UNITED NATIONS **E**

Economic and Social Council

WORKING PAPER NO 2

7 April 2006

ENGLISH ONLY

ECONOMIC COMMISSION FOR EUROPE **STATISTICAL COMMISSION**

CONFERENCE OF EUROPEAN STATISTICIANS

Group of Experts on Consumer Price Indices

Eighth Meeting
Geneva, 10-12 May 2006

STABILITY OF CPIs IN CONSIDERATION OF CHANGES IN A SAMPLE SIZE*

Invited paper submitted by the Regional Statistical Office of Poland

The meeting is organised jointly with the International Labour Office (ILO)

* This paper has been prepared by Waldemar Dubla, Anna Cynkier and Arkadiusz Grabski, Regional Statistical Office of Poland, at the invitation of the Secretariat. Paper posted on Internet as received from the authors.

Joint UNECE/ILO Meeting on Consumer Price Indices
(Geneva, 10-12 May 2006)

Stability of CPIs in consideration of changes in a sample size[1]

[1] By Regional Statistical Office in Lodz, Poland: Waldemar Dubla – deputy director, Anna Cynkier - specialist, Arkadiusz Grabski - specialist.

The views expressed are those of the writers and may not be considered as an official position of the Polish Central Statistical Office.

1. Introduction

Statistical surveys in which data are collected by questioners belong to the most expensive among these conducted by official statistics. That is why a thorough analysis should be done to determine sample size. The analysis presented below gives information on that in relation to practical solutions accepted in Polish official statistics.

Two possible results of such analysis *may* (but not *have to*) lead to the conclusion that the sample size should be enlarged (if the index numbers are sensitive to the change of sample size) or decreased (if not). Of course one should be very careful making decision as reduction of sample size *always* results in decreasing of expected precision. Moreover in the future there will probably appear demand for information on index numbers for small areas that may lead to quite contrary decisions. That's why the paper should be treated as an initial attempt of analysis of the problem and not as a proposition of decision to be taken up.

The most evident way of sample size reduction is a random procedure. The paper present somewhat different approach assuming that among random sub-sets of reduced commodities there can be found some giving better results than the others.

2. Basic information on the data used for simulations

Monthly data on prices of products and services included in CPI calculation for 2004 were analyzed. The data collection was conducted by specialized questioners in 307 areas (PODs – Price Observation Districts) regularly distributed all over the country (towns or districts in the largest cities).

The list of products, identical for all PODs, included 1768 products and services observed all the year long. For most of them prices were observed once a month. It does not concern fruits and vegetables observed three times a month.

The list was than divided into 312 homogeneous elementary groups of consumption (297 groups including at least one product). The number of elements in groups was distributed as it is shown in the table.

Number of	
groups	**commodities in groups**
64	1
57	2
37	3
28	4
17	5
15	6
12	7
14	8
9	9
5	10
5	11
7	12
4	13
1	14
2	15
2	16
1	17
2	18
3	19
2	21
1	23
1	25
1	27
1	30
1	37
1	39
2	56
1	72
1	90
297	**1777***

* 1777-9=1768 – 9 (9 products belong to more than one group)

Apart from the above there were also 131 seasonal products observed only for some months.

3. Methodology of CPI calculation.

To calculate CPI Central Statistical Office of Poland uses a standard procedure similar to the one utilized by most countries. The general idea of this procedure is as follows.

First, monthly mean values for prices in PODs are calculated in relation to products observed by questioners depending on the number of observations during a month. For products with unique (centrally observed) prices weighted averages are calculated. These are used to calculate index numbers for products in PODs as relations of calculated monthly averages to average prices from the previous year. For each product a country index number is calculated as a geometrical mean value for all the PODs. These

are the base to calculate aggregate index on the lowest level of aggregation (for elementary groups of products) – it is a geometrical mean value of separate index numbers for products in the group.

In following stages the weights representing shares of the groups in total consumption are introduced to calculate index numbers for higher levels of aggregation – up to total country's CPI.

