental SD-based indexes		
H	Introduce scanner data with food products and a small sample of the larger retailers.	January 2024
H	Have the SRC focusing on coding the food products.	March 2022
H	Meet with SRC to clarify the characteristics of the data.	March 2022
M	Created a variable identifying the region and the district to which each observation pertains.	March 2022
M	Download data at least every week to save, analyzed and classified to the COICOP	Continuous
M	Use the first three weeks for the index calculation.	Continuous
M	Create a conversion table to link the product code to the National breakdown of the COICOP.	September 2022
M	Sum the turnover and the quantities of the weekly data.	Continuous
M	Choose the compilation method.	September 2022

A. Data Specifications

8. BNS is expecting to have access to scanner data thru an online platform made available by the SRC. This system is currently a work in progress and should be implemented by

the first quarter of 2022. The SRC is currently completing the data received from the retailers with introducing product codes.

9. It is recommended to introduce scanner data in a stepwise approach beginning with food products and a small sample of the larger retailers. To speed up the process of having access to data, BNS should recommend to the SRC to focus on coding the food products.

10. A meeting with SRC should be held to clarify the characteristics of the data. Sample data should be obtained as soon as possible to guarantee that the data structure fits the needs of the CPI compilation. It is crucial to confirm that the product codes being introduced by SRC respect tight product specifications. For example, a bottle of water of Brand A with 0,5 liters and exactly the same but with 1 liter are different products and should have different codes. Also, should be confirmed that the key variables as listed in Table 2 will be available.

Table 2. Key Variables of a Scanner Data File

Variable description
Start date of the week
End date of the week
Code of the product
Brand of the product
Quantity in the package
Product description
Turnover
Total quantities sold
Sold by units or by weight
Region
District
Retailer

11. Currently data will be available for three retailers. BNS will also receive information on the name and retailer address and with that a variable should be created identifying the region and the district to which each observation pertains.

12. Data should be downloaded from the online system at least every week. This practice will create a redundancy to make the process safer. Data are downloaded, analyzed, and classified to the COICOP on a weekly basis. The first three weeks should be used for the index calculation.

Recommended Actions:

- Introduce scanner data with food products and a small sample of the larger retailers.
- Have the SRC focusing on coding the food products.
- Meet with SRC to clarify the characteristics of the data.

- Created a variable identifying the region and the district to which each observation pertains.
- Download data at least every week to save, analyzed and classified to the COICOP
- Use the first three weeks for the index calculation.

B. Data Analysis and COICOP Classification

13. At this stage the mission could not provide extensive training in data cleaning and analysis. However, some general recommendations were provided. All data handling should be done in R since this allows for reproducibility and is less prone to errors. The following steps can be performed as soon as data are available:

- Checking the file structure
- Checking the size of the files
- Checking the number of items per period and district/region
- Checking for duplicate records
- Checking for missing values
- Checking for non-positive values in quantities sold or turnovers
- Checking the sum of turnover over time.

14. A conversion table should be created to link the product code to the National breakdown of the COICOP. This table is very time consuming in the beginning. Nevertheless, will become increasingly faster with time since only new products in the market need to be classified manually. The conversion table should have four columns: Product code, Product description, COICOP code (National Classification level), COICOP description.

15. The product labels can also be used for the classification using machine learning and text mining techniques. Nevertheless, it is recommended to use supervised machine learning for which the learning dataset is essential.

Recommended Action:

- Create a conversion table to link the product code to the National breakdown of the COICOP.

C. Methodology

16. The turnover and the quantities of the weekly data are summed for the three weeks by product code and district. After the unit price is calculated by dividing the turnover by the quantities.

17. There are a number of methods used to compile CPI with scanner data. The mission focused on the two methods mostly used in other countries: the dynamic basket approach and the multilateral method, weighted time product dummy.

