

The Norwegian CPI

The System for Data Validation and Editing

Prepared by Tom Andersen Langer, Statistics Norway

Abstract

The current IT-system for data validation and editing was introduced in the late 1990-ies being part of a general reengineering process for the CPI system. A number of innovations were introduced including a system for automatic data validation and editing. The system utilises several approaches and methods in identifying likely types of errors in data. The main tool for flagging such errors is the HB method designed for testing the distributions of price changes on detailed product level utilising the median and the quartiles. From the outset the HB method has been combined with a normalised test (Chebychef) utilising the mean and the standard deviation from the same distributions. A basic decision principle is that observations to be subject for inspection should be flagged using both methods.

The editing procedures are semi-automatic in the sense that the system provides a suggestion how to handle a specific case. The suggestions given are based on a predefined set of algorithms selected according to a procedure catalogue for this purpose. All flagged observations will be subject for inspection.

This paper starts out providing some background on the CPI data. This is to give the reader a basis for understanding the Norwegian system for data validation and editing system. The paper continues discussing some of the main elements in the system for data validation.

1. The CPI data capture system

The CPI uses a Laspeyres type of formula (Lowe) in a system based on a chained index with annual links. July each year is the price reference period for the short term index which is estimated for a 12 month period. Each month the aggregates produced in the short term index are linked to the respective long term index aggregates. The reference year of the long term index is 1998.

The basic sample of units comprises some 2200 establishments each participating for a 6 year period. The sample covers units located in 8 geographical areas (regions). Each respondent participates for a 6 year period i.e. 1/6 of the sample is substituted every year – a process starting in January each year. Sampling of units is based on a PPS-procedure (probability proportional to size).

No price collectors are involved in regular data capture. However, in front of starting regular reporting each new respondent is visited and training provided in using the reporting tool preferred. The training also comprises assistance on selecting products for regular reporting as well as how to handle situations when products are temporarily or permanent out of sale, special sales activities, product changes etc.

The prices reported are defined as point-in-time observations relating to 15th each month or in practise the mid week of the month. The basket of non-food products comprises approximately 750 specified products and the number of observations each month adds up to 39 000.

As concerns food and beverage products scanner data is collected comprising some 300 000 observations and 3000 products each month. In this context the product is defined utilising the product specific EAN code and additional information in text describing the product.

All data capture tasks are located to a specialised unit for this purpose in Statistics Norway. Three separate data capture methods are in use:

- a) Internet based reporting available for most types of respondents
- b) Postal reporting using a traditional paper questionnaire. When returned from the respondent the questionnaires are scanned and interpreted using an optical reading tool.
- c) Scanner data reported from the large retail chains mainly covering products relating to COICOP 0 Food and beverages.

The c) category comprises administrative data from government or semi-government institutions for product categories like automobiles and alcoholic beverages. It should be added that the category c) data reported are not a part of the system for validation and editing described in this paper.

The 27th each month is a deadline for the revision process. The data capture unit delivers a data set comprising all information reported which provides the starting point for the regular validation and data editing processes. Not responding units might later on be fined – according to the Statistical Law.

The monthly revision process continues for the next two weeks until the release of the index on the 10th each month.

2. The main elements of the Norwegian system for data validation and editing

The CPI system consists in large of the four elements below.

- 1. Handling doublets, likely key punch errors and reported decimal errors.
- 2. Identifying or flagging extreme observations
- 3. The treatment of partial and total non response
- 4.1 Controls and editing on regional level
- 4.2 Macro control

Table 1 provides a background and sketches the proportions of the data set being handled on a monthly basis in the system.

Table 1. CPI Revision System. Some key figures 2006:08 - 2007:07. Monthly average

COICOP	A. No of observations	B. Non response	Total = A + B	C. No of interventions by SM	C, pct of Total	B, pct of Total
Total¹	33 166	5 756	38 922	561	1,4	14,8
2	1 681	281	1 962	7	0,4	14,3
3	6 072	1 357	7 429	194	2,6	18,3
4	675	137	812	13	1,6	16,9
5	6 615	1 201	7 816	110	1,4	15,4
6	4 014	520	4 534	45	1,0	11,5
7	2 445	220	2 665	21	0,8	8,3
8	267	55	322	30	9,3	17,1
9	3 534	702	4 236	52	1,2	16,6
11	2 750	390	3 140	32	1,0	12,4
12	5 113	893	6 006	57	0,9	14,9

¹ Excluding COICOP 1 Food and beverages and group 10 Education
SM - Subject matter specialist

As is emphasised in appendix A the COICOP group 1 Food and beverages is not part of the regular system for data validation and editing described in this paper. This group is solely based on scanner data which in number of observations per month are 10 times the total number of observations given

in table 1. A separate data validation system has been developed for this group. For more background information on the organisation of the CPI and the data capture system – see appendix A.