CPIs are calculated for every month according to Laspeyres formula. Weights represent the expenditure structure observed in previous year Household Budget Survey.

The above procedure was applied in simulations with some simplifications implied mainly by technical reasons:

⇒ Index numbers for prices in PODs were calculated as relations to January 2004 observations instead of to previous year averages.

⇒ In case of necessary substitution of product disappearing from the market the procedure was as follows:

Example

In period t+2 commodity „A" is substituted by commodity „B". Technically the index number is than calculated in the following way.

Period	**t**	**t+1**	**t+2**
Observations of „A" prices in POD	$P_{A,t}$	$P_{A,t+1}$	-
Observations of „B" prices in POD	-	$P_{B,t+1}$	$P_{B,t+2}$
Index numbers	x	$P_{A,t+1}/ P_{A,t}$	$P_{B,t+2}/P_{B,t+1}^{(*)}$

(*)

$$\frac{P_{B,t+2}}{P_{B,t+1}} = \frac{P_{B,t+2}}{P_{A,t+1}} \cdot \frac{P_{A,t+1}}{P_{B,t+1}}$$

⇒ Prices that could not be observed for a given period were not included (i.e. by imputation) into calculation of the group index. In case of no observation in the group index value = 1 was assumed.

⇒ No calculations were made for groups represented by no product (they were not included into calculated indices).

4. Reduction Methods

Monthly indices for prices assuming prices for January 2004 = 100 were inputs for analysis. Seasonal products were excluded (131 products belonging to 43 groups, of which 7 groups entirely consisting of seasonal commodities).

Reduction was conducted inside groups. It was arbitrary assumed that reduction is allowed only for groups consisting of at least 5 products. It reduced the eliminating procedure to 100 groups as the whole picture was as it can be observed in the table below:

Groups consisting of	Number of	
	groups	commodities
< 5 commodities	190	407
>= 5 commodities	100	1239
Total	**290**	**1646**

Elimination procedure was conducted using two methods (in a couple of variants):

⇒ *Method I – correlation analysis,*

⇒ *Method II – least discrepancies method.*

Two parameters determining reduction variants were accepted for Method I mainly maximal percentage of commodities to be removed in the group (U = 20, 25 and 30%) and minimal correlation coefficient (C). Only the first one (U = 10, 15, 20 and 25%) was assumed in case of Method II.

5.1. Method I

One may argue that commodities belonging to homogeneous group behave on the market in a similar way. So probably at least some of them represent similar information like the rest of the group representatives. In that case these showing the highest average correlation with the set of the others can be removed with a very slight impact on the aggregate group index number.

Two conditions were assumed to determine products to be eliminated:

⇒ **condition 1:** the selected products show the highest average correlation with the rest of products in the group;

⇒ **condition 2:** the selected products show the correlation exceeding accepted level with an assumed number of product in the group (that means only these products are accepted that show high correlation with the greatest number of other products in the group)

The calculations were additionally complicated by the fact that some prices did not show any change during the year 2004. In those cases constant prices products were randomly reduced (there were 146 such products).

5.2. Method II

It can be easily proved that the lowest impact on the aggregate index number would be observed in case of removing the products whose price indices are close to average in the group. To identify these the following measures have been calculated:

⇒ average monthly price index number for every group of products I_G,

⇒ differences between monthly price index number for products and index for the group: $R_{I_{P_i,G}} = I_{p_i} - I_G$,

⇒ the sum of differences in 11 analyzed months ($R_{I_{P_i,G}}$),

Commodity *i* may be removed if

$$R_{I_{P_i,G}} \geq ZR_{I_{P_i,G}},$$

where ZR denotes the level imposed by assumed level of U.

5. Results

To analyze received results indicators have been calculated for each case taken into account, for aggregate yearly index as well as for monthly indices. Differences between original indices (non-reduce indices) and indices incorporating reducing factor influence are presented in the tables and figures together with the number of removed products.

