Yayın Geliş Tarihi (Submitted): **30/05/2023**

Yayın Kabul Tarihi (Accepted): **13/07/2023**

Makele Türü (Paper Type): **Araştırma Makalesi – Research Paper**

Please Cite As/Atıf için:
Şen, E. and Arslan, O. (2023), A comparison of imputation methods in CPI calculations used by IWGPS organizations and imputation methods of robust cellwise outlier and missing data, *Nicel Bilimler Dergisi*, 5(2), 196-228. *doi: 10.51541/nicel.1307183*

A COMPARISON OF IMPUTATION METHODS IN CPI CALCULATIONS USED BY IWGPS ORGANIZATIONS AND IMPUTATION METHODS OF ROBUST CELLWISE OUTLIER AND MISSING DATA

Elif Şen[1] and Olcay Arslan[2]

ABSTRACT

In this study, imputation methods used by IWGPS (*) organizations in CPI (Consumer Price Index) calculations in case of missing data are discussed. Depending on the development of technological devices, methods suitable for the demand of collecting data and producing statistics from the field immediately, in a way that can be adapted to the data collection tools of statistics offices have been proposed. While the immediate imputation advantages of the proposed methods are mentioned, the proposed imputation results are compared with the method results used in the current practice and imputation results of cellwise outlier and missing data in the statistical computer programming language. The $method3(i_m\text{üd}19)$ proposed to assist the imputation tools used in CPI calculation all over the world and produced from the statistics is intended to provide convenience to all users. It can also be considered as a common weighted imputation method for both cellwise outlier and missing data case.

Keywords: Cellwise Outlier, Missing Value, Imputation, Robust Estimation Methods, CPI Imputation Methods of IWGPS Organizations

[1] Corresponding Author, Dr., Turkish Statistical İnstitue, Diyarbakır Regional Office, Diyarbakır, Türkiye. ORCID ID: https://orcid.org/0009-0008-0267-9287
[2] Prof. Dr., Statistics, Faculty of Science, Ankara University, Ankara, Türkiye. ORCID ID: https://orcid.org/0000-0002-7067-4997

IWGPS KURULUŞLARININ KULLANDIĞI TÜFE HESAPLAMALARINDAKİ İMPUTASYON YÖNTEMLERİ İLE SAĞLAM HÜCRESEL AYKIRI DEĞER VE KAYIP VERİ İMPUTASYON YÖNTEMLERİNİN BİR KARŞILAŞTIRMASI

ÖZ

Bu çalışmada IWGPS (*) kuruluşları tarafından kayıp veri durumunda TÜFE (Tüketici Fiyat Endeksi) hesaplamalarında kullanılan imputasyon yöntemleri ele alınmaktadır. Teknolojik cihazların gelişime bağlı olarak, istatistik ofislerinin veri derleme araçlarına uyarlanabilecek şekilde, anında alandan veri derleme ve istatistik üretme talebine uygun yöntemler önerilmiştir. Önerilen yöntemlerin anında imputasyon avantajlarından bahsedilmekle birlikte öneri imputasyon sonuçları, mevcut uygulamada kullanılan yöntem sonuçlarıyla ve istatistik paket programındaki hücresel aykırı değer ve kayıp veri imputasyon sonuçları ile karşılaştırılmıştır. Tüm dünyada TÜFE hesaplamasında kullanılan ve istatistiklerden üretilen imputasyon araçlarına yardımcı olmak için önerilen $yöntem3(i_m üd19)$ ün, tüm kullanıcılara kolaylık sağlaması amaçlanmıştır. Hem hücresel aykırı değer hem de kayıp veri durumu için ortak bir ağırlıklı imputasyon yöntemi olarak da düşünülebilir.

Anahtar Kelimeler: Hücresel Aykırı Değer, Kayıp Veri, Değer Atama, Sağlam Tahmin Yöntemleri, IWGPS kuruluşlarındaki TÜFE Değer Atama Yöntemleri.

1. INTRODUCTION

It is of great importance to obtain complete data for statistics and statistical applications. Similarly, the problem of outliers is also encountered in the dataset obtained by the researchers during the study process. The missing value and outlier structures in the dataset are completely different situations. First of all, the concepts of outlier and missing value encountered in the literature are mentioned. While there are many studies in multivariate datasets, cellwise and casewise outlier problems are generally together or only casewise outlier, few investigations have been encountered for the robust method used in imputation only in the case of a cellwise outlier. In this study, methods of coping with the

problem encountered only in the case of cellwise outlier in the multivariate dataset are discussed.

The dataset containing missing data in "Consumer Price Index (CPI) Manual, Concepts and Methods 2020" published by IWGPS organizations has been discussed. This problem in the manual, which is encountered in the situation of missing data in the multivariate dataset, to is dealt with and missing data here is, evaluated as an outlier; imputation process has been done with robust outlier imputation methods. In other words, for this missing dataset, which was converted to outlier at the general price level with zero (0) price assignment to missing data, robust outlier imputation was made to missing data with a computer programming language. Alternative methods have been proposed for the imputation methods applied in the case of missing data in CPI calculations by IWGPS organizations included in the manual. With these proposed methods, imputations were calculated for missing data in the same dataset. For the same dataset, the results of imputation methods in the CPI calculations used IWGPS organizations, the results of robust outlier and missing data imputation in the computer programming language, and the results of the proposed methods imputation were compared.

Among the proposed methods, the results of $method3$ (i_mü$d19$) were similar to the results of robust imputation. The proposed method depending on CPI can be adapted according to the data structure in other panel data type studies. In case of a cellwise outlier in the multivariate dataset, instead of rejecting the outliers, it is recommended to use the last observation, weighting with $method3$ to never be zero, and the specified marking imputation method to be used in other panel data studies.

2. METHODOLOGY

The concepts of missing data and outlier are two completely different problems that statistical science has to overcome.

The situation of missing data in the dataset for various reasons creates a problem when performing data analysis. Loss of information based on causes should be completed with appropriate techniques. Over time, as the interest in this subject increased, the number of studies also increased. Even though missing data analysis studies started in the 1930s, major breakthroughs have began with the advent of maximum likelihood estimation techniques and multiple imputation in the 1970s, along with advances in computer technology. According to

Rubin (1976), in some articles on multivariate normality (Wilks 1932, Anderson 1957, Afifi and Elashoff 1966, Hocking and Smith 1968, Hartley and Hocking 1971), the supposition about the process causing the missing data looks like to be that each value in the dataset is evenly missing. He also said in other articles on analysis of variance (Hartley 1956, Healy and Westmacott 1956, Wilkinson 1958, Rubin 1972, 1976) that the worthies of the dependent variables are presumed to be missing data regardless of the values to be observed. Rubin's (1976) classification of missing data mechanisms has increased the interest in missing data research. When we look at the literature, studies on imputation started with Rubin's work in 1976. The terminology of missing data and imputation used today was first used by Little and Rubin (Rubin, 1976).

Similar to the problems faced by researchers in this data collection process for missing data, outlier problems are also encountered in the dataset. Outlier value is “defined as an observation in a dataset that appears to be inconsistent with the rest of that dataset” (Barnett and Lewis, 1978). As it is known, it is assumed that the observations in the datasets have the same structure, in other words, they come from the same distribution. With the help of graphical methods and statistical tests, information about outliers can be obtained. However, when more than one variable is considered, it is very difficult to determine the observations that do not fit the general trend of the data in a multivariate structure and it is necessary to use different methods to determine outliers. After determining the outlier in the dataset; appropriate method selection is made by removing the outlier from the dataset, using statistical methods that will reduce the effects of the outlier or assigning a new value instead of an outlier. In the analysis phase, after the outlier was detected and determined, as a result of the recent investigations, the researchers decided that instead of removing the outliers from the dataset; They tried to minimize the negative effects of imputation value assignment or outliers on parameter estimation with robust estimation methods. It should not be forgotten that in addition to the problems of determining the outlier and imputation, the estimations to be made in the presence of this outlier are another subject of study. In order to cope with this problem, there are studies on robust estimator methods that help us make meaningful estimations about the population in the statistical research to be done.

In the literature review, in order to deal with the cellwise and casewise outlier problem in multivariate structure; methods that take into account correlation structures, methods for estimating inverse of covariance and covariance matrix, predictions of location and scatter

matrix and their robust predictions, generalized *S*-estimators, *M*-estimators, *Lasso* type variable selection operators etc methods are proposed. However, the common thought that draws attention here is; cellwise and casewise outliers in multivariate datasets are generally considered together and large literature studies are encountered for casewise outliers. It has also been observed that there is insufficient literature study for cellwise outlier only and that robust methods can deal with casewise outlier while cellwise outlier are not easily resolved. Accordingly, in this study, robust methods have been tried to be used to overcome this problem in case of solely cellwise outlier in the multivariate dataset.