18. The dynamic basket approach follows a matched model with a monthly re-sample. Every month a new sample is drawn by using the market shares of the products. This process is made after the data are classified with the COICOP and at the lowest level, i.e., the national classification. Excel files were provided with examples and the formula that is used for the sampling is the following:

$$\frac{s_m + s_{m-1}}{2} > \frac{1}{n * \lambda}$$

λ	1,25
n	Number of matching products
s_m	Market share of a product in month m
s_{m-1}	Market share of a product in month m-1

19. The sampling uses the products which were transacted in two consecutive months. Its market shares are calculated, and the product is included if its share is higher then the threshold. The threshold varies according to the number of products existing in consecutive months for that specific COICOP.

20. After selecting the products to be included in the sample the index compilation follows as the regular CPI. Monthly changes of prices are calculated, and a Jevons (geometric average) is used to aggregate the prices to the national COICOP level. An index by national COICOP level is obtained that will be included in the regular CPI.

21. Products that were in the previous sample and went missing will continue being included in the sample and their prices are estimated. In the dynamic approach context, the concept of missing product relates to a product that was selected for the sample in some period and in the subsequent periods was not transacted. While the estimations for the in-person price collection are made for three months only, except for the seasonal products. For scanner data all missing products are estimated for 14 months. The methodology for the price estimations is the same as for the in-person collected prices: the previous price multiplied by the rate of change of the same COICOP.

22. Outliers and dumped prices are also identified and replaced by estimated prices. Thresholds should be created for price variations higher than 300 percent and lower then 75 percent. Dumped prices are those when turnover decreases more than 75 percent and prices more than 70 percent. A filter for these should also be created. When outliers and dumped prices are identified the "wrong" price is replaced by an estimated price.

23. The use of multilateral methods has been increasing since it makes use of all the data available, without sampling, and uses the turnover as weights without causing downward bias. There are several methods but one of the most currently used is the WTPD.

24. The WTPD method uses a weighted least squares regression, with turnover shares of products acting as weights:

$$S_i^t = \frac{p_i^t q_i^t}{\sum_{i \in N_t} p_i^t q_i^t}$$

The model is defined as follows:

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t$$

[0 , T] is the time window

N is the number of available products

D_i^t is a dummy variable that takes the value 1 to indicate the presence of product i in period t and 0 otherwise

D_i is a dummy variable that takes the value 1 if the observation relates to product i and 0 otherwise

25. A regression is estimated using weighted least squares and the indices are derived by exponentiating the parameters δ^t. In practice, a time window of 25 months is recommended. Therefore, 25 months of data are pulled in a regression using a weight least square function. From this regression, coefficients for the time dummy's (the periods/months) are obtained. The indexes are the calculated by exponentiating these coefficients.

26. A splicing method should be used when using rolling window methods as it is the case. Several splicing methods exist however the mission recommends the one exemplified in the CPI manual and that can be seen in the example bellow. Note that the example uses a window of 13 months instead of 25:

Table 3. Splicing Method

Period	1	2	3	4	5	...	11	12	13	14	15
First Compilation Round in Period 13	100	100.7	100.6	101.6	102.7	...	104.3	106	103.8		
Second* Compilation Round in Period 14		100	100.2	101.1	102.2	...	103.8	105.5	103.3	104.6	
Third Compilation Round in Period 15			100	101	102	...	103.5	105.3	103.2	104.4	104.1
Published Index (movement splice)	100	100.5	100.6	101.6	102.7	...	104.3	106	103.8	**105.1**	**104.8**

The splicing starts in period 14 (shown in bold). The published indices for periods 1–13 are obtained at the first compilation round. The published index in period 14 is obtained by applying the change between period 13 and period 14 indices of the second compilation round to the published index of period 13 (103 8 × 104.6/103 3 = 105.1). The published index in period 15 is obtained by applying the change between period 14 and period 15 indices of the third compilation round to the published index of period 14(105 1 × 104 1/104 4 = 104 8).