5.1. Method I

For correlation method the following cases were regarded:

Parameters	Variants			
	1	2	3	4
U (%)	20	25	30	20
Correlation variant	A	A	A	B

The fourth variant is the most restrictive as it needs maximum 20% of products reduction and implies minimum (at least ½ of average in the group) correlation coefficient with all the products in the group. The scale of reduction can be observed in the table below:

Reduction	Variants			
	1	2	3	4
In absolute values	188	256	306	152
% %	11	14	17	9

Comparison of original index with indices calculated in various variants implies the following conclusions:

⇒ all the index numbers for reduced sets of products are lower than original index by -0.0063 to -0.0104;

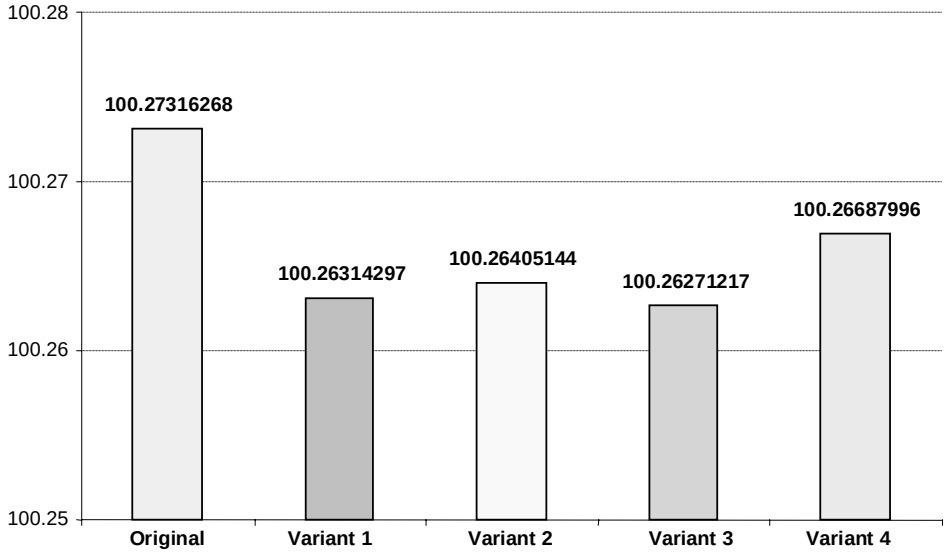
⇒ variant 4 leads to the result closest to original index number;

⇒ results in variant 2 are close to these in variant 4 but the number of removed products is much larger (by 104 elements).

Synthetic results are presented in the table and figure below.

	Annual index numbers (% %)				
	original	Reduction variants			
		1	2	3	4
Index value	100.27316268	100.26314297	100.26405144	100.26271217	100.26687996
Difference	-	-0.01001971	-0.00911124	-0.01045051	-0.00628272

Fig 1. Method I
Yearly index numbers in 2004 (original and after reductions)



As for monthly indices most important conclusions are following:

⇒ all the „reduced" indices are lower than original;

⇒ no matter what variant is considered reduced indices reach maximum difference for May/April;

⇒ the smallest difference (-0.0002) is observed for variant 4 and the largest (-0.0477) for variant 3;

⇒ monthly indices support conclusion coming from yearly results mainly that slight differences in variant 2 are accompanied with relatively high number of reduced commodities.

The table below presents the calculated index numbers.

Period	Monthly index numbers				
	original	reduced (variants)			
		1	2	3	4
Feb/Jan	100.1231	100.1101	100.1071	100.1014	100.1161
Mar/Feb	100.2903	100.2864	100.2743	100.2713	100.2901
Apr/Mar	100.5351	100.5263	100.5156	100.5149	100.5303
May/Apr	100.9131	100.8875	100.8786	100.8654	100.8921
Jun/May	100.327	100.3210	100.3197	100.3155	100.324
Jul/Jun	100.2138	100.2024	100.2011	100.1964	100.2044
Aug/Jul	99.6874	99.6755	99.6703	99.6662	99.6808
Sep/Aug	99.9891	99.9802	99.9760	99.9721	99.9847
Oct/Sep	100.4926	100.4837	100.4878	100.4871	100.488
Nov/Oct	100.2219	100.2178	100.2174	100.2165	100.2194
Dec/Nov	100.2163	100.2115	100.2092	100.208	100.2135

In the following table discrepancies between original and reduced indices are presented.