2.1. Missing Data

The two main methods encountered in the literature as a resolution method in case of missing data are defined as data deletion and imputation methods as follows.

2.1.1. Data Deletion Methods

Removing missing data observations from the dataset can lead to a considerable reduction in the size of observations, thus data deletion will reduce the power of the statistical analysis to be performed. This problem will affect us to make meaningful estimations about the population in the statistical research to be done. This negative situation in real life is an example of the structure we encounter as the problem of reducing sample size in the theoretical science of statistics. It is clear that missing data causes a reduction in n sample dimension. It is conversant that if homogeneity of population is high, sample size to be taken decreases. However, as the differentiation between the units in the population grows, it is necessary to take a large sample size for accurate statistical analysis. The large sample size reduces standard error. The larger the standard error, the less its ability to estimate the parameter of the population or to represent the population. That is, if there are more missing cells, n sample size decreases; can cause bias, it can increase the standard error, which reduces the ability to represent population or estimate the population parameter. In the literature, there are two basic methods as data deletion method, Listwise/Casewise data deletion and Pairwise data deletion method (Tabachnick and Fidell 1996; Schafer, 1999; Osborne, 2013).

2.1.2. Data Imputation Methods

Approximate imputation methods in missing data emerge as a way for researchers to both save time and effort, and to protect the data they collect. However, according to Little and Rubin (1987), unconscious imputations instead of missing data do not eliminate existing problems, but also create new problems that are more difficult to solve (Little and Rubin, 1987). Today, many different statistical packages provide users with the opportunity to deal with missing data in different ways. Since the analyses are based on data, it is not yet possible to use precise statements about which method would be more appropriate for which amount of missing in which variable structure. The imputation methods used in the case of missing data are shortly described under the following subheadings.

2.1.2.1. Mean Imputation in Case of Missing Data

A simple to apply method for estimating missing data is the mean imputation method. The missing data is completed by assigning the overall sample mean of all other data instead of the missing data. The mean imputation method first creates the imputation cells using some auxiliary variables and after modifies the missing values in each cell with the sample mean. Also, missing data is able to be estimated using mode or median values. The fact that the median is not affected by outliers is a robust feature compared to the mean; in datasets with outliers, it would be beneficial to use median instead of mean.

2.1.2.2. Hot/Cold Deck Imputation in Case of Missing Data

Hot deck imputation is an important method that enables the missing value to be obtained from the dataset without any other mathematical and statistical information, and this is an intuitively meaningful method for many practitioners. We can think of it as a "hot" imputation because it can be processed instantly. In this method, the missing data is filled with the value obtained from an estimated distribution for the missing value from the existing data. This is done by estimating mean or mode of the data belonging to a population. Random Hot Deck imputation; the missing data is completed by randomly selecting one of the other data in the same population. Cold Deck imputation has a similar method to hot deck imputation; however, the selected welding must be different from the existent data welding (Acuna 2004).

2.1.2.3. Last Observation Carried Forward (*LOCF*) in Case of Missing Data

LOCF imputation method is a specific case of hot deck imputation. *LOCF* method detects the first missing data and uses the cell worth just before missing value to determine missing data. The procedure is reiterated for the next cell with the missing data until all missing data are assigned. It is not recommended to use this method (Molnar et al., 2008).

2.1.2.4. Regression Imputation in Case of Missing Data

Regression imputation supposes that the value of one variable varies linearly with other variables. This method uses estimates from the linear regression model, rather than completing the missing values with statistics. Because it is based on regression models, *X*and *Y* relationships are preserved. Therefore, it can be said that it provides an advantage over more simple imputation methods such as mean imputation. This method is based on the supposition of linear relationship between variables. However, the relationship between variables is generally non-linear. For this reason, linearly estimating missing data would be a biased model (Peng and Lei 2005).

2.1.2.5. Maximum Likelihood Methods (*ML*) in Case of Missing Data

Edgewort first used the maximum likelihood in 1908. In 1921, Fisher found the general formula for the variance of the estimator found by this method, and this method gained even more importance (İnal ve Günay, 1993). Allison (2001) revealed that maximum likelihood estimation and multiple imputation methods are more successful than traditional methods (Allison, 2001).

ML method uses all observed data in a dataset to generate the first order and second order moment estimates. It does not assign a prediction to any data, but instead creates a vector of mean-covariance matrix for variables in a dataset. This method is an improvement of the Expectation Maximization (*EM*) approach. Having an acknowledged statistical basis is an advantage. The supposition of original data distribution and the supposition of incomplete random missing is a disadvantage (Peng and Lei 2005).

2.1.2.6. Expectation Maximization Algorithm in Case of Missing Data

EM algorithm to get *ML* estimates in case of missing data is a much common method (Dempster et al., 1977; McLachlan and Krishnan 1997). It consists of an expectation stage and a maximization stage, thus is called *EM*. The *E*-stage works out the expectancy of all given efficient statistics data, given the observed data and the available parameter predictions. The *M*-stage updates parameter predictions via the maximum likelihood approach based on the available values of the complete efficient statistic. The algorithm is afterwards repeated multiple times in an iterative process until the difference between the last two successive parameter predictions converges to a certain point (Allison, 2001; Hu and Salvucci, 2001).

2.1.2.7. Multiple Imputation in Case of Missing Data

Multiple imputation (*MI*) estimates missing data with the help of random dataset. By repeating the process, completed datasets are obtained. The result is obtained by creating a single dataset by taking the averages of the obtained datasets. *MI* generally gives better and unbiased results than simple imputation methods. A negative feature of *MI* is that it requires more processing load and time for data processing and calculating estimates (Osborne, 2013).

2.2. Outlier

When the history of outliers was investigated, discrepant observations were first put forward by Bernoulli (1777) (Bernoulli, 1777). Chauvenet (1863) suggested rejecting values that do not meet his criteria (Chauvenet, 1863). Stone (1868) and Wright (1884) also proposed test statistics to reject outliers (Stone, 1868; Wright, 1884). In addition, various statistical tests have been proposed to identify and reject outliers in regression models. Anscombe (1960), Anscombe and Tukey (1963), Tietjen, Moore and Beckman (1973), Lund (1975), Rosner (1975), Ellenberg (1976), Cook and Weisberg (1980, 1982), Cook and Prescott (1981), Barnett and Lewis (1984) are the main ones (Çil, 1990). Apart from these, the determination of outliers is given by Grubbs (1969), Kale (1979), Hawkins (1980), Prescott (1980) and Rider (1933). A historical review of these methods is given by Stigler (1973, 1975) and Harter (1978) (Beckman and Cook, 1983). Instead of rejecting outliers, Glaisher (1872-73) after finding an approximate most likely outcome; has proposed a smaller weighting of these values than the other values (Glaisher, 1872-73)). Stone (1873) proposed

an alternative weighting method to Glaisher's method (Stone, 1873). Another weighting method was proposed by Newcomb (1886) (Newcomb, 1886).

In the 20th century, we have encountered studies of removing and not removing outliers from the dataset. Researchers emphasizing the importance of transforming data into information in the 21st century do not find it right to remove outliers from the observation set; they suggest using some estimation values (imputation) instead of these values or reducing the effects of outliers on parameter estimation with robust estimation methods.

2.2.1. Outlier Determination

Before determining the analysis methods to be used to identify outliers, the structure of the data should be considered. Let's try to give some examples of methods of determining the outlier affecting the dataset depending on the structure it is in. There are two commonly used methods for identifying outliers in univariate data: *Z-score* and Inter Quartile Range Method (*IQR*). In multivariate data; It can be determined by classical determinants such as Cook Distance, Mahalanobis Distance, Robust Distance Functions and Minimum Covariance Determinant Method. Outlier detection methods are also available, such as Standardized Error Terms, Student's Type Error Terms, Diagonals of Hat Matrix, DFBETAS Measure, DFFITS Measure, COVRATIO Measure and FVARATIO Measure, which are shown based on regression error terms. The methods shown based on the regression residuals are quite efficient in detecting outliers.

2.2.2. Outlier Analysis

After identifying the outliers in a dataset, researchers use these data; can be removed/deleted from the dataset or they can impute or can use statistical methods to reduce their effects.

