Measurement of two services components of the basket of the Brazilian Consumer Price Index using the Continuous National Household Sample Survey (PNADC) as data source

Lincoln T. da Silva[*], *Vladimir G. Miranda*[†], *Marcelo R. do Amaral, Plinio L. dos Santos, Gustavo V. Leite*
Department of Price Indexes - Directorate of Surveys - Brazilian Institute of Geography and Statistics - IBGE

Abstract

In developing countries such as Brazil, the measurement of services for the CPI may present extra difficulties such as the presence of services in the basket which parts belong to informal economy. In such instances, data sources are even more scarce and difficult to obtain, and the desired information usually rely on some sort of household survey. In this work we present the methodology developed to obtain the inflation for services of home maintenance and housekeeper in the Brazilian CPI using the Continuous Brazilian National Household Sample Survey as the data source. Since this survey was not originally designed to give estimates for the CPI, the methodology developed required a combination of sampling reweighting and structural time series modelling techniques to derive robust CPI measures. Sampling reweight is used to generate adequate weights for the desired domains for the CPI and the adoption of structural time series modelling is employed to deal with the small samples sizes obtained. Finally, we present the price microdata generated for home maintenance and housekeeper services which are available for researchers interested to work with such data sets.

1. Introduction

The Brazilian System of consumer prices indices (SNIPC) maintained by the Brazilian Institute of Geography and Statistics (IBGE) is responsible to provide the official measures of consumer price indicators for Brazil [IBGE, 2013]. Among the indices contained in the system is the extended consumer price index (IPCA) which is used as a macroeconomic indicator by the Brazilian Central Bank in order to define the national monetary policy.

The IPCA follows the same methodological pillars as other CPIs all over the world which aim to approach the cost of living of a reference population based on a fixed basket of goods and services and on the matched model approach [IBGE, 2013, ILO, 2004]. Though the use of a fixed basket is a simplification adopted in order to allow the compilation of consumer inflation, this approach is not free of important challenges. The measurement of services components of the CPI basket is one of such case examples.

The determination of prices of a given service is a hard task since usually require the specification of a large amount of characteristics. A poor description can lead to a sample with high price dispersion. In developing countries such as Brazil, additional difficulties present in some kinds

---

[*] The views expressed in this paper are those of the authors and not necessarily reflect views of IBGE.
[†] Corresponding author: vladimir.miranda@ibge.gov.br

of services which are mainly based on Labour's informal market. In such instances, there is no source of price information available but the individuals that offered the services (or the consumers who hired them). In such circumstances household surveys are necessary in order to derive such informations.

In the IPCA's basket two important *subitems*[1] are representative of the above picture, housekeeper (ED) and services for home maintenance (MORD), amounting to approximately, 4% and 1%, respectively, of the expenditures of the whole basket [IBGE, 2019]. Though in recent years there have been a pressure to increase the formalization of such activities due the promotion of labour's laws that established that such workers should also share the same rights as other categories, the degree of informality in such sectors is still high, specially in periods of economical recessions.

In Brazil, IBGE's Monthly Employment Survey (PME) [IBGE, 2007] used to be the survey responsible to collect and provide information on the brazilian's labour force. The PME was based on a household sample survey conducted in six brazilian's metropolitan regions and was the source used to generate the inflation for the ED and MORD subitems of the IPCA's basket. However, this survey ceased in 2016 and the informations provided by the PME should be replaced by the continuous household sample survey (PNADC). The PNADC is a survey whose sample design comprises the whole national territory and provides estimates for a much finer and richer geographical level than the PME. Despite the different geographical coverage between the PME and the PNADC, another important difference is that the PNADC sample design was planned to provide estimates in a quarterly basis against the monthly periodicity of the PME.

Due such changes, the incorporation of the PNADC as the source responsible to provide the information for the calculation of the ED and MORD subitems' inflation is not straightforward. In this work the methodological treatment developed to incorporate the PNADC data for the calculation of the inflation of ED and MORD is described. The paper is structured as follows: in Section 2 a brief discussion on the elementary formulas used in the IPCA is presented in order to derive the parameters of interest of our problem. In Section 3 some important details of the PNADC sample design are described and how this is determinant for the choice of a estimator for the monthly price relatives. The considerations for the chosen estimator are also presented. Since the sample design of the PNADC was not planned to give estimates for the CPI subitems ED and MORD, it is necessary to perform a reweighting treatment in the PNADC original weights in order to correctly derive the weights for our domain of interest. These steps are presented in Section 4 which also describes the outlier detection and treatment techniques adopted in the data set. In Section 5 the time series approach necessary to deal with the volatile series resulting from the small data sets resulting from the PNADC estimates is described. In Section 6 the final results of the method are presented and also the price microdata available for researchers. Finally in Section 7 the conclusions of the work are disclosed.

2. Statement of the problem: Population parameters of interest.

To determine the parameter of interest for the estimation of the inflation of a given subitem of the IPCA, it is necessary to take into account the elementary level formula used to calculate the inflation of the subitem (see Ref. [Miranda et al., 2019] for more details on the calculation).

---

[1] In the classification system adopted by the SNIPC the lowest level where weights are available from the household budget survey is called "subitem" [IBGE, 2014b].

In the IPCA, the prices variation of a given subitem k between subsequent months t and $t-1$ are obtained by the use of a two-stage-process elementary level formula given by [IBGE, 2013]:

$$R_{t/(t-1),k} = \prod_{t=1}^{n'_k} (P_{t,k,k'}/P_{t-1,k,k'})^{1/n'_k}. \tag{1}$$

where in Eq. (1) $P_{t,k,k'}$ is the arithmetic mean of the prices of products within a given "elementary products aggregate" k' which belongs to the subitem k. $n_{k'}$ is the number of elementary products aggregates within the subitem k.

Hence, as stated by Eq. (1), the price relative $R_{t/(t-1),k}$ between subsequent months t and $t-1$ of the subitem k is given by a "two-stage" averaging process: In the first step an arithmetic mean of similar products within an elementary aggregate is taken while in the second a geometric mean of the ratios of mean elementary aggregates prices is performed.

Since the subitems of interest of this work (housekeeper and home maintanance and repair) are considered to be constituted of a single elementary products aggregate, the subindex k' referring to the elementary products aggregates can be dropped and the price relatives given by Eq. (1) can be simplified to a Dutot-like elementary formula [ILO, 2004]

$$R_{t/(t-1),k} = P_{t,k}/P_{t-1,k} = \frac{\left(\sum_{l \in U_{t,k}} p_{t,k,l}\right)/N_{t,k}}{\left(\sum_{l \in U_{t-1,k}} p_{t-1,k,l}\right)/N_{t-1,k}}, \tag{2}$$

where in Eq. (2) $p_{t,k,l}$ are the prices of a service of a given subitem k offered by an individual l and $N_{t,k}$ is the number of individuals offering services for the subitem k at time t in the population of individuals $U_{t,k}$.

The quantity expressed by Eq. (2) is the parameter of interest for the determination of the inflation of the subitems ED and MORD. In the next section the challenges to construct an estimator for such parameter, taking into account the peculiarities of the PNADC complex sample design, are presented.

3. Considerations for the choice of the estimator

As mentioned above, the PME [IBGE, 2007] was the data source previously adopted to provide estimates for the inflation of the subitems ED and MORD. With the extinction of such survey, this data source was supposed to be replaced by the PNADC [IBGE, 2017]. In order to understand the choice of the estimators proposed below, it is important to keep in mind the main methodological characteristics of the PNADC, which we briefly discuss in the following paragraphs[2].