Before going into the main elements some additional background should be provided. The main arguments for introducing automatic procedures were that extremes observations should be identified and handled in a standardised way and independent of individual judgements. The experiences from running the system for some years provide support for saying that this target has been achieved. An equally important ambition emphasised was that the system should allow for a more efficient use of the competence of the subject matter specialists in this field. Even on this point one can say that the ambition has been met.

3.1 Handling likely key punch errors or decimal errors

This is a minor element of the system measured in numbers of observations being subject for inspection each month, though still very important. The purpose is to have an initial cleaning of the data set removing likely errors. The decimal error refers to cases where a current month observation is e.g. 10 or 100 times the last month observation or the July observation. The observations being subject for inspection are detected automatic.

3.2 Identifying / flagging extreme observations

The HB method is used as the main tool in this process step. This method is combined with a normalised test based on the same set of data. A basic decision principle is that observations to be subject for further inspection / editing should be flagged in both methods. In this part we will concentrate on the HB method and the CPI set up for this test.

Some sampling issues

To optimise the properties of the two methods in use, the group of observations being tested should be fairly homogenous. In addition the number of observations in the group should be large enough to allow for robust estimation of the median and the quartiles (HB method) and mean and standard deviation (the normalised test).

A typical data set being subject for analysis in CPI comprises all observations for a specific product e.g. pair of shoes from outlets located in a given region (regional product group). The CPI works with 8 regions.

In some cases the number of observations in the CPI regional product group might not be sufficient to provide robust estimates for the median and the quartiles. In such cases the group is expanded to cover data for all 8 regions.

Standard test variables

The system works with two standard test variables.

T1 Price current month / price for preceding month

T2 Price current month / price in the reference period (July)¹

It is the current month observation that is subject for control. The reason for having two test variables is that the causes behind extreme changes normally are manifold covering aspects like seasonality, product changes, sales offers, new product etc. The intention with T2 is mainly to cover seasonality aspects throughout the 12 month period.

¹ July each year is the price reference period in the short term index. See appendix A for some more about this.

For a given regional product group a distribution of price relatives is established for each of the two test variables.

Transforming the test variable distributions

As a standard procedure two transformations are made in the CPI system. The first step implies making the distributions symmetric around the median value of the original price relative distribution. The second step involves taking the price levels into consideration. This involves a U parameter. The outcome of the second step is an effect distribution using the terminology introduced by the Hidiriglou and Berthelot.

Accept intervals

The accept intervals for the T1 or T2 distributions of a CPI regional product group could be written as below:

$$\text{Lower level (T1)} = E^m - C \cdot \max(E^m - E^{q1}; A \cdot E^m)$$

$$\text{Upper level (T1)} = E^m + C \cdot \max(E^{q3} - E^m; A \cdot E^m)$$

Where E refers to effect, m to the median, q1 and q3 to the quartiles while C and A refers to 2 of the 3 parameters in use. The C parameter is used to expand the interval around the E^m , while the A parameter is introduced to handle cases where the $E^m - E^{q1}$ or $E^{q3} - E^m$ is equal to 0. This could especially be relevant for prices which might remain unchanged over time.

The impact of the parameters

The C, A and U parameters contributes in different ways to fixing the upper and lower levels and thus the number flagged observations. Table 2 shows some results from a test performed on CPI data.

Table 2 HB method and the effect of parameters

All tests are performed on distributions based on price changes compared to last month (T1).