2.2.2.1. Data Deletion Methods

Extract outlier from the dataset may result in data loss where that data is correct. For this reason, any data that is not certain to be an erroneous data should be excluded from the dataset. Time dependent, with few observation opportunities, for example, considering the multivariate data researched for cancer and its treatment; removing the observations

belonging to the row containing outliers from the dataset can lead to a serious reduction in the number of observations. Whether the outlier is removed from the dataset, the sample size n decreases; this can cause bias, increase the standard error, which reduces the ability to estimate the parameter of the population or to represent the population.

2.2.2.2. Robust Estimators in Case of Outliers

Instead of removing outliers from the dataset; it should try to avoid the negative effects of outliers on parameter estimation by imputation or robust estimation methods. In case of outliers, the main estimation methods encountered in the literature, in which situations they will be used and how they will be applied are briefly explained below (Huber, 1981).

L-estimators (Linear order statistics estimators)

It can be said that the use of L-estimators, estimators in the form of linear combinations of ordered sample values, is a generalization of the α-trimmed mean and winsorized mean approaches. Suppose the observations are ordered, $x_1 \leqq x_2 \leqq \cdots \leqq x_n$ then the statistic $T = n^{-1}\left(gx_{g+1} + x_{g+1} + x_{g+2} + \cdots + x_{n-h} + hx_{n-h}\right)$ is called the "Winsorized mean", it is acquired by winsorizing g the leftmost observation and h the rightmost observation (Huber, 1964).

Here the aim is to reduce the influence of outliers on the estimated value of μ, to use estimators in the form of linear combinations of ordered sample values $\mu = \sum c_i x_i$ where the weights of c_i are lower at the ends than those in the middle of the dataset. An example of such 'linear ordinal statistical estimators' (called 'L-estimators' by Huber, 1972) is the sample median with $c_i = 0$for all but the middle or two middle rank observations (Barnett and Lewis, 1978).

M-estimators (Maximum likelihood type estimators)

M-estimators, which Huber (1972) calls maximum likelihood type estimators, are obtained by solving an equation in the form of $\sum_{i=1}^{n} \psi(x_i - \mu) = 0$ to obtain an estimator μ where $x_1, x_2, \ldots, x_n$ is a random sample and $\psi(u)$ is a weight function with desired properties. For example, if $|\psi(u)|$ is small for large $|u|$, μ will reduce the effect of outliers in the sample and will protect against outliers. Special $\psi(u) = f'(u)/f(u)$ selection, where if the distribution has a probability density function of the form $f(x - \mu)$ it will give μ as a solution to the likelihood equation (Barnett and Lewis, 1978).

R-estimators (Rank test estimators)

First of all, Hodges and Lehmann (1963) stated that μ forecasters can be acquired by rank test methods such as the Wilcoxon test. Such R-estimators are generally robust. An example is given by Huber (1972) for a two-sample rank test for position shift. If the samples are $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_n$, the test statistic is $W(x| \ |1, x_2, \dots, x_n; y_1, y_2, \dots, y_n) = \sum_{i=1}^{n} J[i/(2n+1)]V_i$, where $J[i/(2n+1)]$ is an appropriately chosen function of the empirical distribution function for the composite sample. If the ith ordinal value in the composite sample is one of the x values, it is $V_i = 1$ (otherwise $V_i = 0$). We can derive an estimator μ as the solution to the equation $W(x| \ |1-\mu, x_2-\mu, \dots, x_n-\mu; -x_1+\mu, -x_2+\mu, \dots, -x_n+\mu) = 0$ and the asymptotic behavior of μ is derived from the power function of the test. For symmetrical distributions, the R-estimators for a suitable selection of $J(.)$ can be asymptotically efficient and normally distributed (Barnett and Lewis, 1978).

S-estimation method

Rousseeuw and Yohai (1984), generalizing the method that attempts to minimize a robust measurement of the scattering of LTS and LMS residues; described S-estimators corresponding to $Minimize S(\beta)$, where $S(\beta)$ is a certain kind of robust M-estimation of the scale of $r_1(\beta), \dots, r_n(\beta)$ residues. It has been shown that the breakpoints of S-estimators can also reach 50% with an appropriate selection of the relevant constants. The scale estimation $\hat{\sigma} = s\left(r_1(\hat{\beta}), \dots, r_n(\hat{\beta})\right)$ and $\underset{\beta}{Minimizes}\left(r_1(\beta), \dots, r_n(\beta)\right)$ are defined by minimizing the distribution of residuals. This estimator is called the S-estimator because it is produced from a scale statistic. The distribution $s\left(r_1(\beta), \dots, r_n(\beta)\right)$ is defined as the solution of the $\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{s}\right) = K$equation. In fact, the s given in this equation is an M-estimator of the scale. It can be said that S-estimators actually have the same asymptotic performance as regression M-estimators (Rousseeuw and Leroy, 1987).

MM-estimation method

Yohai (1985) presented a proposal for high-breakdown estimators such as LMS and LTS, which he called MM-estimators for higher efficiency. Yohai's estimators are defined in

three stages. In the first step, a high breakout estimate β , such as LMS or LTS, is calculated. Next, on the $r_i(\beta\ )$ residues, an M-estimation of the s_n scale with a 50% breakdown is calculated from the robust fit. Eventually, the MM-estimator $\hat{\beta}$ is described as any solution of $\sum_{i=1}^{n}\psi(r_i(\beta)/s_n)x_i = 0$ that satisfies $S(\beta) \leq S(\beta\ )$, where $S(\beta) = \sum_{i=1}^{n}\rho(r_i(\beta)/s_n)$. That is, the M-estimator is calculated with a function that gives zero weight to very high residuals. In the given equations, ρ is symmetric and continuously differentiable and $\rho(0) = 0$. $c > 0$ such that ρ is certainly strictly increasing over $[0, c]$ and constant over $[c, \infty|$. $E_H[\|x\|] < \infty$ and ρ are non-increasing functions for derivatives $\rho' = \psi$ and $\psi\,(u)/u, u > 0$ (Rousseeuw and Leroy, 1987).

3. EXAMPLE OF COMPARISON OF ESTIMATED VALUES IN CASE OF CELLWISE OUTLIER AND MISSING VALUE

Imputation methods used in consumer price index (CPI) calculations in case of missing data are discussed. In the calculation of these missing values with application of the programming language; missing data imputation, regression imputation and robust outlier imputation methods were used. Here, when making a robust outlier imputation; in the case of missing data in the multivariate dataset, zero (0) is assigned to the missing cells, at the general level of prices, the zero (0) price of the product for the consumer is considered an outlier. Thus, the robust outlier imputation method is applied to zero (0) outlier prices. Alternative methods have been proposed to the imputation techniques in case of missing data encountered in the CPI calculations in practice. All imputation results were compared.

As an example in this study, in CPI calculations; for the methods and methodology applied in the case of no product found during the survey period, “Consumer Price Index Manual, Concepts and Methods 2020” published jointly by the “Intersecretariat Working Group on Price Statistics (IWGPS)” organizations has been discussed.

Manual, consists of exhaustive knowledge and explications on compilation of the CPI. As deciding how to deal with different issues in the calculation of the CPI, Manual ensures also a conspectus of the methods and practices that national statistical offices (NSO) should take into consideration (Anonymous, 2020).

The index that gives the rate of change in the prices of goods and services subject to consumption over time is called CPI. The classical phases of computation are to calculate the

mean of the price change between periods for products and use the mean amount that the household spends on these products as the weight. It should be shared with the public as soon as probable (Anonymous, 2020).

Data Editing. There are three stages: (a) detection of probable errors and outliers, (b) confirmation and correcting of data, (c) output revise (Anonymous, 2020).

Temporarily Missing Products. Providing that the price statistician trusted that a missing item will be found again within an acceptable time, has three choices.

(a) Using matching pairs, the type where the price is missing is skipped to maintain a similar comparison with the previous period. The base index uses only data that the surveyor has acquired completely the same diversity of prices in the current and previous periods.

(b) Carryforward method is recommended solely in case of stable or arranged prices. Spite of that situation ensures price continuity in unobservable dates, it causes the movements to be biased in the absence of prices. If prices are generally increasing, the trend will be fall, and if prices are decreasing, the trend will be rise. Carryforward method is not commended, especially where periodic movements in the price index are significant or when there is lofty inflation.