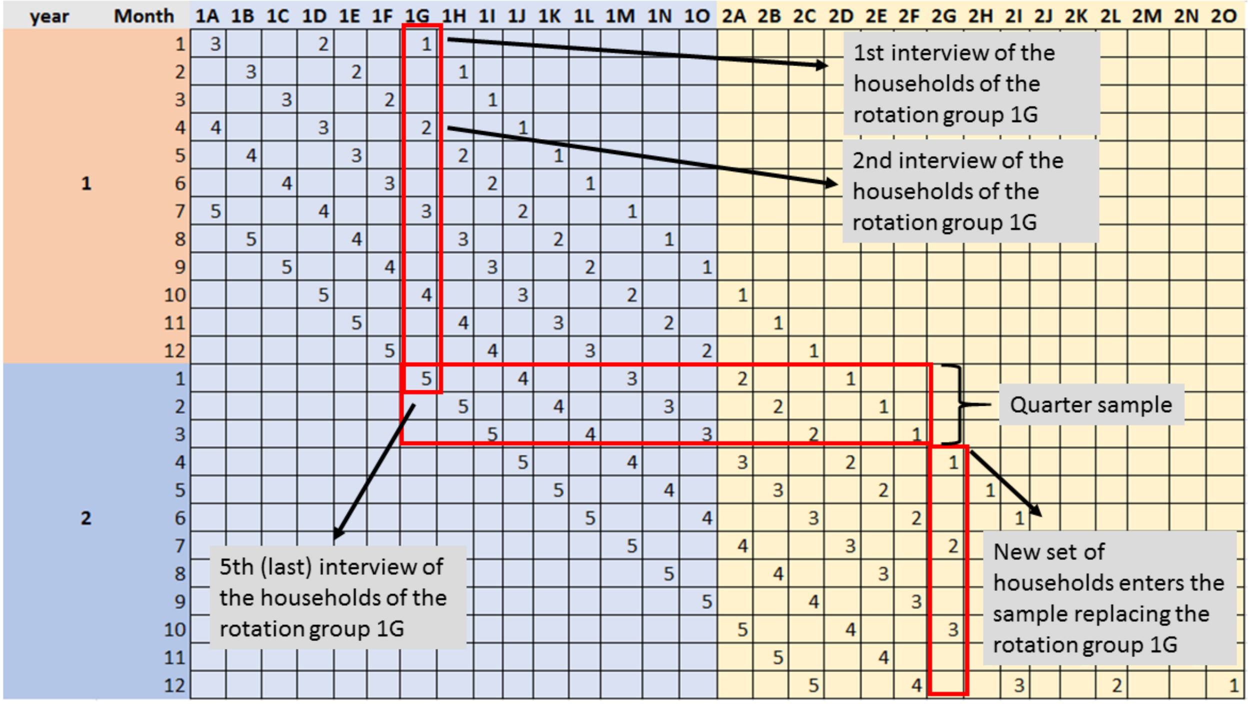
The PNADC is a survey based on a probabilistic sample of households [Freitas and Antonaci, 2014] whose goal is to continuously produce informations on the population labour's force, associated with its demographic and educational characteristics and to provide support for the study of social-economic development of the country. The PNADC has a very robust sample which allows

---

[2] For further details please refer to [IBGE, 2014a, 2017, Freitas and Antonaci, 2014].

the construction of informations with high accuracy for a fine level of geographical coverage such as the capitals of brazilian municipalities.

The PNADC sample was designed in order to produce quarterly informations. This fact implies that it takes three months in order to cover the data collection of the whole country sample. In order to guarantee that the observed estimates variations between following quarters are not highly volatile, a partial sample rotation scheme among consecutive quarters is adopted. In such scheme, a set of selected households is interviewed five times at most. Also, there is a quarter "break" between consecutive interviews of a given household, as described in Figure 1. In the rotation scheme adopted, the primary sampling units (PSUs) of each stratum are split into 15 rotation groups (denoted by the letters A, B, . . . , O in Figure 1). In each quarter a new sample of households is selected within the SPUs of three such groups. For the remaining 12 groups, a new interview is performed. Such rotation process provides a sample superposition of 80% between consecutive quarters and 20% between the same quarter in consecutive years.



year	Month	1A	1B	1C	1D	1E	1F	1G	1H	1I	1J	1K	1L	1M	1N	1O	2A	2B	2C	2D	2E	2F	2G	2H	2I	2J	2K	2L	2M	2N	2O
1	1	3			2			1																							
	2		3			2			1																						
	3			3			2			1																					
	4	4			3			2			1																				
	5		4			3			2			1																			
	6			4			3			2			1																		
	7	5			4			3			2			1																	
	8		5			4			3			2			1																
	9			5			4			3			2			1															
	10				5			4			3			2			1														
	11					5			4			3			2			1													
	12						5			4			3			2			1												
2	1							5			4			3			2			1											
	2								5			4			3			2			1										
	3									5			4			3			2			1									
	4										5			4			3			2			1								
	5											5			4			3			2			1							
	6												5			4			3			2			1						
	7													5			4			3			2								
	8														5			4			3										
	9															5			4			3									
	10																5			4			3								
	11																	5			4										
	12																		5			4			3			2			1

1st interview of the households of the rotation group 1G

2nd interview of the households of the rotation group 1G

Quarter sample

5th (last) interview of the households of the rotation group 1G

New set of households enters the sample replacing the rotation group 1G

Figure 1: 1-2(5) Rotation scheme of the PNADC sample. The top rectangle illustrates that a set of households from the rotation group 1*G* are interviewed for the first time in month 1 of the first year. After the first interview in month 1 the second interview only occurs in the next quarter in month 4. The same same pattern proceeds for the remaining interviews until a household is interviewed for the fifth and last time in 12 months after the first interview. After the fifth interview of the group of households of the rotation group 1*G* in month 1 of year 2 this group of households leave the sample and is replaced by a new set of households from the rotation group 2*G* that will perform their first interview in month 4 of year 2 (bottom rectangle), a quarter after the last interview of the households of group 1*G*. The rectangle in the middle of the Figure illustrates that in a given quarter the sample is balanced such that there are always individuals answering for the first, second, third, fourth and fifth time the survey. This guarantees that the sample is refreshed parsimoniously with a 80% superposition between subsequent quarters and 20% in a year window.

In order to develop estimators for the mean prices and price relatives given in Eq. (2) one should note that, within the PNADC nomenclature, the combinations of products x local of interest for the price index are the workers employed and belonging to the categories of housekeepes or home and maintenance repairs of a given reference month and region of interest.

Monthly estimates for the price's population averages $P_{t,k}$ can be obtained from a subsample from the PNADC where the domains of interest are the group of workers which represent the subitems ED and MORD at each month. The sample estimators for such population means are given by:

$$\hat{P}_{t,k} = \left(\sum_{l \in a_{t,k}} w_{t,k,l} \times p_{t,k,l} \right) / \sum_{l \in a_{t,k}} w_{t,k,l} \tag{3}$$

where $a_{t,k}$ is the sample of people surveyed in month t belonging to the subitem k (ED or MORD), $p_{t,k,l}$ is the price charged by the worker l from the subitem k in month t, and $w_{t,k,l}$ is the corresponding sampling weight to select the worker l from the subitem k in month t.

The estimator given by Eq. (3) has appealing characteristics since its use does not require the "creation" of special weights or any intervention in the production routines of the estimates from the PNADC sample. Besides, it allows the use of the complete PNADC monthly samples in order to produce estimates for the domains defined by the workers belonging to the ED and MORD groups.

By taking the ratio of mean prices for subsequent months using the estimator given by Eq. (3), a prices relative estimator can be derived as:

$$\hat{R}_{t/(t-1),k} = \hat{P}_{t,k} / \hat{P}_{t-1,k}. \tag{4}$$

Though the use of the estimator given by Eq. (4) is appealing, a careful analysis of the PNADC sample characteristics reveals that such choice is not advisable. Due the sampling rotation process of the PNADC discussed above, the samples collected between subsequent months have no superposition. As a consequence, the approximate variance of the estimator given by Eq. (4), writes

$$V(\hat{R}_{t/(t-1),k}) = V(\hat{P}_{t,k}) + V(\hat{P}_{t-1,k}) - 2Cov(\hat{P}_{t,k}, \hat{P}_{t-1,k}). \tag{5}$$

The lack of superposition between individuals in the samples of subsequent months implies that the variance given by Eq. (5) is "maximal" as a consequence of the null covariance in Eq. (5). Another important point is that the subsamples collected each month amount to approximately 1/3 of the whole quarter's sample of the PNADC.

The combination of the reduced size of the month samples and lack of superposition among samples of subsequent months implies that the relatives estimates given by Eq. (4) have a high variance producing a highly volatile series of estimates.

Preliminary analysis (not shown) confirmed that estimates calculated from Eq. (4) produce a volatile series. As an outcome of this analysis, it was decided that the adoption of the estimator given by Eq. (4) was not appropriate.