Estimations are based on data from February - March 2004

The effect of parameter C - with A = 0,05 and U = 0,5

<u>C-values</u>	<u>No of flagged extremes</u>
8	322
12	249
16	204

The effect of parameter U - with A = 0,05 and C = 12

<u>U-values</u>	<u>No of flagged extremes</u>
0,0	244
0,5	249
1,0	282

The effect of parameter A - with U = 0,5 and C = 12

<u>A-values</u>	<u>No of flagged extremes</u>
0,0	249
0,5	243
1,0	239

This test indicates that the impact of the A parameter is very small measured in changes in the number of flagged observations. This result can not be generalised due to that it might be dependent of the test period selected.

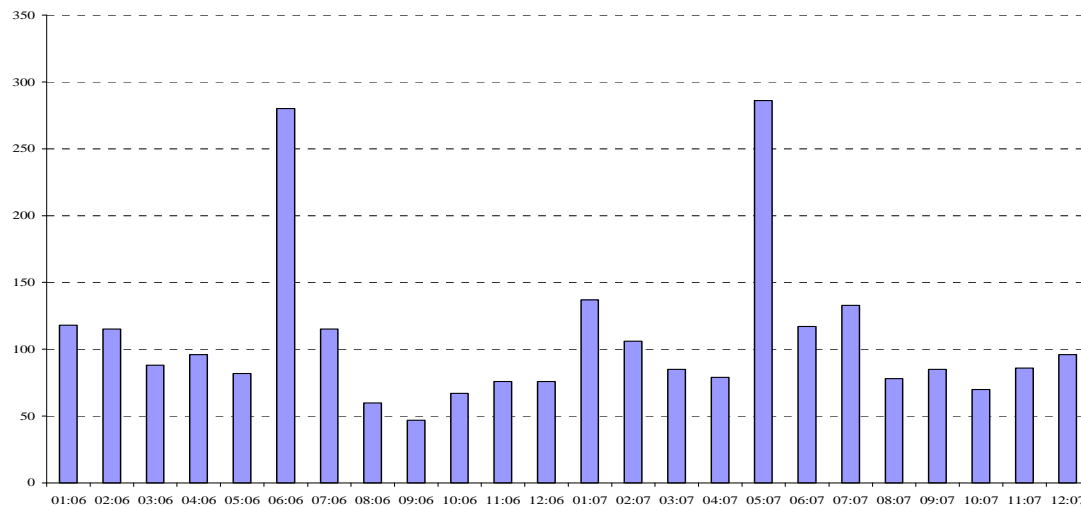
The variations in the C parameter shows a much larger impact measured in changes in the number of flagged observations. This is as one would expect simply due to that purpose of this parameter is to expand / delimit the width of the accept interval.

The U parameter is introduced in the transformation step 2 to allow for taking the price level into consideration. A U parameter = 0 implies that the price level does not influence the effect distribution while a U = 1 allows for maximum influence.

The CPI method and flagged observations

In the graph presented below a time series for flagged observations per month in CPI is presented.

CPI-method 2006:01 – 2007:12. Number of extremes flagged per month



For the period 2006 – 2007 the number flagged observations are fairly stable around 100 on a monthly basis though with June / July figures being larger. When compared with the numbers of flagged observations in table 2 (March 2004) the differences is substantial. The main reason for the drop in the number of flagged observations after 2005 is that the food and beverages group (based on scanner data) were taken out of the regular system for data validation and editing.

Editing flagged observations

Flagged extremes are passed on to subject matter specialists using a computer based system for editing. The system offers automatically suggestions on how to make corrections for the flagged observations based 4 predefined algorithms.

Three of the algorithms provide estimated growth rates for the regional product group or any other group selected. The growth rates estimated are arithmetic averages or geometric averages from the non flagged data from the same group. The last algorithm provides an estimate of the average price of the group concerned. The algorithm selected by the system depends in practise on the data available for the flagged observation. The suggestion offered to the subject matter specialist is selected according to a predefined procedure catalogue for this purpose. The same type of algorithms are also used when handling partial and total non response.

The subject matter specialist decides on what type of action to be taken. All actions taken are recorded allowing for an analysis of types errors identified and how they are handled.

One of the experiences made is that a non negligible number of the flagged observations reflect real price changes. Due to this the subject matter specialist are advised to utilise all additional information available on each observation concerned. In the questionnaires the respondent is required to provide qualitative information providing causes for large changes in price levels. This might be information telling that the current month price refers to a new product, that quality is changed, reflects sales prices etc. This type of information might be very important when taking correct decisions.