

DISTR.
GENERAL

CES/AC.49/1997/20
13 August 1997

Original: ENGLISH

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

INTERNATIONAL LABOUR ORGANIZATION

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/ILO Meeting on consumer price indices
(Geneva, 24-27 November 1997)

Item 9 of the provisional agenda

PROBLEMS OF DATA COLLECTION AND QUALITY CONTROL OF DATA

Paper submitted by the Czech Statistical Office

I. Problems of data collection

1. For calculating the CPI index, more than 88.800 individual prices are used in the database each month. When identifying more precisely what this huge sum of data is composed of as far as the collection and control purposes are concerned, the following specification can be made:
2. Certainly, not all of those 88.800 prices change every month. Some 19 percent of them moved in April 97, for instance. About 8 percent of the total volume of items are still centrally controlled prices. These prices are for example natural gas, electricity, public transport, telecommunication services, etc. Each month check on the validity of these prices is made.
3. Further, the data (except centrally controlled prices) come from around 10,000 outlets, and collected are for 749 items chosen by purposive sampling.
4. In terms of geography, the prices of these items are collected in 41 regions plus Prague by 53 trained collectors. The price collectors have to check usually 3 prices per item, in more inhabited areas 5 prices per item, and in Prague up to 16 prices per item. The collectors receive no data by mail. Telephone is not widely used.
5. The software used for data capture has been maintained by a well-established professional external company. In 1998 we plan to remake this program ourselves by adding some more identifiers to help us have a better

control over the places of data collecting. The program should enable us to use pen-pad computers, too. Another strength of the new program should be the use of our own software prepared specially for the purposes of the Czech Statistical Office with lots of connections to databases used in other statistical fields than price statistics.

6. Up to 1995, the above-mentioned external company also maintained the databases. This has been changed. As a result, headquarters is able to have more effective control over the data since.

7. The database is maintained by the system Oracle 7, the operating system is Novell 4.1, the working stations are equipped with the processor Intel 80486 or Pentium, 16 MB RAM.

8. The price collectors are employees of the Czech Statistical Office. They are regularly trained using meetings with the headquarters staff. These meetings take place at least twice a year and the following topics are mostly discussed:

- explaining misinterpretations connected mainly with item varieties;
- assistance in solving technical problems, if any;
- assistance and encouraging of the collectors to build up good Statistical Office image in the eyes of the outlets.

A brief guide to consumer price collection provides help to the staff, too.

II. Control of the data

9. On the 20th day of each month, the price collectors send their separate databases by means of modem devices to a CSO-CZ Regional Division. There are 7 days to check up the whole database and within next 7 days mistakes, if any, have to be corrected. Each implausible data is discussed with the price collector prior to its possible correction.

10. The majority of control tasks take place with the help of report sheets made by prepared programs. This is especially the report of maximum and minimum prices/indices, program of the completeness of notes and comments, unless they seem plausible. The price collector is contacted by phone and he/she either has to give a clear explanation or has to return to the outlet, verify the price, and call the result back.

11. In addition to these programs, all members of headquarters controlling staff are able to make their own ad hoc queries in the Oracle database through SQL link (Structured Query Language). This is done by defining the query in the Database Desktop options either in Paradox or in Quattro-Pro. This facility is also very comfortable not only when working with both the individual prices and computed indices, but also when measuring the efficiency of the collectors. Such queries, as for example the number of prices not changed in a given region of a given collector for a given period

of time, are answered within a few minutes and return precise identification of such a case.

12. Another facility, used usually to find quickly an extreme data in the database, is the SPSS package connected with Oracle Database. The power of this approach in control procedures rests mainly on the construction of immediate box-plots and scatter-plots to visualize implausible extremes and identify them straightforwardly.

13. Naturally, not all controlling mechanisms can be left to computers. That is why the controls on the spot in the outlets are made on a regular basis. The Czech Statistical Office staff visit collectors to ensure that the methodology of data collection is complied with. They concentrate their attention mainly to uncover discrepancies between the methodology and the collecting routine. They have to check whether the appropriate variety corresponds with its definition. Varieties are characterized by the brand or specific suppliers of goods or services, by weight or other parameter of quantity, including conversions to standard measurement units, by allowed variations of quality from a standard product. All these characteristics are precisely fixed in the collector's guide and the collectors are obliged to take them into account.