An alternative estimator can be developed from the inspection of Eq. (5). Such expression suggests that estimators with superposed samples can be significantly benefited from a variance reduction, and such reduction is as large as the samples degree of superposition.

A natural choice of estimators for the parameters given by Eq. (2) that takes into account the longitudinal aspect of the PNADC and the rotation scheme is one based on the ratio of mean prices of months separated by a quarter.

$$\tilde{R}_{t/(t-1),k} = (\tilde{P}_{t,k}/\tilde{P}_{t-3,k})^{1/3}. \tag{6}$$

Since the ratio of mean prices between months t and $t-3$ gives a quarterly variation, the cubic root in Eq. (6) is necessary for the extraction of monthly variations. One should note that the mean prices in Eq. (6) are obtained taking into account the matching of individuals in the samples in months t and $t-3$ and are given by

$$\tilde{P}_{t,k} = \left(\sum_{l \in e_{t,k}} w_{t,k,l} \times p_{t,k,l} \right) / \sum_{l \in e_{t,k}} w_{t,k,l}, \tag{7}$$

where in Eq. (7) $e_{t,k}$ is the sample of individuals of the subitem k in the month t which have a match in month $t-3$. Due the rotation process, at each month at most 4/5 of the monthly sample would be available for matching and for the calculation of the mean prices and prices relatives estimates given by Eq. (7) and Eq. (6), respectively.

Despite the reduced size of the available monthly sample left for the estimation of quarterly prices relatives given by Eq. (6), the reduction in the variance due sample superposition should compensate the lost in sample size and still leads to estimates with a reduced variance respective the estimator give by Eq. (4).

Another important point to note is that the estimator proposed by Eq. (6) is a consistent estimator for the population quantity

$$(P_{t,k}/P_{t-3,k})^{1/3} \tag{8}$$

rather than the population target parameter given by Eq. (2). Such fact amounts in the estimator given by Eq. (6) being a biased estimator for the parameter Eq. (2), where the asymptotic bias is given by:

$$Vies(\tilde{R}_{t/(t-1),k}) \cong (P_{t,k}/P_{t-1,k}) - (P_{t,k}/P_{t-3,k})^{1/3} \tag{9}$$

Table 1: Summary statistics of the bias and absolute bias obtained by comparing the true IPCA series and the estimates obtained using the estimator Eq. (6).

Summary Statistics	Bias	Absolute Bias
Min	-0.4	0.0
Q1	-0.1	0.04
Median	-0.01	0.1
Mean	0.0	0.12
Q3	0.1	0.18
Max	0.44	0.44

Although there is a bias "gain" in adopting the estimator given by Eq. (6), the balance of the bias-variance trade-off, expressed by measuring the mean square error, suggests that the use of the estimator given by Eq. (6) is preferred.

To illustrate the estimation bias issue, data from the official series of the IPCA[3] is used in order to evaluate the bias incurred by the use of the estimator in Eq. (6), in comparison with the "monthly-direct" estimator of prices. Table 1 displays the summary statistics of the bias distribution and of the absolute bias calculated through the use of Eq. (9) on the IPCA data, considering the time window between Jan-2017 through Apr-2017.

One can note from the analysis of Table 1 that the distribution of the absolute bias stand around $0,1\%$ ($\approx 0,18\%$) for about 50% (75%) of the series values. The summary statistics of the bias also reveal important features such as a null mean of the bias of the series. This indicates that the bias is non-cumulative throughout the time and cancel itself as time evolves, a highly desirable aspect for the inflation measure.

Figure 2 shows the evolution of the bias estimates for the whole time window considered. One can note the bias fluctuation and alternating between positive and negative values in agreement with the analysis of the summary values of Table 1.

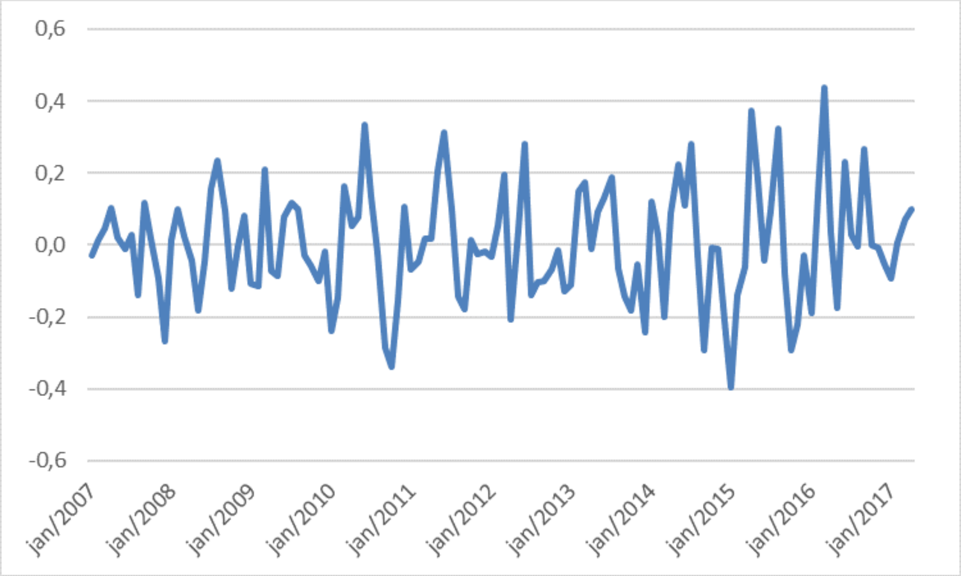


Figure 2: Series of the estimated bias incurred in using the estimator given by Eq. (6) for the calculation of the IPCA inflation in the ten years time window interval comprising Jan-2007 through Jan-2017.

Based of the previous discussion and data presented above, where it was found that the bias incurred in using the estimator given by Eq. (6) is small, this estimator was selected to calculate the inflation of the subitems ED and MORD.

---

[3] The official series are available on-line in the IBGE official website and can be obtained in the link.

4. Selection and treatment of the PNADC data in order to use the adopted estimator

4.1. Reweighting

As mentioned above, the available data for a month t used for the calculation of the estimator given by Eq. (6) constitute a subsample of the whole quarter's sample of the PNADC. Further reduction in the sample occurs due non-responses and the matching procedure between months t and $t-3$. In order to derive estimates for our domains of interest it is necessary to take into account the sample design of the PNADC and such sample reductions. Since only a subsample of the PNADC is used for the derivation of estimates of the inflation of the subitems ED and MORD, a reweighting process is necessary in order to derive the correct inflation measures.

In the following it is presented the reweighting steps and adjustments in the PNADC sample weights for the calculation of the ED and MORD inflation estimates. In order to simplify the notation, all the expressions presented below refer to a single generic sample stratum used for the selection of a PNADC PSU.

4.1.1. Calculation of the quarterly weights

The PSU's standard weight to estimate characteristics produced at a given quarter of the PNADC is given by

$$w_{gi} = \frac{1}{m_g} \times \frac{N_g}{N_{gi}}, \tag{10}$$

where in Eq. (10) w_{gi} is the weight of the PSU i of the rotation group g in a PNADC quarter containing the months t, $t-1$ and $t-2$;

m_g is the number of PSUs selected from the rotation group g;

N_{gi} is the total number of permanent private households (occupied, unoccupied or closed) in the PSU i from the rotation group g, according the informations from the national households frame *Cadastro Nacional de Endereços para Fins Estatísticos* - CNEFE in the moment of the PSUs selection;

N_g is the total number of permanent private households (occupied, unoccupied or closed) in the rotation group g, according the informations from the national households frame *Cadastro Nacional de Endereços para Fins Estatísticos* - CNEFE in the moment of the PSUs selection.

The standard weight of each household is given by the reciprocal of the probability of selection of a household and can be decomposed into a portion relating to the PSU's selection and another one corresponding to the selection of the households within the PSU. Such decomposition writes

$$w_{gij} = w_{gi} \times w_{j|gi} = \frac{1}{m_g} \times \frac{N_g}{N_{gi}} \times \frac{N_{gi}^*}{n_{gi}} \tag{11}$$

$w_{j|gi}$ is the standart weight of the household j in the PSU i and rotation group g in a given quarter of the PNADC;

N^*_{gi} is the total number of permanent private occupied or closed households in the PSU i and rotation group g, according the informations from the national households frame *Cadastro Nacional de Endereços para Fins Estatísticos* - CNEFE in the moment of the PSUs selection.

n_{gi} is the number of households selected from the PSU i and rotation group g.