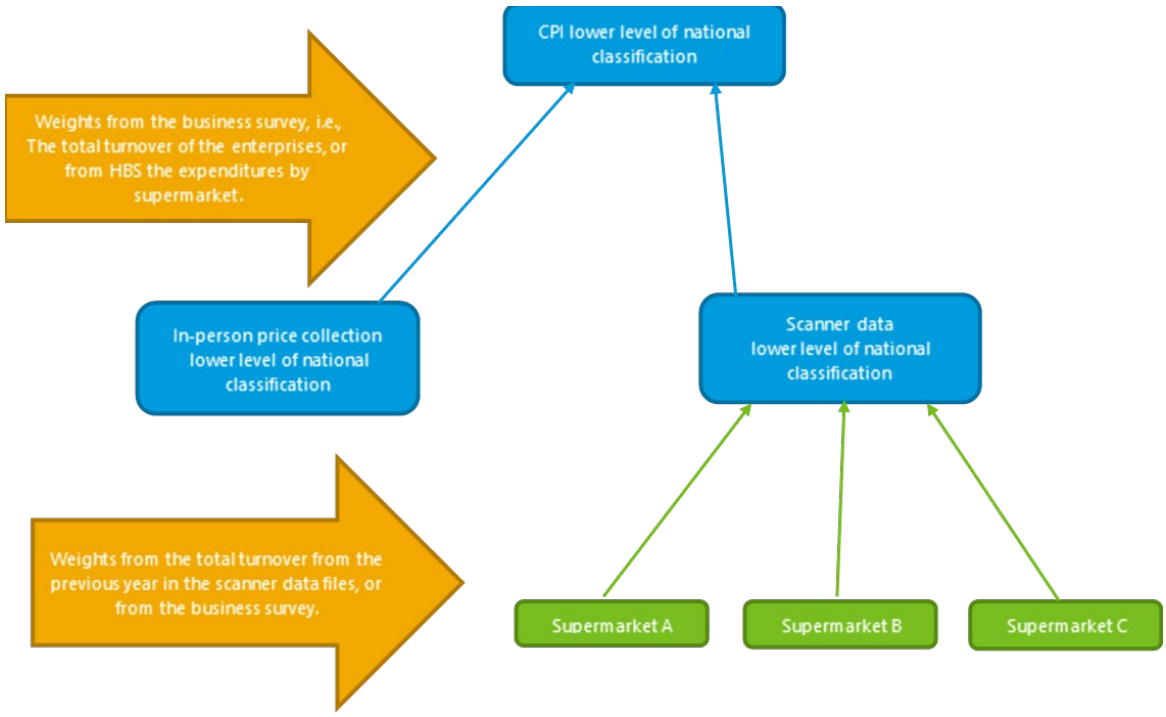
Source: Table 10.2 of the CPI Concepts and Methods Manual – 2020.

27. Once a decision is taken on the choice of methodology, and a sufficient amount of data have been received, a follow up mission will be needed to provide practical guidance. To begin the calculation of experimental indices with scanner data BNS should decide the method taking into account that the experiments with the dynamic basket can be started with 6

months of data and WTPD needs a minimum of 13 months of data. It is a current practice to begin with the dynamic approach and move to a multilateral index after some time say 2 years.

28. Once sub-indices are obtained from the scanner data these need to be aggregated with the sub-indices produced from in-person price collection. The mission provided examples in excel and the figure 1 bellow shows the aggregation schema.

Figure 1. Combining Different Data Sources



Recommended Actions:

- Create a conversion table to link the product code to the National breakdown of the COICOP.
- Sum the turnover and the quantities of the weekly data.
- Choose the compilation method.

D. Officials Met During the Mission

Name	Position, Institution
Kuanyshbek Zhakypbekuly	Director of the Price Statistics Department
Altynai Kumasheva	Head of the Consumer Price Statistics Directorate, National Statistics Bureau of Kazakhstan
Assel Zhunussova	Chief expert, National Statistics Bureau of Kazakhstan
Kanat Ukibasov	Chief expert, National Statistics Bureau of Kazakhstan
Meirambek Zhumadildin	Chief expert, National Statistics Bureau of Kazakhstan