(c) Imputation, the best resolution until now to assign a price. Imputation uses the ideally present knowledge to ensure an unbiased forecast of price action. There are basically two options:

(i) Missing price imputes with "relevance" to the mean price variation for the prices current in the base total (overall mean imputation). This supposes that the price variation for the missing item, if it were stock in the store, would be equal to the mean change in prices in the base total. If the base sum is fairly homogeneous, this may be an acceptable supposition.

(ii) Imputing the missing price by referring to the mean price variation (class mean imputation) for the prices of comparable species at alternate alike workplace. This symbolizes a more certain mate between the lost product and the products that provide the assigned price (Anonymous, 2020).

Table 1. Temporarily Missing Price Observations and Imputed Prices in Table 6.3 in the Consumer Price Index Manual 2020 (Anonymous, 2020).

Table 6.3 Temporarily Missing Price Observations and Imputed Prices

Outlets	Price Reference Period							
	Dec.-19	Jan.-20	Feb.-20	Mar.-20	Apr.-20	May.-20	Jun.-20	Jul.-20
A Supermarket	5.25	5.25	5.49	5.49	5.49	5.49	5.49	5.49
B Supermarket	5.10	5.10	5.10	5.25	5.25	5.25	5.25	5.25
C Supermarket	5.20	5.20	5.20	5.20	5.20	5.25	5.25	5.25
D Independent Trader	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
E Independent Trader	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
F Independent Trader	5.99	5.99	5.99	**6.13**	**6.15**	**6.17**	6.25	6.25
Overall Mean Imputation: F Mar:May								
Geometric Mean: A:E			5.54	5.67	5.69	5.71		
Geometric Mean: A:F	5.49	5.57	5.61	5.74	5.76	5.78	5.79	5.84
Short-Term Price Relatives: A:F	1.00000	1.01371	1.00748	1.02377	1.00352	1.00365	1.00198	1.00864
Long-Term Indices as Product of Short Term	100.00	101.37	102.13	104.56	104.92	105.31	105.52	106.38
Targeted Imputation: Independent Traders								
D Independent Trader	5.49	5.49	5.49	5.65	5.75	5.80	5.80	6.00
E Independent Trader	5.99	6.50	6.50	6.90	6.90	6.90	6.90	7.00
F Independent Trader	5.99	5.99	5.99	**6.26**	**6.32**	**6.34**	6.25	6.25
Geometric Mean: A:C & D:F	5.49	5.57	5.61	5.76	5.79	5.81	5.79	5.84
Short-Term Price Relatives: A:F	1.00000	1.01371	1.00718	1.02731	1.0044	1.00377	0.99753	1.00808
Long-Term Indices as Product of Short Term	100.00	101.37	102.13	104.92	105.38	105.78	105.52	106.37

Bold: Imputed values

Table 1 above in Anonymous 2020 (Table 6.3), It is an example of missing data and imputation of the workplace sales price of a specified product type, compiled from the field in order to calculate the CPI. If we evaluate this table as an $X_{n\times p}$ dimensional matrix; n row workplaces in the dataset; the p column shows the monthly sales price values of the specified product of the relevant workplace. Bold cells are cells that have been imputed. In this example, the price values for March, April and May 2020 are temporarily not compiled missing data from the F-Independent Trader workplace. For convenience, when the panel data entry arrangement in the shape of $isyerleri \times aylar$ of the data in Table 6.3 in the 2020 Manual; F-Independent Trader can defined as 6th workplace and cells 44th, 45th and 46th for the times March, April and May 2020, which is the missing cell of this workplace. Panel data entry arrangement in the shape of $isyerleri \times aylar$ of the data in Table 6.3 in the 2020 Manual, and results of imputed value of 44th, 45th and 46th missing cell by all methods are given in Table 3.

Geometric Averages

With the launch of the CPI Manual in 2004, great importance was given using the geometric mean when weights for individual prices are not available in the CPI base indices. The geometric price index is known as the Jevons price index. The formula "Jevons index number" is denoted by $I_j(I^0, I^t) = \prod_{i=1}^{nN} \left(\frac{p_i^t}{p_i^0}\right)^{1/n}$,where $p = price$ (Anonymous, 2020).

The geometric mean of workplace sales price of the product type specified with $X_{n\times p}$, represents the unit price of the relevant product, item or service types mentioned in the CPI calculation. Unit item prices, whose geometric mean is calculated and weighted; It is calculated as the index value by chained Laspeyres formula. In the CPI, the change of the index value at time t, obtained by weighting the geometric mean, compared to the previous $t-1$ time; it is called monthly rate of change (%). In Table 6.3 in the manual, monthly and index number change is named as "Short-Term Price Relatives: A:F" and "Long-Term Indices as Product of Short Term", respectively.

The Treatment Applications in the Manual

Overall Mean Imputation

In March-May excluding 6th workplace data, where there is primarily missing data, Geometric mean of February, March, April and May 2020 prices are calculated. For the missing data estimation, the short-term, monthly change rates of the full data are calculated and then the 6th workplace for the time of March-20 forecast is calculated. The missing data is completed with a similar step. The missing cells are imputed as the monthly change multiplied by the last observation, as well as the rate of change of the general mean. It is clear that Short-Term Price Relationships: A:F monthly change is for example 5.7934/5.781966=1.00198 for June 2020 and 5.84/5.79=1.00198 for July 2020. Long-term indices, that is, the index number, are calculated as 100x1.01371=101.37 with the change in the monthly rate of change by taking Base=100.

Targeted Imputation (class mean imputation)

The method is similar to the overall mean calculation method. Here, instead of the overall mean, the target is the mean within its own class. Geometric mean of February, March, April and May 2020 prices of Independent Traders D and E, excluding F, are calculated. By calculating the short-term, monthly change rates, the estimate of the 6th workplace for the time of March-20 is calculated. Missing data is completed with a similar step.

Carryforward Imputation

When the price of the missing type is registered again, there is probably to be a great compensatory stage replace in the index to remand to its appropriate value. If the type keeps unpriced for some time, the negative impact on the index will become more and more severe. The lost price in March 2020 is imputed with previous price of 5.99. Other missing data is completed with a similar step. In June, at the revert of the type, there will be a stepwise rising in price from 5.99 to 6.25 from May to June. In the abstract, carryforward is not recommended in the Manual (Anonymous, 2020).

In these examples only the F-Independent Trader data was missing, what if the missing had been more! There is no guarantee that the data will always be complete. Considering that both methods in Table 6.3 are assigned according to weights; If there are more missing, that is, more random missing cells, the prediction values generated from the current part will behave according to the structure of the existing data. That is,n sample volume reduction in case of more missing cells; can cause bias, increase the standard error, which reduces the ability of the population to estimate its parameter or to represent the population. What if there was a method that instantly imputes hotly and does not have to wait for the current data to be completed and the imputation to be calculated!

It was discussed in the previous section that the carryforward method used in the CPI application is a generalization of the Hot Deck method encountered in methodology. Recommended first method as an alternative for a rapid imputation for instant data collection; let be the weighted imputation method associated with monthly rate of change in same month of previous year. The purpose here is; while collecting data from the workplace, it is to send the data to the index calculators instantly, hotly. In this way, analysis for data quality is carried out instantly and rapid return is provided, detection of records whose department label

price update is neglected in the workplace thanks to the instant comparison opportunity, get instant predictions, and so on, meeting the fast needs that keep up with today's conditions are intended. Today, technological equipment and facilities are available in statistics offices so that there is instant data flow over mobile data with the data collection tool. It is obvious that the carryforward method, which is observed in the case of a temporary missing price, is a method for fast data flow under time constraints, does not require analysis, and fills the missing cell instantly. However, this method will either increase or decrease the geometric mean during periods of high change. In today's world after Covid-19, where inflation changes have increased significantly, weighted and marked according to the monthly change rate in the same month of the last year, especially Coicop food classification, method that can be calculated and imputed in the instant data compilation tool is defined as shown in Equation (1). Where $x_{t,miss}$: missing cell in t time, $x_{t-1,last}$: $(t-1)$ last observation in time,mr_{t-12}: monthly rate change of the previous year (monthly rate of change) show and is obtained from Table 2. While proposed 1st method is called $method1$, we can also name it as i_{mon} since it is imputed according to monthly change.

$$method1:\ i_{mon}: x_{t,miss} = x_{t-1,last} + \left(x_{t-1,last} * mr_{t-12}\right) \quad (1)$$

The basic idea here is to use the time series feature of the CPI. Under the assumption of the same product tracking, we can assume that the products exhibit the same price change behavior in certain periods of the year. For instance, the change in the prices of tomato product in November last year is expected to be in the same direction in November this year. We have $t-1$ change rates, where we can look at the monthly change in the previous year's trend of increase or decrease. Under the assumption of the same trend of change for the same product and the same time period; signs of these changes, marking the forecast as up or down; the proportional value of the monthly change of the previous year also provides the opportunity to give weighting. It will be sufficient for statistics offices to define this multiplicative weighting to the imputation of carrying forward the last observation with a command to be added to the software update in the field data collection tools. Instantly, the imputation of the temporary missing data will be completed with its weight, and it will quickly reach the index calculator via mobile data.