4.1.2. Calculation of the monthly weights

In order to use data from a single month t of the PNADC, the weights correction is given by [IBGE, 2015]:

$$w^t_{gij} = w_{gij} \times \frac{\sum_{g,i,j} w_{gij}}{\sum_{g,i,j,} w_{gij} \times I(j \in a_t)}. \tag{12}$$

where in Eq. (12) w^t_{gij} is the standard weight of the household j, within the PSU i, rotation group g and month t. Note that the superscript t was introduced to denote the reference month.

The numerator in the right hand side of the fraction in Eq. (12) denotes the total standard weight of the households of a PNADC's quarter and the denominator is the total standard weight of households surveyed in a month t.

4.1.3. Calculation of the monthly weights considering non-response adjustment

The adjustment in the weights due non-response losses writes:

$$w^{t_a}_{gij} = w^t_{gij} \times \frac{n^*_{gi}}{n^{**}_{gi}}, \tag{13}$$

where in Eq. (13) $w^{t_a}_{gij}$ is the standard weight adjusted by non-responses of the household j, from the PSU i and rotation group g in the month t. The supersxript t_a is used to denote month's t relative weight adjusted for non-responses.

n^*_{gi} is the number of households with dweller selected from the PSU i in the rotation group g;

n^{**}_{gi} is the number of households selected and successfully interviewed within the PSU i and rotation group g.

For the strata where there are no matched households, this stage is performed in the post stratification of all the strata contained in the post stratum.

4.1.4. Calculation of monthly weights for matched households

In order to deal with the sample modification due the matching, the weights adjustment writes:

$$w^{t_p}_{gij} = w^{t_a}_{gij} \times \frac{\sum_{g,i,j} w^{t_a}_{gij}}{\sum_{g,i,j} w^{t_a}_{gij} \times I(j \in e_t)} \tag{14}$$

where in Eq. (14), $w_{gij}^{t_p}$ is the matched-adjusted weight of the household j, from the PSU i, rotation group g, within the month t for the households with at least one of its dwellers matched between months t and $t-3$. The superscript t_p is used to denote the quantities associated with the matched households in month t.

The numerator in Eq. (14) fraction is the total non-response-adjusted weight in the month t and the denominator is the total non-response-matched-adjusted weight which takes into account households with at least one dweller matched between months t an d $t-3$.

4.1.5 Post stratification of the monthly weights for the matched households

As in the PNADC the weights are adjusted by post stratification considering the estimated population in the strata by the 15th of each reference month [IBGE, 2015], such adjustment is also maintained in the present context. In this case, the adjustment is performed in each stratum contained in the post stratum b. After this adjustment, the final weight for an individual l, belonging to the household j, in the PSU i, rotation group g with a match between months t and $t-3$ is given by

$$w_l^{t_{SNIPC}} = w_{gij}^{t_p} \times \frac{P_b}{\hat{P}_b}. \tag{15}$$

P_b is the population estimate produced by the IBGE for the post stratum b for the day 15 of the reference month and $\hat{P}_b$ is the population estimate obtained for the households within the post stratum b with at least one dweller matched between months t and $t-3$ in the reference month.

4.2. Outliers detection and treatment

Since the estimator chosen is based in prices means, the estimates obtained for the income information from the PNADC are subject to distortions due to outliers. Due this fact it is necessary to develop an outliers' detection system and establish some criteria for the treatment of the detected observations.

The first step is the definition of the criteria used to classify an observation as an outlier. For the income distributions considered here, the outliers are determined by the boxplot approach which states that values below (above) an lower (upper) bound threshold LI (LS) are classified as outliers of the distribution. LI and LS are defined as

$$LI = Q1 - 2(Q3 - Q1) \,\text{and}\, LS = Q3 + 2(Q3 - Q1), \tag{16}$$

where in Eq. (16), $Q1$ ($Q3$) refer to the 1st (3rd) quartile of the income distribution.

In each moment t, two prices distributions are considered separately, the prices distributions obtained at the months t and $t-3$ after the matching process.

For the calculation of the quartiles and thresholds LI and LS in Eq. (16), it is necessary to take into account the weight of the matched individuals as defined by Eq. (15) in order to obtain the appropriate prices distribution. Another important point is that due the rotation process and the non-response adjustments, the weight of an individual l matched between months t and $t-3$ may be different in these two months. Hence, in order to guarantee that the calculated inflation is solely due to a pure prices variation the weight calculated in the most recent month t for an individual l

matched between months t and $t-3$ is adopted in order to derive the prices estimates for both matched months.

The second step is the treatment of the detected outliers. In this work, the treatment adopted is the use of the Winsorization technique of the detected outliers. Such treatment states that the detected outliers are replaced/imputed by the value of LI (LS) if it lies below (above) LI (LS).

The boxplot detection system assumes that the distributions under consideration follow approximately a normal distribution. Since prices distributions are in general asymmetrical, a Box-Cox transformation [Box and Cox, 1964] is applied to the income distributions in order to approximate them to a normal distribution and allow the implementation of the boxplot detection process.

Mathematically, the outliers treatment rules are expressed as

$$p'_{t,k,l} = \begin{cases} LI & \text{if } p'_{t,k,l} < LI \\ p'_{t,k,l} & \text{if } LI \leq p'_{t,k,l} \leq LS \\ LS & \text{if } p'_{t,k,l} > LS, \end{cases} \tag{17}$$

where in Eq. (17) $p'_{t,k,l}$ are the prices obtained after the Box-Cox transformation and are given by [Box and Cox, 1964]

$$\mathcal{F}_{BC}(p_{t,k,l}) = p'_{t,k,l} = \begin{cases} \frac{p_{t,k,l}^{\lambda_{BC}}-1}{\lambda_{BC}} & \text{if } \lambda_{BC} \neq 0 \\ \ln(p_{t,k,l}) & \text{if } \lambda_{BC} = 0. \end{cases} \tag{18}$$

$p_{t,k,l}$ amounts to the price of the service offered by the individual l from subitem k in the month t. The parameter λ_{BC} defines the Box-Cox transformation ($\mathcal{F}_{BC}$) to be adopted in the data set and is estimated for each of the prices samples analysed.

Note that for the calculation of the prices relatives estimates from the estimator given by Eq. (6) it is necessary to apply such treatment for two prices distributions apart, the income distribution for the matched individuals found in month t and those at month $t-3$. Due the sample rotation, individuals matched in a given month t are different when the match is between $t-3$ and t from t and $t+3$. Hence, the detection process for the individuals of a given month t needs to be conducted twice, one for each distribution that results from the matching step.

Based on the final weight of an individual $w_l^{t_{SNIPC}}$ given by Eq. (15) and derived in the previous section, and after the definition of the outlier detection and treatment processes, the mean prices and prices relatives estimates for the subitems ED and MORD are given by

$$\tilde{R}_{t/(t-1),k} = (\tilde{P}_{t,k}/\tilde{P}_{t-3,k})^{1/3}, \tag{19}$$

with

$$\tilde{P}_{t,k} = \left(\sum_{l \in e_{t,k}} w_{t,k,l} \times p_{t,k,l} \right) / \sum_{l \in e_{t,k}} w_{t,k,l} = \left(\sum_{l \in e_{t,k}} w_l^{t_{SNIPC}} \times \mathcal{F}_{BC}^{-1}(p'_{t,k,l}) \right) / \sum_{l \in e_{t,k}} w_l^{t_{SNIPC}}, \tag{20}$$

$$\tilde{P}_{t-3,k} = \left(\sum_{l \in e_{t,k}} w_l^{t_{SNIPC}} \times \mathcal{F}_{BC}^{-1}(p'_{t-3,k,l}) \right) / \sum_{l \in e_{t,k}} w_l^{t_{SNIPC}} \tag{21}$$

where in Eqs. (20) and (21) the sets $e_{t,k}$ indicate that the sums are restricted to those individuals l belonging to the subitens k of interest (ED or MORD) and matched between months t and $t-3$. $\mathcal{F}_{BC}^{-1}$ denotes the inverse of the Box-Cox transform given by Eq. (18). Note again that the weights adopted for the matched individuals that appear in Eqs. (20) and (21) are the ones given at month t for both in the calculation of $\tilde{P}_{t,k}$ and $\tilde{P}_{t-3,k}$, as discussed previously.

