Distr.
GENERAL

CES/AC.68/2002/10
18 February 2002

Original: ENGLISH

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD)

CONFERENCE OF EUROPEAN STATISTICIANS

COMMISSION OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint ECE/Eurostat/OECD
Meeting on National Accounts
(Geneva, 24-26 April 2002)

OUTLIER DETECTION OF MULTIVARIATE EXTRAPOLATED ADMINISTRATIVE DATA IN BELGIAN NATIONAL ACCOUNTS

Invited Paper submitted by National Bank of Belgium*

Introduction

1. The aim of this paper is to present the main results of an outlier detection system used in Belgian national accounts with regard to the main variables of the production account of S11. The focus is on results, not on theoretical derivations.

2. The estimation of the variables output (P1), intermediate consumption (P2) and value added (B1g) of sector S11 relies for a considerable part (20% of total value added) on extrapolating partial administrative data. Because extrapolating data always carries some risk of erroneous extrapolation, a robust and preferably automated outlier detection system is needed. As the variables under review are clearly related, a multivariate analysis is preferred.

---

* Prepared by Mr. Michel Soudan, Department of Statistics, National Bank of Belgium.

3. The structure of the paper is straightforward. Firstly, some background is provided on the Belgian system of estimation after which the data under review is summarised using some basic statistics and a principal components analysis. Secondly, the use of the Mahalanobis-distance in outlier detection is explained. The last part is concerned with how this concept has been implemented in Belgian national accounts.

Background

4. Estimates of the variables P1 and P2 of S11 are obtained using both quasi-exhaustive administrative sources and surveys. The most important of these sources are: companies' annual accounts, VAT declarations, social contribution declarations and structural enterprise surveys.

5. All economic units are classified according to institutional sector, industry, region and category. Initial P1, P2 and B1g estimates for S11 are therefore estimated on a very detailed level: by industry (249 industries), district (NUTSIII, 44 districts) and "category" (10 categories).

6. Region is included as a classification variable in order to ensure maximal coherence between national and regional accounts. However, the inclusion of this classification variable since 2000 has led to a nineteenfold increase in the number of aggregates: from about 1 300 to over 25 000. Manual inspection is therefore no longer an option.

7. The category indicates the size of the company (large enterprises vs. small and medium sized enterprises) and the data that is available from administrative sources. For some categories, only information on value added or wages is available. For other categories information on P1, P2 and B1g is available.

8. If information is only available on one of the three main variables of the production account, extrapolation based on observations of a similar group of entreprises is used to arrive at estimates for all three variables. Similar means that entreprises are classified in the same sector, industry and region. They are only different with regard to category.

Data exploration

9. Outlier analysis should take into account two different kinds of outliers. Firstly, P1 and P2 should be approximately linearly related (on an aggregate level) and this relation ought to show some stability over time. Any changes should be gradual. Secondly, the evolution of aggregates should be within reasonable limits. Aggregates which show extremely high or low growth rates arouse suspicion.

10. To allow for such an analysis, both the variables P1 and P2 should be included. In order to include the dimension "time", the analysis should take into account variables from more than one year. To perform a multivariate analysis, the variables are to be grouped into vectors.

11. The statistical population under review is therefore the set of complete, strictly positive vectors which contain the following variables:

$X = (P1_t \;\; P2_t \;\; P1_{t-1} \;\; P2_{t-1})'$

12. For the period 1998-9 X becomes $(P1_{1999} \;\; P2_{1999} \;\; P1_{1998} \;\; P2_{1998})'$. B1g is not included as a variable as it can be derived immediately from P1 and P2. The population contains 23 432 vectors out of a possible 28 390 (83%). The difference between the two figures is explained by the fact that individual characteristics are not stable over time. For instance, certain combinations (sector, industry, region, category) may exist in one time-period but not in the other because of reclassifications, mergers and acquisitions or changes in availability of data. Also, two new categories were created in 1999 for which no information was available in 1998. This implies that a considerable number of aggregates still need to be checked manually, mainly through ratio checks.