Table 2. G20 CPI all-items –group of twenty- mexico example (https://ec.europa.eu/eurostat, 2023) (Anonymous, 2023)

Data extracted on 08/05/2023 09:51:42 from [ESTAT]

Dataset:	**G20 CPI all-items - Group of Twenty - Consumer price index [PRC_IPC_G20__custom_4733163]**
Last updated:	**19/04/2023 11:00**
Time frequency	**Monthly**
Unit of measure	**Monthly rate of change**

TIME	Year	1	2	3	4	5	6	7	8	9	10	11	12
European Union - 27 countries (from 2020)	2019	-0.8	0.3	0.9	0.7	0.2	0.1	-0.3	0.1	0.2	0.2	-0.2	0.3
European Union - 27 countries (from 2020)	2020	-0.7	0.2	0.5	0.2	0.0	0.3	-0.2	-0.3	0.0	0.2	-0.3	0.3
Mexico	2019	0.1	0.0	0.4	0.1	-0.1	0.1	0.4	0.0	0.3	0.5	0.8	0.6
Mexico	2020	0.5	0.4	0.0	-0.1	0.4	0.5	0.7	0.4	0.2	0.6	0.1	0.4

To examine the application of $method1$), let's take the example of Mexico from the monthly rate of change data in the dataset created and exported for the G20 CPI - Twenty Group - Consumer Price Index 2019 and 2020 on the Eurostat page. The missing data estimation of the 6th workplace in Table 1 for March, April and May 2020 can be completed by using the values of 0.4, 0.1 and -0.1 respectively in March, April and May 2019 as the weighting criteria and weighting them as shown in Table 3, like the imputation calculations in the manual.

Table 3. Panel data entry arrangement in the shape of $isyerleri \times aylar$ of the data in Table 6.3 in the 2020 Manual, and results of imputed value of 44th, 45th and 46th missing cell with all methods

			current methods			proposed methods					R programming language imputation				
	isyerleri	aylar	overall	target	carry	method 1	method2	method3	method3_1	method3_2	R_miss	ImpReg	ImpMI	R_outlier	ImpImrob
1	1	12	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
2	1	1	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
3	1	2	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
4	1	3	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
5	1	4	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
6	1	5	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
7	1	6	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
8	1	7	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
9	2	12	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
10	2	1	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
11	2	2	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10	5.10
12	2	3	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
13	2	4	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
14	2	5	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
15	2	6	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
16	2	7	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
17	3	12	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
18	3	1	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
19	3	2	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
20	3	3	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
21	3	4	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20	5.20
22	3	5	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
23	3	6	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
24	3	7	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25	5.25
25	4	12	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
26	4	1	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
27	4	2	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49	5.49
28	4	3	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65	5.65
29	4	4	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75	5.75
30	4	5	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80
31	4	6	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80	5.80
32	4	7	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00
33	5	12	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
34	5	1	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50
35	5	2	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50	6.50
36	5	3	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
37	5	4	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
38	5	5	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
39	5	6	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90	6.90
40	5	7	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
41	6	12	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
42	6	1	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
43	6	2	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99	5.99
44	6	3	**6.13**	**6.26**	**5.99**	**8.39**	**7.61**	**6.183074**	**6.105651326**	**6.105651326**	***NA***	**6.1908**	**5.99**	**0.00**	**6.320582**
45	6	4	**6.15**	**6.32**	**5.99**	**9.22**	**9.23**	**6.376148**	**6.223535578**	**6.223535578**	***NA***	**6.2108**	**6.25**	**0.00**	**6.306284**
46	6	5	**6.17**	**6.34**	**5.99**	**8.30**	**7.61**	**6.183074**	**6.103375288**	**6.34369528**	***NA***	**6.2308**	**6.50**	**0.00**	**6.291986**
47	6	6	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25
48	6	7	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25

Bold: Imputed values

When imputed with$method1$), as cells 44th, 45th and 46th, the missing cells for the 6th workplace at the time of March, April and May 2020 are calculated 8.39, 9.22 and 8.30, respectively. The geometric mean for A:F is calculated as 6.05, 6.16 and 6.08 for the time March, April and May 2020, respectively. The monthly rate of change for A:F are calculated as 1.07859, 1.01899, 0.98558 and 0.95378 for the time March, April, May and June 2020, respectively. Index number change for A:F are calculated as 110.16, 112.25, 110.63 and 105,52 for the time March, April, May and June 2020, respectively.

After $method1$ imputation, it is clearly seen that; March, April and May 2020 missing data imputations for 6th workplace it is higher than the overall mean and targeted imputation methods; in June monthly change it is lower than the overall mean and the targeted imputation method.

Similarly, the second method proposed as an alternative; let the imputation method be marked based on the monthly rate of change of the previous year and weighted with the Gold Ratio. The method that can be calculated and imputed in the instant data collection tool, which is weighted with the golden ratio and marked as before, is defined as shown in Equation (2), especially for Coicop food classification. Where $x_{t,miss}$: missing cell in t time, $x_{t-1,last}$: $(t-1)$ the last observation in time, sgn): show the sign of the monthly rate change of the previous year, with the sign function sgn. Let's take the Gold Ratio as $\varphi = 1.618033$. While proposed 2st method is called $method2$, we can also name it as i_φ since it is imputed according to φ value.

$$method2:\ i_\varphi : x_{t,miss} = x_{t-1,last} + (sgn(mr_{t-12}) * \varphi) \quad (2)$$

The basic idea here is that the perfection of the golden ratio in the universe, is to investigate the effect in case of missing data. Similarly, we have proportional changes in our hands, where we can look at the monthly change of the previous year, increasing or decreasing trend. The signs of these changes are marking us in the forecasted upward or downward direction; The Gold Ratio value also provides the opportunity to give weighting.

To examine the implementation of $method2(i_\varphi)$, let's take the example of Mexico in Table 2. Here, using the positive/negative signs of 0.4, 0.1 and -0.1 values, respectively, in March, April and May 2019, with the weighting criterion φ, and weighted them as show in

Table 3, as in the previous imputation calculations, the 6th workplace in Table 6.3 in March, April and May The 2020 missing data forecast can be completed.

When imputed with $method2(i_{\varphi})$, as cells 44th, 45th and 46th, the missing cells for the 6th workplace at the time of March, April and May 2020 are calculated 7.61, 9.23 and 7.61, respectively. The geometric mean for A:F is calculated as 5.95, 6.16 and 5.99 for the time March, April and May 2020, respectively. The monthly rate of change for A:F are calculated as 1.06123, 1.03568, 0.97132 and 0.96776 for the time March, April, May and June 2020, respectively. Index number change for A:F are calculated as 108.38, 112.25, 109.03 and 105,52 for the time March, April, May and June 2020, respectively.

After $method2$ imputation, it is clearly seen that; March, April and May 2020 missing data imputations for 6th workplace it is higher than the overall mean and targeted imputation methods; in June monthly change it is lower than the overall mean and the targeted imputation method.

Third suggested $method3$; let the imputation method be marked based on the monthly rate of change of the previous year and weighted with $m\ddot{u}d19$. The method that can be calculated and imputed in the instant data collection tool, which is weighted with $m\ddot{u}d19$ and marked as previous, is defined as shown in Equation (3), specifically for Coicop food classification. Here $x_{t,miss}$: missing cell in t time, $x_{t-1,last}$: $(t-1)$ last observation in time, sign function sgn with $sgn(mr_{t-12})$: a show the sign of the monthly rate change of the previous year. The defined weight is determined as $m\ddot{u}d19 = 0.193074$. While proposed 3st method is called $method3$, we can also name it as $i_m\ddot{u}d19$ since it is imputed according to $m\ddot{u}d19$ value.

$$method3:\ i_m\ddot{u}d19: x_{t,miss} = x_{t-1,last} + (sgn(mr_{t-12}) * m\ddot{u}d19) \quad (3)$$

The basic idea here is a weighting that does not affect the fraction of the currency much, instead of the high Gold Ratio weight, and it is enough to move the stagnation in the carryforward imputation a little bit. Similarly, we have proportional changes that we can look at the monthly change of the previous year for an increase or decrease trend. The signs of these changes are marking us in the forecasted upward or downward direction; The $m\ddot{u}d19$ value also provides a lesser weighting opportunity than the Gold Ratio. In addition, the value

of $m\ddot{u}d19$ can be recommended as a weighting ratio for imputation studies with other weighting.