In Figures (3)-(4) the series of prices relatives estimates $\tilde{R}_{t/(t-1),k}$ obtained for the subitems ED and MORD using the PNADC data and the methodology described above is displayed. One can note that such results present fluctuations that are common to surveys with small sample sizes (note for instance the abrupt change in the values obtained for the MORD subitem in the mids 2014 in Figure (4)). Thus, the use of such estimates is still not satisfactory for the measurement of inflation of the subitems ED and MORD. To deal with this volatility problem a time series modelling treatment is adopted. It is decided to use the estimates of the trend component of such series of estimates $\tilde{R}_{t/(t-1),k}$.

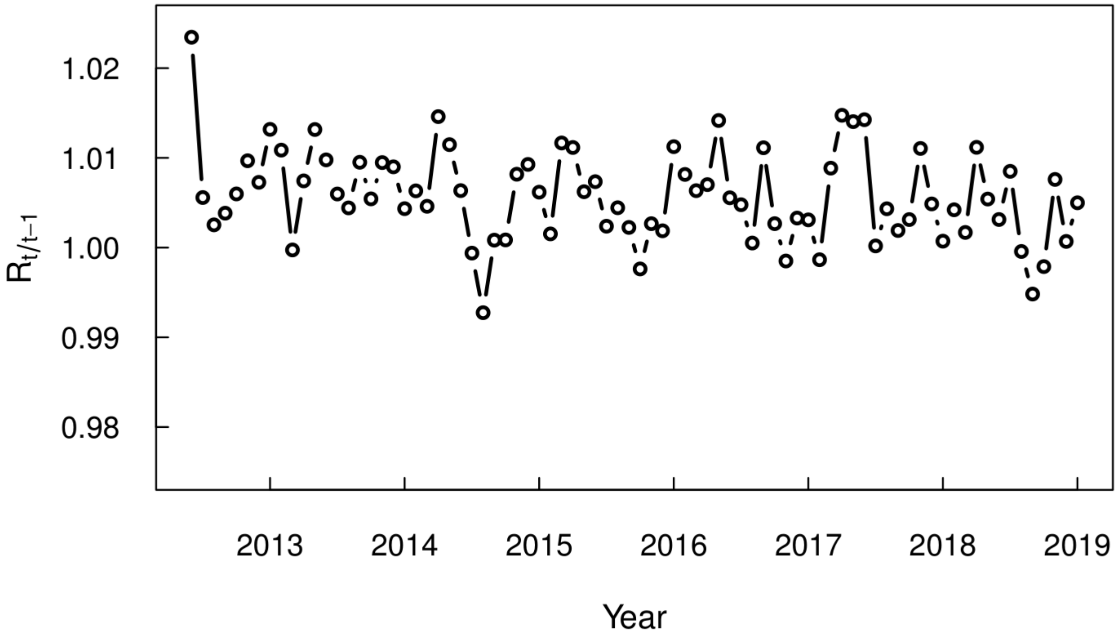


Figure 3: PNADC series of estimates for the subitem ED.

5. Time series modelling for estimating the trend of the prices relatives series of the ED and MORD subitems

In this section the time series approach adopted in order to derive the estimates of the trend component $\tilde{R}_{t/(t-1),k}$ of the series of prices relatives of the subitems ED and MORD is described. The approach adopted consists in the use of structural times series modelling which allow to

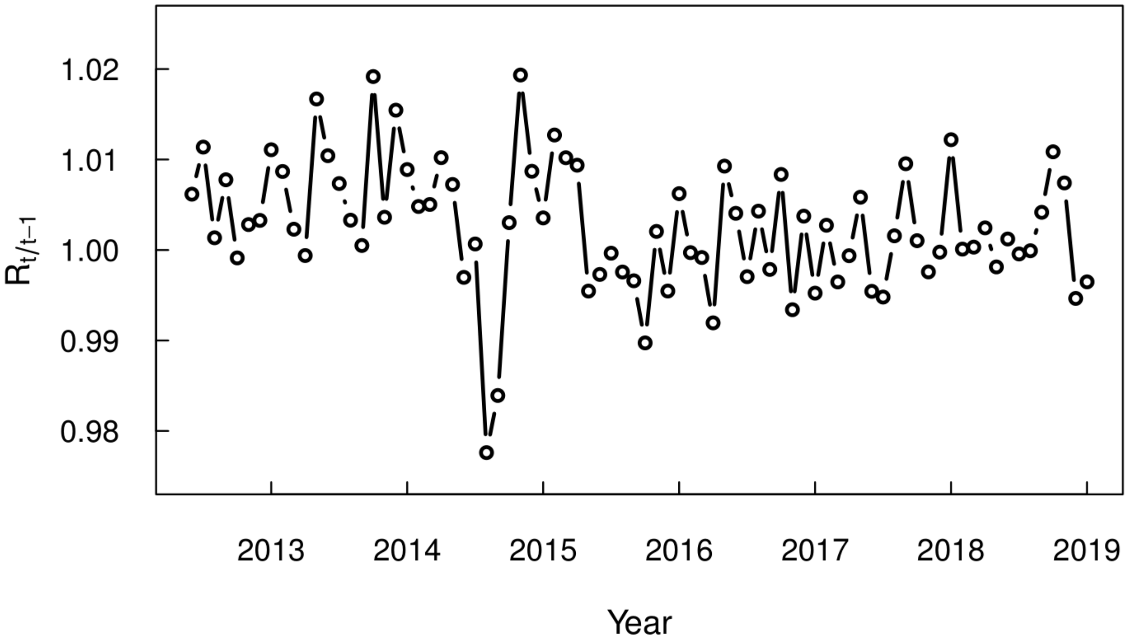


Figure 4: PNADC series of estimates for the subitem MORD.

express the time series via non-observable components of trend, seasonal and irregular which evolve stochastically in time.

5.1. Structural times series models

The structural time series models [Harvey, 1989] allow the decomposition of a time series in unobservable components which however have a clear interpretation such as trend, seasonal, cicle and irregular. The trend component captures the direction and level of the series. The seasonal represents the periodical movement of the series (intra-annual in the present context). The cicle is associated with recurrent movements of long range and the irregular component reflects non-systematic movements and residual fluctuations not explained by the other components of the model. The additive decomposition approach, used to represent a series with simple variations, as in the present context, writes

$$ST_t = Trend_t + Seasonal_t + Cicle_t + Irregular_t \tag{22}$$

where ST_t is an observation of the time series in the moment t and the remaining components are as described above.

In the structural model the components are usually described stochastically and hence each component is influenced by a random perturbation term (random error). The focus of the modelling process is the estimation of the components of interest, which in the present case is the trend component. In this way, excluding the cicle term, the desired component is expressed as

$$Trend_t = ST_t - Seasonal_t - Irregular_t \tag{23}$$

To estimate the series components, the structural models are expressed in the state space formulation [Harvey, 1989] and the Kalman filter is used [Kalman, 1960] which is an approach that establishes a recursive procedure allowing the construction of inferences about the unobserved components of the time series.

In this way, the main feature of the state space approach lies on the ability to provide updated estimates of the unobserved components and to make predictions on future observations.

5.2. State space representation of structural time series models

In the present Section, the notation $\tilde{R}_t$ is adopted such as to denote a generic time series or to make connection with the time series of the price relatives estimates for a given month t, for the subitem k, $\tilde{R}_{t/(t-1),k}$.

Consider a time series $\tilde{R}_t$, with $t = 1, ..., T$. Denoting by $\boldsymbol{\alpha}_t$ the vector[4] of unobserved components $\tilde{R}_t$, denominated as the vector of states. The modelling focus is the implementation of inferences over the vector of states $\boldsymbol{\alpha}_t$ based on the information contained in the observed values $\tilde{R}_t$. The current state of the system can be defined as the minimal set of information that, together with future observations, is able to describe the future behavior of the system.