Period	Discrepancies in variants			
	1	2	3	4
Feb/Jan	-0.0130	-0.0160	-0.0217	-0.0070
Mar/Feb	-0.0039	-0.0160	-0.0190	-0.0002
Apr/Mar	-0.0088	-0.0195	-0.0202	-0.0048
May/Apr	-0.0256	-0.0345	-0.0477	-0.0210
Jun/May	-0.0060	-0.0073	-0.0115	-0.0030
Jul/Jun	-0.0114	-0.0127	-0.0174	-0.0094
Aug/Jul	-0.0119	-0.0171	-0.0212	-0.0066
Sep/Aug	-0.0089	-0.0131	-0.0170	-0.0044
Oct/Sep	-0.0089	-0.0048	-0.0055	-0.0046
Nov/Oct	-0.0041	-0.0045	-0.0054	-0.0025
Dec/Nov	-0.0048	-0.0071	-0.0083	-0.0028
Average difference	**-0.0098**	**-0.0139**	**-0.0177**	**-0.0060**
Maximum	-0.0256	-0.0345	-0.0477	-0.0210
Minimum	-0.0039	-0.0045	-0.0054	-0.0002

6.2. Method II

The reduction scale for least discrepancies method is presented in the table below.

	Reduction in variants			
	1 (10%)	2 (15%)	3 (20%)	4 (25%)
In absolute values	67	135	210	273
% %	4	8	12	15

Comparison of original index number with reduced sample size indices lead to the following main conclusions:

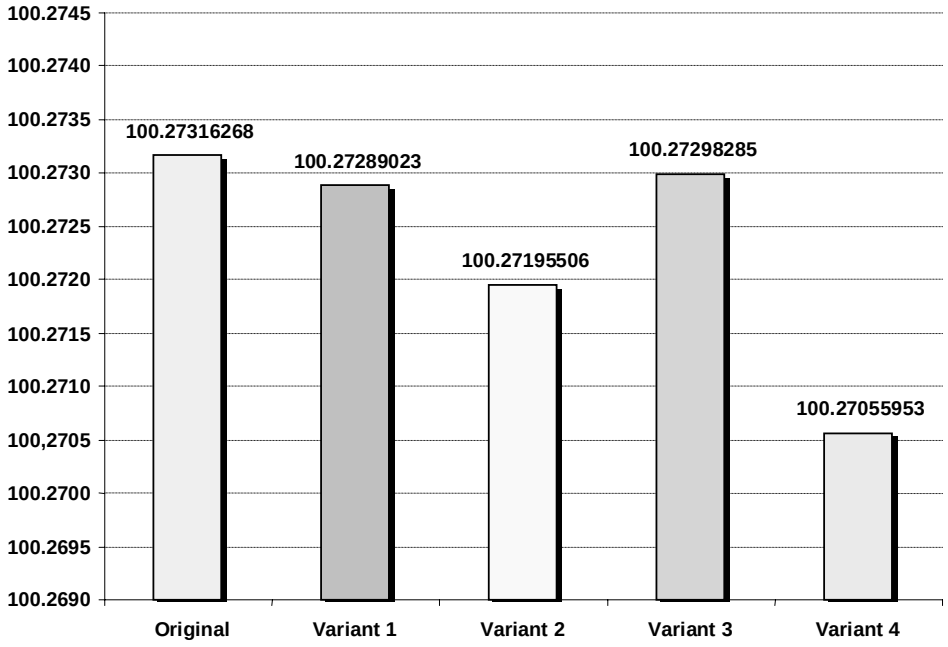
⇒ all the reduced indices are slightly lower than original index (by -0.002603% to -0.000179%);

⇒ the smallest difference is observed in case of variant 3 (!); this variant allows for much higher reduction than variant 2 and especially variant 1.

Results are presented in table and figure below.