To examine the application of $method3(i_m\ddot{u}d19)$, let's take the example of Mexico from the monthly rate of change data in the dataset created and exported for the G20 CPI - Twenty Group - Consumer Price Index 2019 and 2020 on the Eurostat page. Here, using the positive/negative signs of 0.4, 0.1 and -0.1 values, respectively, in March, April and May 2019, with the $m\ddot{u}d19$ weighting criterion, and weighted them as show in Table 3, as in the previous imputation calculations, the 6th workplace in Table 6.3 in March, April and May 2020 missing data estimation can be completed.

When imputed with $method3(i_m\ddot{u}d19)$, as cells 44th, 45th and 46th, the missing cells for the 6th workplace at the time of March, April and May 2020 are calculated 6.183074, 6.376148 and 6.183074, respectively. The geometric mean for A:F is calculated as 5.75, 5.80 and 5.78 for the time March, April and May 2020, respectively. The monthly rate of change for A:F are calculated as 1.02517, 1.00808, 0.99792 and 1.00180 for the time March, April, May and June 2020, respectively. Index number change for A:F are calculated as 104.70, 105.55, 105.33 and 105,52 for the time March, April, May and June 2020, respectively.

After $method3$ imputation, it is clearly seen that; March, April and May 2020 missing data imputations for 6th workplace it is higher than overall mean imputation method, values around the targeted imputation method; monthly change in June it is lower than the overall mean and the targeted imputation method.

Finally, let's consider the change rates by completing the missing cells in this table with the robust outlier imputation and missing data imputation method in programming language.

Imputation with the mice command in RStudio, one of the programming language, is done in Appendix A **(see Appendix A)**. The disadvantage is that all data is expected to complete.

To examine the “mice (MI)” application in the programming language; The data estimation of March, April and May 2020 missing values in Table 6.3 is completed with the 2nd iteration value results in the "ImpMI" column in Table 3.

When imputed with R_Mice_Impute MI, as cells 44th, 45th and 46th, the missing cells for the 6th workplace at the time of March, April and May 2020 are calculated 5.99, 6.25 and 6.50, respectively. The geometric mean for A:F is calculated as 5.90, 6.07 and 5.93 for the

time March, April and May 2020, respectively. The monthly rate of change for A:F are calculated as 1.01977, 1.01006, 1.00962 and 0.99348 for the time March, April, May and June 2020, respectively. Index number change for A:F are calculated as 104.15, 105.20, 106.21 and 105,52 for the time March, April, May and June 2020, respectively.

Here it is clearly seen that; March, April and May 2020 missing data imputations for the 6th workplace have taken values around overall mean and targeted imputation method; monthly change in June it is lower than the overall mean and the targeted imputation method.

In addition, the missing data regression imputation with the mice by “norm.predict” is done in Appendix B **(see Appendix B**). The disadvantage is waiting for all data to complete.

To examine the “mice(reg)” application in the programming language; The data estimation of March, April and May 2020 missing values in Table 6.3 is completed as in the "ImpReg" column in Table 3.

When imputed with R_Mice_Impute Reg, as cells 44th, 45th and 46th, the missing cells for the 6th workplace at the time of March, April and May 2020 are calculated 6.1908, 6.2108 and 6.2308, respectively. The geometric mean for A:F is calculated as 5.90, 6.07 and 5.93 for the time March, April and May 2020, respectively. The monthly rate of change for A:F are calculated as 1.02538, 1.00347, 1.00358 and 1.00051 for the time March, April, May and June 2020, respectively. Index number change for A:F are calculated as 104.72, 105.08, 105.46 and 105,52 for the time March, April, May and June 2020, respectively.

Here it is clearly seen that; March, April and May 2020 missing data imputations for the F-Independent Trader, high values compared to the overall mean imputation method, low values compared to the targeted imputation method; in June monthly change it is lower than the overall mean and higher than the targeted imputation method.

Finally, the imputation with the “lmrob” command with RStudio is done in Appendix C (see Appendix C). The downside is that all data is expected to complete.

To examine the “lmrob (Computes fast MM-type estimators for linear (regression) models)” application in the programming language; The data estimation of March, April and May 2020 missing values in Table 6.3 is completed as in the "Implmrob" column in Table 3.

When imputed with R_outlier Impute lmrob, as cells 44th, 45th and 46th, the missing cells for the 6th workplace at the time of March, April and May 2020 are calculated 6.320582, 6.306284 and 6.291986, respectively. The geometric mean for A:F is calculated as 5.77, 5.79

and 5.80 for the time March, April and May 2020, respectively. The monthly rate of change for A:F are calculated as 1.02894, 1.00255, 1.00266 and 0.99888 for the time March, April, May and June 2020, respectively. Index number change for A:F are calculated as 105.08, 105.35, 105.63 and 105,52 for the time March, April, May and June 2020, respectively.

Here it is clearly seen that; March, April and May 2020 missing data imputations for the 6th workplace higher than the overall mean, values around the targeted imputation method; in June monthly change it is clearly seen that it is lower than the overall mean and higher than the targeted imputation method.

With the overall mean, target mean, forward carry imputation used in current practice, three alternatively proposed methods and let's also compare the descriptive statistics values of the imputation examples of the computer programming language:

```
> summary(ybkp)
 isyerleri      aylar      overallGmean     targGmean       carryGmean       method1          method2          method3
 1:8       1      : 6   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100
 2:8       2      : 6   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250
 3:8       3      : 6   Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490
 4:8       4      : 6   Mean   :5.723   Mean   :5.733   Mean   :5.713   Mean   :5.879   Mean   :5.848   Mean   :5.729
 5:8       5      : 6   3rd Qu.:6.032   3rd Qu.:6.062   3rd Qu.:5.990   3rd Qu.:6.062   3rd Qu.:6.062   3rd Qu.:6.046
 6:8       6      : 6   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :9.220   Max.   :9.230   Max.   :7.000
           (Other):12
   method3_1        method3_2         R_miss          ImpReg          ImpMI         R_outlier        Implmrob
 Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :5.100   Min.   :0.000   Min.   :5.100
 1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250   1st Qu.:5.250
 Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490   Median :5.490
 Mean   :5.723   Mean   :5.728   Mean   :5.695   Mean   :5.727   Mean   :5.729   Mean   :5.339   Mean   :5.733
 3rd Qu.:6.026   3rd Qu.:6.026   3rd Qu.:5.990   3rd Qu.:6.048   3rd Qu.:5.992   3rd Qu.:5.990   3rd Qu.:6.062
 Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000
                                 NA's   :3
```

```
> describe(ybkp,trim=0.05,type=3)
             vars  n mean   sd median trimmed  mad min  max range  skew kurtosis   se
isyerleri*      1 48 3.50 1.73   3.50    3.50 2.22 1.0 6.00  5.00  0.00    -1.34 0.25
aylar*          2 48 4.50 2.32   4.50    4.50 2.97 1.0 8.00  7.00  0.00    -1.31 0.33
overallGmean    3 48 5.72 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.86    -0.45 0.08
targGmean       4 48 5.73 0.58   5.49    5.71 0.43 5.1 7.00  1.90  0.81    -0.61 0.08
carryGmean      5 48 5.71 0.56   5.49    5.69 0.43 5.1 7.00  1.90  0.92    -0.28 0.08
method1         6 48 5.88 0.92   5.49    5.78 0.43 5.1 9.22  4.12  1.84     3.20 0.13
method2         7 48 5.85 0.84   5.49    5.77 0.43 5.1 9.23  4.13  1.80     3.68 0.12
method3         8 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.83    -0.55 0.08
method3_1       9 48 5.72 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.86    -0.44 0.08
method3_2      10 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.52 0.08
R_miss         11 45 5.69 0.57   5.49    5.66 0.43 5.1 7.00  1.90  1.00    -0.25 0.09
ImpReg         12 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.51 0.08
ImpMI          13 48 5.73 0.57   5.49    5.70 0.43 5.1 7.00  1.90  0.84    -0.55 0.08
R_outlier      14 48 5.34 1.50   5.49    5.51 0.43 0.0 7.00  7.00 -2.66     7.18 0.22
Implmrob       15 48 5.73 0.58   5.49    5.71 0.43 5.1 7.00  1.90  0.81    -0.61 0.08
```

```
> matrixplot(ybkp, interactive = T)
```

"with matrixplot {VIM} a matrix graph is created in which all cells of a data matrix are visualized with rectangles. Existing data is coded according to a continuous color scheme, while missing/imputed data is visualized with a clearly distinguishable color" (RStudio).