The state space formulation is defined by the two following equations:

$$\tilde{R}_t = \mathbf{Z}_t \boldsymbol{\alpha}_t + \epsilon_t \quad \text{with } \epsilon_t \sim N(0, \mathbf{H}_t) \tag{24}$$

$$\boldsymbol{\alpha}_t = \mathbf{T_t} \boldsymbol{\alpha}_{t-1} + \mathbf{U}_t \boldsymbol{\xi}_t \quad \text{with } \boldsymbol{\xi}_t \sim N(0, \mathbf{Q}_t) \tag{25}$$

Equation (24) is denominated observation (or measurement) equation and represents the relation between the observations and the actual state of the non-observable components. Equation (25) is denoted the transition (or state) equation and describes the stochastic time evolution of the non-observable components allowing the incorporation of variables that allow the specification of dynamic processes of the components [Durbin and Koopman, 2001].

The matrices that determine the system equations $\mathbf{Z}_t$, $\mathbf{T}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$ and $\mathbf{U}_t$ generally rely on a set of unknown parameters that need to be estimated and whose format is detailed in the Section 5.2.2 below, after the description of the basic structural model adopted.

The stages of the prediction process of $\tilde{R}_t$ and estimation of the unobservable components $\boldsymbol{\alpha}_t$ are performed recursively based on the Kalman filter [Harvey, 1989] algorithm described below.

5.2.1 The Kalman filter

The Kalman filter (KF) is a recursive algorithm conditioned to a particular set of informations. If such set includes the whole past and excludes current values of the observed data, $\mathbf{D}_{t-1} = (\tilde{R}_1, ..., \tilde{R}_{t-1})$, allows to perform predictions. When new observations of $\tilde{R}_t$ are available, the set of informations also include, besides the whole past, the current values of the observed data $\mathbf{D}_t$, and hence can be expressed as a sign extraction problem.

---

[4] Vectors and matrices are denoted by the use of bold letters.

The algorithm has three stages:

Forecast: $P[\boldsymbol{\alpha}_t|\tilde{R}_{t-1}]$, allows to obtain an estimate for the components of the vector of states without the incorporation of the most recent observations;

Updating prediction: $P[\boldsymbol{\alpha}_t|\tilde{R}_t]$. Also known as sign extraction process and allows to recover, in the moment t, information about a non-observable component of the system by using all the informations available until the moment t;

Smoothing: $P[\boldsymbol{\alpha}_t|\tilde{R}_T]$. Provides an estimate in time t through the use of all the available information, prior and after the moment t (where a series of size T is being considered).

The smoothing step is associated with an important feature of the Kalman filter: the ability to provide a retrospective smoothed estimate for the non-observable components when a new information is available. The smoothing process allows to revise the past vector of states inferences by making use of the informations available until time T[5]. In this stage, the algorithm is run from the most recent to the oldest observation.

It is important to note that the estimate of the trend component for the problem treated in this work is the one obtained by the end of the smoothing process.

The estimation of the state space model hiperparameters based on the Kalman filter equations is performed through the use of the maximum likelihood method [Durbin and Koopman, 2001].

5.2.2 Basic structural model

The basic structural model (BSM) contemplates, in the observation equation, the sum of the level and seasonal components. Its general formulation is the following

$$\tilde{R}_t = \mu_t + \gamma_t + \epsilon_t \qquad \text{with } \epsilon_t \sim N(0, \sigma^2_\epsilon), \tag{26}$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \qquad \text{with } \eta_t \sim N(0, \sigma^2_\eta), \tag{27}$$

$$\beta_t = \beta_{t-1} + \zeta_t \qquad \text{with } \zeta_t \sim N(0, \sigma^2_\zeta), \tag{28}$$

$$\gamma_t = \sum_{f=1}^{s/2} \gamma_{f,t}. \tag{29}$$

μ_t refers to the trend (level) component, β_t is the slope component and γ_t refers to the seasonal. ϵ_t refers to the irregular component of the model while η_t and ζ_t refer, respectively, to the level and slope of the random noises of the equation of states.

Here, the seasonal component is treated by the use of trigonometric functions.

---

[5] This revision process is only adopted to provide the current estimates more precise, however due the non-revision policy of the CPI, the published official inflation of past periods are not revised.

$$\begin{bmatrix} \gamma_{f,t} \\ \gamma^*_{f,t} \end{bmatrix} = \begin{bmatrix} \cos(\lambda_f) & \text{sen}(\lambda_f) \\ -\text{sen}(\lambda_f) & \cos(\lambda_f) \end{bmatrix} \begin{bmatrix} \gamma_{f,t-1} \\ \gamma^*_{f,t-1} \end{bmatrix} + \begin{bmatrix} \Omega_{f,t} \\ \Omega^*_{f,t} \end{bmatrix}, \text{ with } \begin{bmatrix} \Omega_{f,t} \\ \Omega^*_{f,t} \end{bmatrix} \sim N(0, \sigma^2_\Omega \mathbf{I}_{2\times 2}). \tag{30}$$

In Eq. (30), $\lambda_f = 2\pi f/s$ and f lie in the range between 1 through $s/2$ for the $\gamma_{f,t}$ equation and in the range 1 through $s/2-1$ for the $\gamma^*_{f,t}$ equation. For month data, $s = 12$.

For this model, the matrices of the system $\mathbf{Z}_t$, $\mathbf{T}_t$, $\mathbf{H}_t$, $\mathbf{Q}_t$ and $\mathbf{U}_t$ follow a deterministic time-invariant process and are given by

$$\mathbf{Z}_t = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & \dots & 1 \end{bmatrix}_{1\times 13}, \tag{31}$$

$$\mathbf{Q}_t = \begin{bmatrix} \sigma^2_\eta & 0 & 0 & \dots & 0 \\ 0 & \sigma^2_\zeta & 0 & \dots & 0 \\ 0 & 0 & \sigma^2_\Omega \mathbf{I}_{11\times 11} & & \end{bmatrix}_{13\times 13}, \tag{32}$$

$$\mathbf{U}_t = \mathbf{I}_{13\times 13}, \tag{33}$$

$$\mathbf{H}_t = \sigma^2_\epsilon, \tag{34}$$

$$\boldsymbol{\xi}_t = \begin{bmatrix} \eta_t \\ \zeta_t \\ \Omega_{1,t} \\ \Omega^*_{1,t} \\ \vdots \\ \Omega_{5,t} \\ \Omega^*_{5,t} \\ \Omega_{6,t} \end{bmatrix}, \tag{35}$$

$$\mathbf{T}_t = \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \mathbf{C}_1 & & & \vdots \\ \vdots & \vdots & & \dots & & \\ 0 & 0 & \dots & & \mathbf{C}_5 & 0 \\ 0 & 0 & \dots & & & -1 \end{bmatrix}_{13\times 13}, \tag{36}$$

where $\mathbf{I}_{d\times d}$ are $d-$dimensional identity matrices and $\mathbf{C}_1 \dots \mathbf{C}_5$ are 2×2 matrices given by

$$\mathbf{C}_f = \begin{bmatrix} \cos(\lambda_f) & sen(\lambda_f) \\ -sen(\lambda_f) & \cos(\lambda_f) \end{bmatrix}. \tag{37}$$

In this formulation the non-observable components matrix writes

$$\boldsymbol{\alpha}_t = \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_{1,t} \\ \gamma^*_{1,t} \\ \vdots \\ \gamma_{5,t} \\ \gamma^*_{5,t} \\ \gamma_{6,t} \end{bmatrix}, \tag{38}$$

and the hiperparameter vector to be estimated is given by $\boldsymbol{\theta}(\sigma^2_\epsilon, \sigma^2_\eta, \sigma^2_\zeta, \sigma^2_\Omega)$.

By representing the model defined in Eqs. (26)-(30) in the state space formulation, and by utilizing the Kalman filter for the estimation of the unobservable components, a smoothed estimate of the trend component μ_t can be obtained for the series of estimates of prices relatives $\tilde{R}_t$. The trend estimates can then be used for the calculation of the inflation of the ED and MORD subitems.

6. Results and prices microdata

6.1. Discussion of the results

Though the PNADC series started in 2012 and the PME ended in 2016, the adoption of the PNADC data for estimation of the inflation of the subitems ED and MORD only occurred in may 2018. Due the revision policy of the CPI the results presented here are limited to the time window starting in may 2018. However, the prices microdata for the whole PNADC series is available for researchers as presented in Section 6.2.