	Annual index numbers (%%)				
	original	**reduced (variants)**			
		1	**2**	**3**	**4**
Index value	100.27316268	100.27289023	100.27195506	100.27298285	100.27055953
Difference	x	-0.00027245	-0.00120762	-0.00017983	-0.00260315

Fig 2. Method II
Yearly index numbers in 2004 (original and after reductions)



Analysis of monthly indices allows for the following general observations:

⇒ index numbers in all the variants concentrate below or over the values of original index;

⇒ absolute values of discrepancies are relatively low and are located between 0.0002 and 0.0159.

The table presents received results.

Period	Monthly index numbers (%%)				
	original	reduced sample size (variants)			
		1	2	3	4
Feb/Jan	100.1231	100.1216	100.115	100.1138	100.1102
Mar/Feb	100.2903	100.2885	100.2885	100.2905	100.2887
Apr/Mar	100.5351	100.5362	100.5333	100.5356	100.5332
May/Apr	100.9131	100.9107	100.9069	100.9034	100.8972
Jun/May	100.327	100.3279	100.3331	100.3379	100.3265
Jul/Jun	100.2138	100.2131	100.2152	100.2182	100.2166
Aug/Jul	99.6874	99.6868	99.6845	99.6828	99.6818
Sep/Aug	99.9891	99.9886	99.9905	99.992	99.9965
Oct/Sep	100.4926	100.4936	100.4922	100.4933	100.4957
Nov/Oct	100.2219	100.2236	100.2227	100.2245	100.2241
Dec/Nov	100.2163	100.2161	100.2145	100.2157	100.2105

The following table presents discrepancies between original and sample size index numbers.

Period	Differences (%%)			
	1	2	3	4
Feb/Jan	-0.0015	-0.0081	-0.0093	-0.0129
Mar/Feb	-0.0018	-0.0018	0.0002	-0.0016
Apr/Mar	0.0011	-0.0018	0.0005	-0.0019
May/Apr	-0.0024	-0.0062	-0.0097	-0.0159
Jun/May	0.0009	0.0061	0.0109	-0.0005
Jul/Jun	-0.0007	0.0014	0.0044	0.0028
Aug/Jul	-0.0006	-0.0029	-0.0046	-0.0056
Sep/Aug	-0.0005	0.0014	0.0029	0.0074
Oct/Sep	0.001	-0.0004	0.0007	0.0031
Nov/Oct	0.0017	0.0008	0.0026	0.0022
Dec/Nov	-0.0002	-0.0018	-0.0006	-0.0058
Average difference	**-0.00027**	**-0.00121**	**-0.00018**	**-0.00261**
Maximum (abs)	0.0024	0.0081	0.0109	0.0159
Minimum (abs)	0.0002	0.0004	0.0002	0.0005

7. Conclusions

⇒ No matter what is the method of the number of representative commodities reduction the differences between resulting indices and original index numbers are very small. One could say then that for this level of aggregation the index value is not sensitive to the change of the number of observed prices. Additional analysis is necessary to determine changes in case of small area indices (lower aggregation levels, regional indices). It would also be of interest to extrapolate calculations outside the estimation period.

⇒ Comparison of two presented methods of reduction leads to the conclusion that can be expressed as follows: better results were obtained for method of minimalization of differences between reduced set index numbers and original indices. Nevertheless both methods possess negative sides.

⇒ The main weakness of correlation method is systematic error generated by the applied way of reducing of the number of products. As on the market price increases are observed much more frequently than decreases the correlation coefficients as a rule more frequently are high in relation to increasing prices. So it is increasing price product which in most cases is eliminated. Applying of the method would be possible only on condition it was seriously modified.

⇒ Second method seems reasonably precise. Nevertheless it is worthwhile to notice the set of eliminated commodities covers mainly these most typical for the group (most representative). That in turn lies in a little bit strange position – in an opposition to *common sense* thinking; one can say that commodities should be chosen to give the best representation of the group. Of course it may be answered that it is the whole set that should be representative not individual commodity. But...