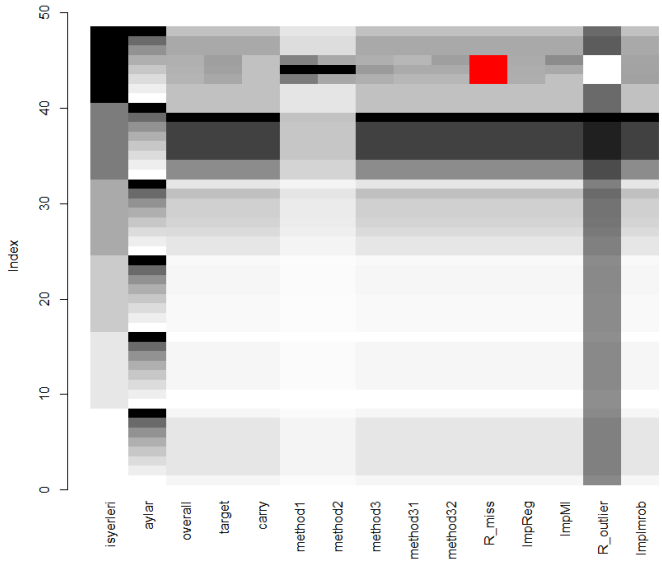

Figure 1. Matrix graph (matrixplot(ybkp)=> Plot Export)

Here, a comparison was made after the imputation of $method3_1$), whose $method3$) is for control purposes and defined by alternative calculation; it has been determined that $method3$ is closer to the targeted mean imputation used in the current application. To examine the effect of weight on the price penny, the $method3_1$) formula defined for control is calculated as shown in Equation (4),

$$method3_1 : i_m \text{üd} 19_1 : x_{t,miss} = x_{t-1,last} + \left(sgn(mr_{t-12}) * x_{t-1,last} * (\text{müd } 19/10)\right) \quad (4)$$

to examine the unmarked effect of the proposed method, the $method3_2$ defined for control is calculated as shown in Equation (5).

$$method3_2 : i_m \text{üd} 19_2 : x_{t,miss} = x_{t-1,last} + \left(x_{t-1,last} * (\text{müd } 19/10)\right) \quad (5)$$

Actually, $method3_1(i_m \text{üd} 19_1)$ imputation is the closest method to the overall mean imputation used in the current application. However, the purpose of marking with mr_{t-12};

due to the CPI data structure, it is to use the change depending on the previous information with the logic of add/subtract that will reflect the +/- marking, that is, the increase/decrease trend compared to the same month of the last year. With this study, it is aimed to make imputation instantly with software rules to be defined in the data compilation tool of the statistics offices and not to wait for the whole data compilation process. Depending on the data structure, the imputation method given with the formula for $method3_1(i_m$üd$19_1)$ can be recommended for other imputation studies where the "general geometric mean" is important and used and there is no time problem in waiting for the compilation of all the data.

Figure 2. Correlation analysis (Data Profiling Report)

4. CONCLUSION

In the CPI calculations, the methods and methodology applied in the case that no products are found during the survey period, contained in the 2020 CPI Manual jointly published by IWGPS organizations, from imputation methods in case of missing data; overall mean, target mean, forward carry imputations are discussed.

With the sample dataset used in the 2020 CPI Manual, three alternative methods have been proposed to the imputation techniques applied in case of missing data in CPI calculations.

In addition, robust outlier and missing data imputation calculations were made with the same data in the RStudio computer programming language. Here, the case of cellwise missing data in the multivariate dataset is evaluated an outlier value. For this, in the application study, zero (0) price is assigned to the missing cell in the multivariate dataset, and at the general level of prices, it is ensured that missing data is defined as outlier value. With the robust estimation method applied only in the case of cellwise outlier, missing data was evaluated as an outlier; and robust outlier imputation is done with "Impute_Lmrob".

Overall mean, target mean, forward carry imputation results from the imputation methods used in the current application of CPI; suggested three methods imputation results; RStudio computer programming language imputation results are compared.

According to the descriptive statistical results, the standard deviation (sd), distribution range, skewness, kurtosis and standard error (se) values in $method1(i_{mon})$ and $method2(i_{\varphi})$; the differences from the programming language robust imputations and $method3(i_m \ddot{u} d19)$ are clearly observed. It can be said that the proposed $method3(i_m \ddot{u} d19)$ shows similar descriptive statistical properties to the programming language robust imputations and is similar to the robust methods according to the results of the correlation analysis.

By assigning a zero (0) price to the missing cell in multivariate dataset, it defines the missing data as outlier at the general price level; a robust outlier imputation calculation has been made. Similar to this robust “Impute_Lmrob” imputation, with the imputation $method3(i_m \ddot{u} d19)$ weighted by $m\ddot{u}d19 = 0.193074$ and marked with monthly change, it can be said that for CPI applications, imputation can be made immediately, without waiting for all data to be collected, with a quick calculation in the data compilation tool.

The proposed $method3_1(i_m \ddot{u} d19_1)$ imputation for control purposes is the closest method to the overall mean imputation used in current practice. However, the purpose of marking with mr_{t-12}; due to the CPI data structure, it is to use the change depending on the previous information with the logic of add/subtract that will reflect the +/- marking, that is, the increase/decrease trend compared to the same month of the last year. Within this study, it is aimed to make imputation instantly with software rules to be defined in the data compilation tool of the statistics offices and not to wait for the whole data compilation process. Depending on the data structure, the imputation method given with the formula for $method3_1(i_m \ddot{u} d19_1)$ can be recommended for other imputation studies where the overall

geometric mean is important and used and there is no time problem in waiting for the compilation of all the data.

In addition, $method3(i_m\ddot{u}d19)$ suggested for statistical data analysis with robust estimation methods in case of cellwise outlier observation; marking on the CPI based on the monthly rate of change of the previous year can be adapted according to the data structure in other panel data type studies. In case of cellwise outlier observation in the multivariate dataset, instead of rejecting the outliers, using the last observation, the process of weighting with $m\ddot{u}d19 = 0.193074$ to never be zero and the specified marking imputation method are recommended to be used in other panel data studies as well.

ETHICAL DECLARATION

In the writing process of the study titled "A Comparison of imputation methods in CPI calculations used by IWGPS organizations and imputation methods of robust cellwise outlier and missing data", there were followed the scientific, ethical and the citation rules; was not made any falsification on the collected data and this study was not sent to any other academic media for evaluation.

ACKNOWLEDGMENTS

This paper "A comparison of imputation methods in CPI calculations used by IWGPS organizations and imputation methods of robust cellwise outlier and missing data" is produced from the "Hücresel aykırı gözlem olması durumunda sağlam tahmin yöntemleri ile istatistiksel veri analizi" doctoral thesis.

REFERENCES

Acuna, E. and Rodriguez, C. (2004), *The treatment of missing values and its effect in the classifier accuracy*. In D. Banks, L. House, F.R. Mc Morris, P. Arabie, W. Gaul (Eds). Classification, Clustering and Data Mining Applications. Springer-Verlag, 639-648, Berlin Heidelberg.

Allison, P. D. (2001), *Missing Data*. Sage Publications, Inc, (Quantitative Applications in the Social Sciences), 104, Pennsylvania, USA.

Anonymous (2020), *Consumer price index manual concepts and methods*. Identifiers: ISBN 978-1-51354-298-0. *https://www.ilo.org› publication › wcms_761444*. Erişim Tarihi: 15.01.2023.

Anonymous (2023), Unit of measure: *Monthly rate of change*. *https://ec.europa.eu/eurostat/databrowser/view/PRC_IPC_G20__custom_4733163/default/table?lang=en Data extracted on 31/01/2023 13:34:36 from [ESTAT] G20 CPI all-items - Group of Twenty - Consumer price index [PRC_IPC_G20__custom_4733163*]. Erişim Tarihi: 31.01.2023.