In Figures (5)-(6) the series of inflation estimates of the subitems ED and MORD are displayed for three different approaches: the one using the "raw" PNADC data (PNADC via the procedures described in Eqs. (17)-(21)), the estimates of the trend component of the time series and estimates derived via the yearly variation of the official brazilian minimum wage[6] (MW). As can be seen, the trend estimates obtained via the time series approach are much more smoother than the results derived by the PNADC via the procedures described in Eqs. (17)-(21), though in this time range the magnitude of the fluctuations of the PNDAC "raw" estimates is smaller than those observed for the whole time window observed in Figures (3) and (4).

In the period shown, the inflation of the subitem ED given by the trend component is always superior to the inflation presented by the minimum wage with a slight decreasing movement after

---

[6] The monthly estimates of the minimum wage are obtained via extracting the 12-th root of the yearly update of the minimum wage (for further details see Ref. IBGE [2016a]).

09/18. One should also note that during the period analysed the trend's monthly prices relatives estimates were always greater than the unity implying in a upward inflationary movement for this subitem in the role time window which is more clearly observed in the index number series plot shown in Figure 7.

For the relatives estimates of the subitem MORD (shown in Figure 6) on the other hand the trend's prices relatives estimates show oscillatory movements. An upward movement of the trend series is observed until 11/18 where the series starts to drop. Another important feature of this subitem is the observation of periods with downwards inflation movements as observed in Figure 7 and which are related with prices relatives lesser than the unity. Also notable is that the trend's prices relatives only exceeds those of the minimum wage estimates during a small window of three months. As observed in Figure 7, the upturn movement is not enough to compensate for the deflationary period of this category and the number index series for this subitem never reaches the minimum wage series in the range analysed.

The results revealed by the trend estimates seem to be in good agreement with the brazilian economical scenario where a high drop in the economy has presented a great impact in the labour market specially for those activities related to the construction sector. The modest recovery of the economy in the last year combined with a longer period of salary drops might explain the upturn movement of the MORD inflation in the period analysed.

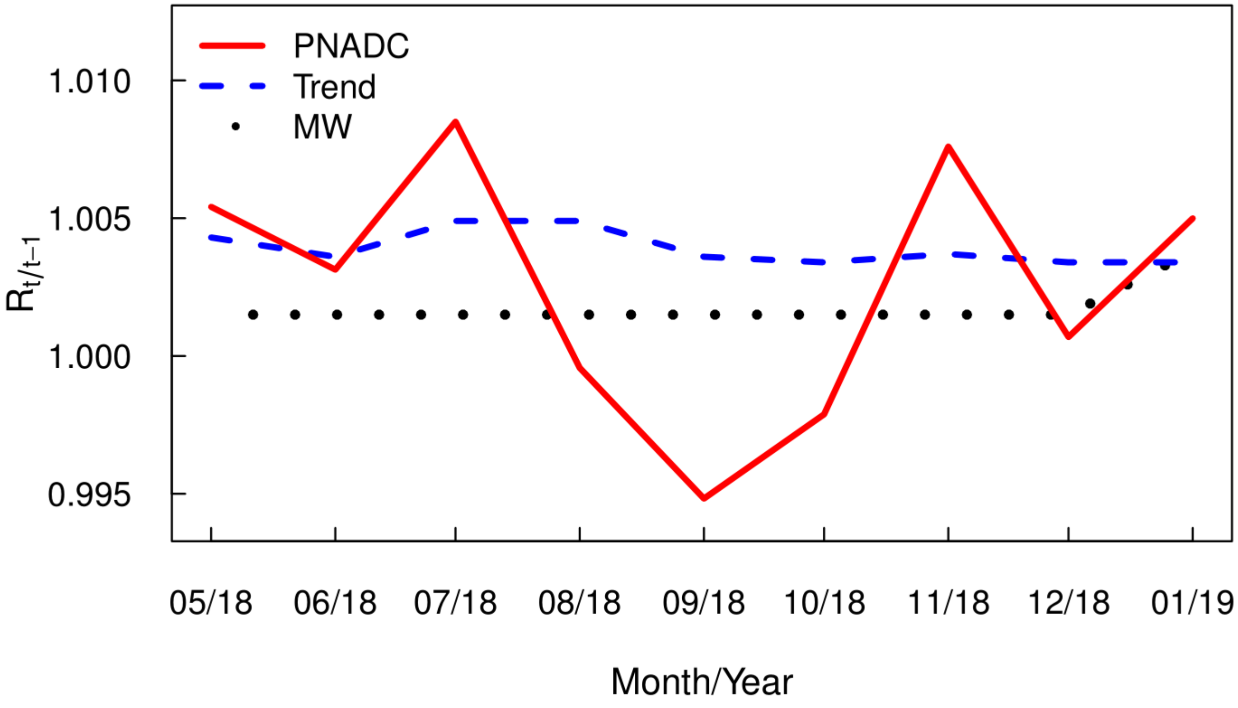


Figure 5: Comparison of estimates for the prices relatives of the subitem ED. The solid, dashed and dotted lines represents, respectively, the PNADC, trend and minimum wage estimates. The PNADC estimate are those derived via the procedures described in Section 4 and the "trend" is the trend component of such series derived by the procedure described in Section 5.

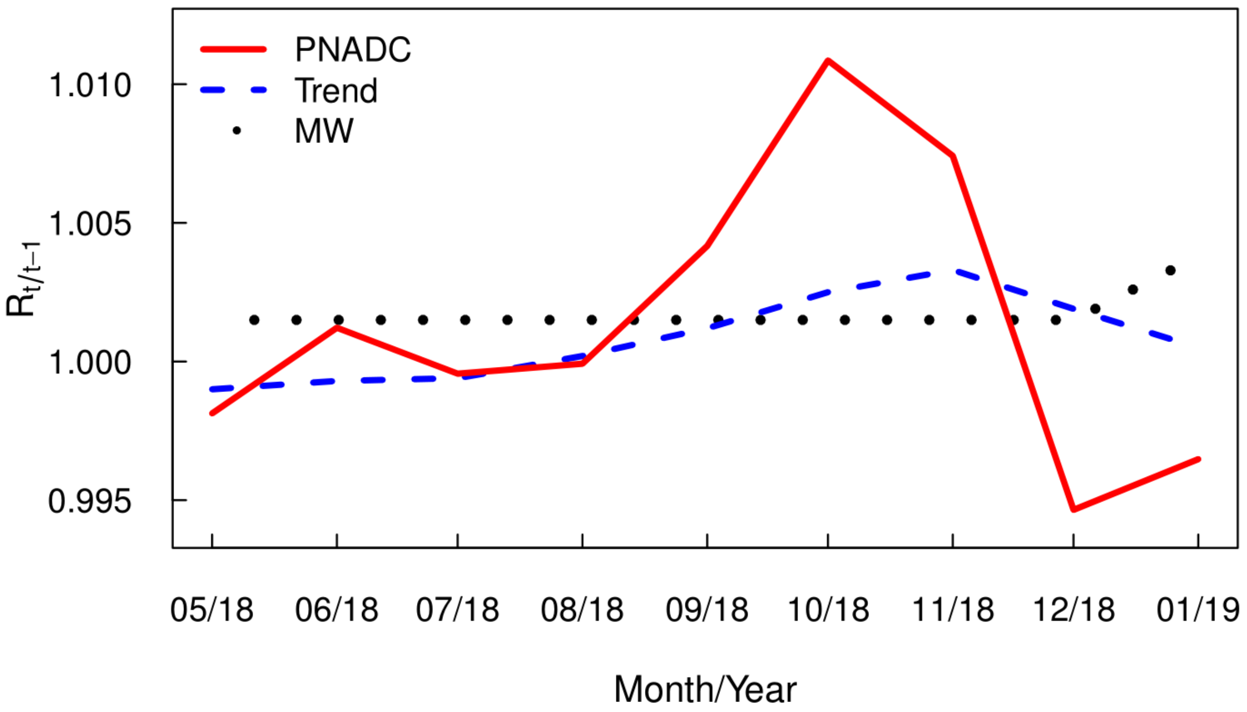


Figure 6: Comparison of estimates for the prices relatives of the subitem MORD. The solid, dashed and dotted lines represents, respectively, the PNADC, trend and minimum wage estimates. The PNADC estimate are those derived via the procedures described in Section 4 and the “trend” is the trend component of such series derived by the procedure described in Section 5.