Barnett, V. and Lewis, T. (1978), *Outliers in statistical data*. John Wiley and Sons, 376, New York.

Beckman, R. J. and Cook, R. D. (1983), *Outlier……..s*. Technometrics, 25 (2), 119-149.

Bernoulli, D. (1977)*The most probable choice between several discrepant observations and the formation there from of the most likely ınduction*, Reprinted in Biometrika, 48, 1-18 (1961, translated by C. G. Allen), London.

Chauvenet, W (1863), *Method of least squares, appendix to manual of spherical and practical astronomy.* Vol 2. Lippincott, 469-566, 539-599, Philadelphia. Reprinted (1960) -5th edn. Dover, New York.

Çil, B. (1990) *Regresyon analizinde tek bir sapan değerin "outlier'ın" belirlenmesine ilişkin metodların mukayesesi*. Doktora Tezi, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara, Turkey.

Glaisher, J. W. L. (1872-73). *On the rejection of discordant observations*, Monthly Notices of the Royal Astronomical Society. 33, 391-402.

İnal, C. and Günay, S. (1993), *Olasılık ve matematiksel istatistik*, Hacettepe Üniversitesi Fen Fakültesi Beytepe Basımevi, 339- 349, Ankara.

Hu, M. and Salvucci, S. (2001)*, A study of imputation algorithms*, 122. Working Paper No. 2001-17, Washington, DC.

Huber, P. J. (1964), Robust estimation of a location parameter. *Ann. Math. Statist.* 35(1), 73-101.

Huber, P. J. (1981)*, Robust Statistics*. John Wiley and Sons, 320, New York.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical analysis with missing data*, (1st Ed.), John Wiley&Sons, 291, New York.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*, (2nd Ed.), John Wiley&Sons, 409, New Jersey.

Molnar, F. J., Hutton, B. and Fergusson, D. (2008), Does analysis using "last observation carried forward" introduce bias in dementia research?. *Canadian Medical Association Journal,* 179 (8), 751–753.

Newcomb, S. (1886), A generalized theory of the combination of observations so as to obtain the best result, *American Journal of Mathematics*, 8 (4), 343-366.

Osborne, J. W. (2013), *Best practices in data cleaning*, Sage Publication, Inc, 275, California.

Peng, Liu and Lei, L.A. (2005), *A review of missing data treatment methods*. Shanghai University of Finance and Economics, 8, Shanghai, P. R. China.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust regression and outlier detection*, John Wiley & Sons, Inc, 341, Canada. ISBN 0471-85233-3.

Rubin, D. B. (1976), Inference and missing data (with discussion), *Biometrika,* 63, 581–592.

Schafer, J. L. (1999), Multiple imputation: A primer, *Statistical Methods on Medical Research*, 8(1), 3-15.

Stone, E. J. (1868), On the rejection of discordant observations, *Monthly Notices of the Royal Astronomical Society*, 28, 165- 168.

Stone, E. J. (1873), On the rejection of discordant observations, *Monthly Notices of the Royal Astronomical Society*, 34, 9-15.

Student, (1927), Errors of routine analysis, *Biometrika,* 19, 151–164.

Tabachnick, B. and Fidell, L. (1996). *Using multivariate statistics* (8th ed.). Pearson Education, 1018, USA.

Wright, T. W. (1884). *A Treatise on the adjustment of observations by the method of least squares*. Van Nostrand, 298, New York.

Appendix A

"mice: Generates Multivariate Imputations by Chained Equations (MICE), Impute the missing data *m* times"

```
> library(plm)

> library(mice)

> pMiss<-pdata.frame(RStudioMiceMiss)

> # "imputing missing values MI with mice for variables in our RStudioMiceMiss and create 10 new imputed datasets (pmm: imputation by predictive mean matching)" (RStudio).

> multiple_imputation = mice(pMiss, seed = 1, m = 10, print = FALSE)

> multiple_imputation[["imp"]]
$R_Mice_Impute
      1    2    3    4    5    6    7    8    9   10
44 5.99 5.99 6.25 5.99 6.25 5.99 5.99 5.99 5.99 5.99
45 5.99 6.25 5.99 5.99 5.99 6.25 5.99 5.99 5.99 5.99
46 6.00 6.50 6.25 5.99 5.99 6.25 6.50 6.25 6.25 5.99
```

Appendix B

> # the missing data regression imputation with the mice (data, method ="norm.predict", seed=1, m=10, print= F) command with RStudio is done as follows

"mice: Generates Multivariate Imputations by Chained Equations (MICE), Impute the missing data *m* times"

> # "using the norm.predict modeling approach via the MICE package to generate precedent values in our RStudioMiceMiss (norm.predict_numeric_Linear regression, predicted values)" (RStudio).

```
> RegMiss_imputed <- mice(pMiss, method="norm.predict" ,seed=1, m=10, print= F)

> RegMiss_imputed[["imp"]]

$R_Mice_Impute
       1      2      3      4      5      6      7      8      9     10
44 6.1908 6.1908 6.1908 6.1908 6.1908 6.1908 6.1908 6.1908 6.1908 6.1908
45 6.2108 6.2108 6.2108 6.2108 6.2108 6.2108 6.2108 6.2108 6.2108 6.2108
46 6.2308 6.2308 6.2308 6.2308 6.2308 6.2308 6.2308 6.2308 6.2308 6.2308
```

Appendix C

```
> library(readxl)

> library(DataExplorer)

> library(ggplot2)

> library(plm)

> library(psych)

> library(dbplyr)

> library(robustbase)

> # lmrob (Computes fast MM-type estimators for linear (regression) models)

> control <- lmrob.control(method = "SMDM",compute.outlier.stats = c("S", "MM", "SMD", "SMDM"))

> fit1 <- lmrob(R_outlier ~ isyerleri*aylar, RStudioLmrobOutlier, control = control)
# "lmrob {robustbase}: The method argument takes a string that specifies the estimates to be calculated as a chain. Setting method='SMDM' will result in an intial S-estimate, followed by an M-estimate, a Design Adaptive Scale estimate and a final M-step" (RStudio).

> summary(fit1)

Call:
lmrob(formula = R_outlier ~ isyerleri * aylar, data = RStudioLmrobOutlier,
    control = control)
 \--> method = "SMDM"
Residuals:
    Min      1Q  Median      3Q     Max
-6.4453 -0.2673 -0.1438  0.3042  0.8519

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.83431    0.24131  20.034  < 2e-16 ***
isyerleri       0.20150    0.06288   3.205  0.00252 **
aylar           0.01058    0.05764   0.184  0.85518
isyerleri:aylar 0.01164    0.01500   0.776  0.44216
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.3966
Multiple R-squared:  0.5372,    Adjusted R-squared:  0.5057
Convergence in 8 IRWLS iterations

Robustness weights:
3 observations c(44,45,46) are outliers with |weight| = 0 ( < 0.0021);
33 weights are ~= 1. The remaining 12 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6779  0.8352  0.9353  0.8968  0.9884  0.9987
Algorithmic parameters:
      tuning.chi1       tuning.chi2       tuning.chi3       tuning.chi4                bb       tuning.psi1
       -5.000e-01         1.500e+00                NA         5.000e-01         5.000e-01        -5.000e-01
      tuning.psi2       tuning.psi3       tuning.psi4        refine.tol           rel.tol         scale.tol
        1.500e+00         9.500e-01                NA         1.000e-07         1.000e-07         1.000e-10
         solve.tol       eps.outlier             eps.x warn.limit.reject warn.limit.meanrw
         1.000e-07         2.083e-03         7.640e-11         5.000e-01         5.000e-01
     nResample            max.it           best.r.s           k.fast.s             k.max     maxit.scale       trace.lev              mts
           500                50                  2                  1               200               200               0             1000
   compute.rd       numpoints fast.s.large.n
            0                10          2000
                psi           subsampling                   cov compute.outlier.stats1 compute.outlier.stats2
              "lqq"          "nonsingular"         ".vcov.w"                       "S"                    "SM"
compute.outlier.stats3 compute.outlier.stats4
               "SMD"                  "SMDM"
seed : int(0)

> augment(fit1)
```

	R_outlier	isyerleri	aylar	.fitted	.resid
44	0.00	6	3	6.320582	-6.32058163
45	0.00	6	4	6.306284	-6.30628392
46	0.00	6	5	6.291986	-6.29198620