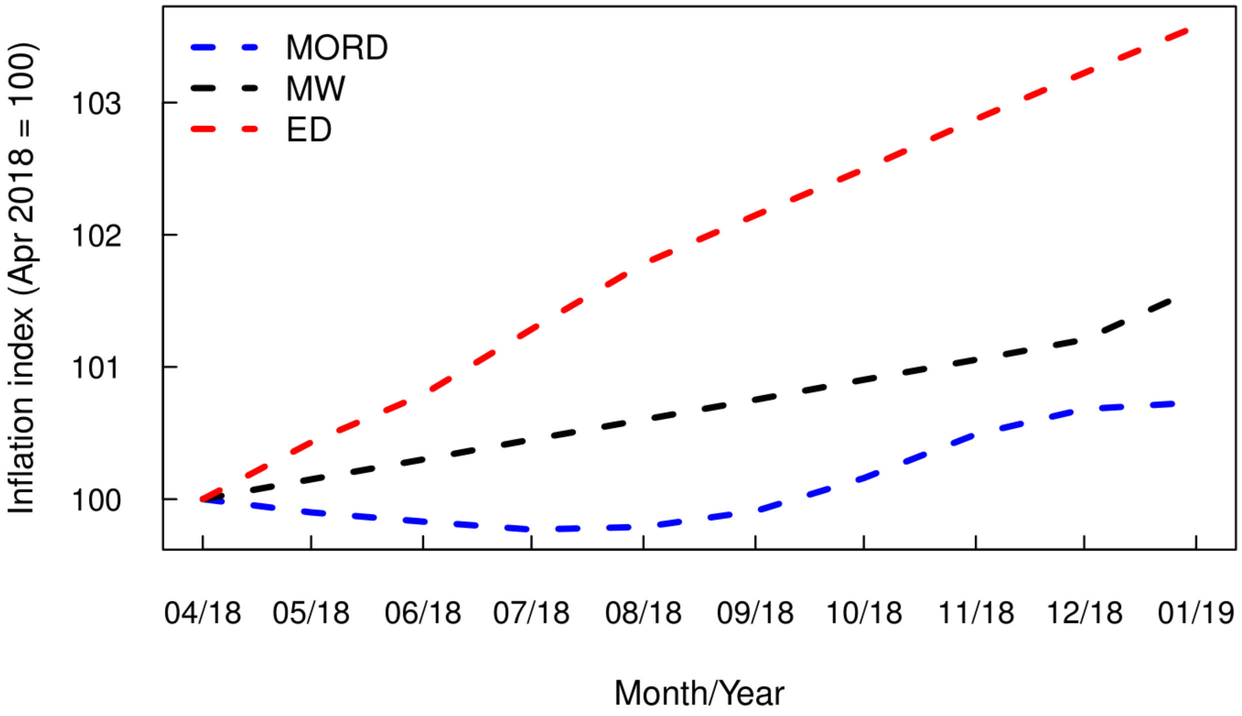


Figure 7: Index number series derived for the series of estimates of the trend components of the subitems ED (red) and MORD (blue). Also shown for comparison the series obtained via the minimum wage estimates.

6.2. Prices microdata

The previous Sections described in detail the methodology derived for the estimation of the inflation of the subitems ED and MORD via the PNADC survey. For researchers interested in working with the prices microdata obtained from the PNADC for the subitems ED and MORD, IBGE has made available monthly files containing the prices of matched individuals in months t and $t-3$. The final SNIPC sample weights for an individual of the domains of interest $w_l^{t_{SNIPC}}$ derived in Section 4 above is also provided in such files. The data available is the one before the outliers detection and treatment process described in Section 4.2 and the time series approach presented in Section 5. The microdata can be downloaded in the link[7].

The data is only available for the national level of aggregation since for finer levels the results still presented high volatility and their use for the calculation of the CPI was not considered appropriate.

7. Conclusions

In this work the problem of compiling the inflation for housekeeper and services for home maintenance, two services components of the IPCA basket that are mainly characterized by informal jobs activities, is presented. The main challenges in moving from the PME survey, previous source of labour's force information, to the PNADC survey are presented. The details for the derivation of an estimator for the inflation of such subitems taking into account the complex sample design and sample rotation scheme of the PNADC are discussed. The main steps necessary for the correct use of the PNADC data are scrutinized. Finally, the structural modelling times series approach adopted for the calculation of the estimates of the trend of the PNADC prices data are discussed. The results obtained via the "raw" PNADC estimator and the trend component estimator are compared with the estimates obtained via the brazilian official minimum wage. The estimates obtained via the time series modelling approach show good agreement with the national economical scenario for the period analysed revealing the power of the method and its superiority respective the minimum wage approach which is insensitive to month conjunctural changes in the economy affecting the labour's market.

This work also serves as a tutorial for users interested in working with the prices microdata presented in Section 6.2 for the ED and MORD subitems.

References

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964. ISSN 00359246. URL http://www.jstor.org/stable/2984418.

J. Durbin and S.J Koopman. *Time Series Analysis by State Space Methods.* Cambridge University Press, 2001.

---

[7] After clicking on the link, click on **Arquivos complementares para o cálculo dos índices do SNIPC** and then select the year of interest in the page. After that a set of txt files for the months of the selected year will be displayed for selection and download.

M. P. S Freitas and A. A Antonaci. Sistema integrado de pesquisas domiciliares amostra mestra 2010 e amostra da pnad contínua. *Textos para discussão Diretoria de Pesquisas*, 50, 2014. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv86747.pdf.

A.C Harvey. *Forecasting, Structural Time Series and the Kalman Filter.* Cambridge University Press, 1989.

IBGE. Pesquisa mensal do emprego. 2 ed, 2007. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv37313.pdf.

IBGE. Sistema nacional de índices de preços ao consumidor: métodos de cálculo. *Série relatórios metodológicos*, 14, 2013. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv65477.pdf.

IBGE. Pesquisa nacional por amostra de domicílios contínua. *Notas metodológicas*, 1, 2014a. URL ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Notas_metodologicas/notas_metodologicas.pdf.

IBGE. Sistema nacional de índices de preços ao consumidor: estruturas de ponderação a partir da pesquisa de orçamentos familiares 2008-2009. *Série relatórios metodológicos*, 39, 2014b. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv86320.pdf.

IBGE. Nota tecnica pnadc 1: Esclarecimentos sobre os resultados da pnad contínua produzidos mensalmente. 1, 2015. URL ftp://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Mensal/Notas_tecnicas/nota_tecnica_01_pnadc_mensal.pdf.

IBGE. Nota técnica 03/2016: empregados domésticos e mão de obra de pequenos reparos. *SISTEMA nacional de índices de preços ao consumidor.*, 2016a. URL ftp://ftp.ibge.gov.br/Precos_Indices_de_Precos_ao_Consumidor/Sistema_de_Indices_de_Precos_ao_Consumidor/Notas_Tecnicas/snipc_nota_tecnica_2016_03.pdf.

IBGE. Pesquisa nacional por amostra de domicílios contínua. *Notas técnicas*, 1.2, 2017. URL https://biblioteca.ibge.gov.br/visualizacao/livros/liv101392.pdf.

IBGE. Metodologia de cálculo da inflação dos subitens do snipc empregado doméstico e mão de obra para reparos domésticos a partir das informações de rendimentos da pnad contínua. *Paper submitted to the 16th Meeting of the Ottawa Group, Rio de Janeiro, Brazil.*, 1, 2018.

IBGE. Ipca month weights. 2019. URL https://sidra.ibge.gov.br/tabela/1419#/n1/all/n7/all/n6/all/v/66/p/last%201/c315/all/d/v66%204/l/,p+t+v,c315/resultado. Last access Feb, 20, 2019.

ILO. Consumer price index manual: Theory and practice. 2004.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME – Journal of Basic Engineering*, (82 (Series D)):35–45, 1960.

Vladimir G. Miranda, Pedro K. da Costa, Rodrigo V. Ventura, and José Fernando P. Gonçalves. Consumer price indices at ibge: 40 years and counting. *Paper submitted to the 16th Meeting of the Ottawa Group, Rio de Janeiro, Brazil.*, May 2019.