13. Because of the huge range present in the variables, they are logtransformed. The smallest aggregate in the population is EUR 34, the largest is EUR 16,5 bln. All variables are expressed in log(BEF). As can be seen from figures 1a to 1d, the distribution of the transformed (and normalised) variables seems to be quasi-normal as a slight right-skewness is present. However, the Kolmogorov-Smirnov test rejects the hypothesis of normality for all four variables with a 1% margin of error.

14. Tables 1 and 2 reveal that the variables are very similar and highly correlated. A principal component analysis based on the correlation matrix shows that the cloud of points in 4-dimensional space can be easily summarised in 2-, if not 1-dimensional space (see table 3).

15. The associated eigenvectors in table 4 reveal that the structure of the cloud of points is straightforward. The first principal component is simply a size variable. The second principal component is a contrast between the two years under consideration. The other components are further contrasts. However, the amount of variation they explain is negligable.

16. It is clear that the logtransformed data form a cloud of points in 4-dimensional space which takes the shape of a very allongated ellipsoid with only very minor departures from its main axis.

Outlier detection method

17. To check for outliers, the Mahalanobis-distance is calculated for each observation. Conceptually, this can be separated into two steps. Firstly, the variables under review are

transformed using the Mahalanobis-transformation which standardises (mean=0, variance=1) and orthogonalises the original data (i.e. their intercorrelations become 0).

$y = C^{-1/2}.(x - x_{avg})$

where

x = original (log-transformed) data-vector
y = transformed data-vector
C = estimated covariance matrix
x_{avg} = estimated vector of averages

18. Secondly, the Mahalanobis-distance between the transformed data and the origin 0 is calculated using the following formula:

$D^2 = y'.y = \sum y_i^2 = (x - x_{avg})'.C^{-1}.(x - x_{avg})$

where D (= $+\sqrt{D^2}$) equals the Mahalanobis-distance.

19. Geometrically, observations with an equal Mahalanobis-distance lie on an ellipsoid. The Mahalanobis-distance is small for observations lying on or close to the principal axis of the ellipsoid. Observations further away from the principal axis have a much higher Mahalanobis-distance. Figure 2 illustrates this point in the two variable cases. The ellipsoid is formed by all observations with a Mahalanobis-distance equal to 1. The Mahalanobis-distance of the (fictional) outlier at (7,5;8,5) –meaning that P2 is 10 times higher than P1– equals 7,12.

Implementation

20. The ultimate aim of an outlier detection system is to single out suspicious observations so that they can be inspected manually. In order to do so, the total population of observations is divided into two according to average size of an observation. Observations of which the average value of the four variables is higher than BEF 1 bln (approx. EUR 25 mln) are labelled large observations. Other observations are labelled small observations.

21. There are 2 558 large and 20 874 small observations. The observations that need to be checked manually are those whose Mahalanobis-distance is in the 5% upper quantile of each group. For large observations, this corresponds to observations with a Mahalanobis-distance higher than 3. For small observations, the cut-off point is 3,8. Table 5 contains the frequency tables for the variables D, D_{large} and D_{small}.

22. An excerpt of a typical output table is given in table 6. Apart from the original data and the Mahalanobis-distance itself, transformed data are also included. As the squared Mahalanobis-

distance is simply the sum of the squared transformed variables, transformed data allow to determine which variable(s) of the four included in the vector is (are) responsible for the high deviation. Note that Y1 is the transformed variable P1_98, Y2 is the transformed P1_99 etc. Transformed variables are expressed in standard deviations.

23. For example, the high Mahalanobis-distance of the first observation is mainly due to the variables P1_98 and P2_98 (Y1 and Y3 resp.). Obviously, P1_98 is much too low given P2_98.

24. The output in table 6 was generated using a com-component imbedded in an excel-macro. This enables everyone in the Statistics Department who has access to excel to generate the output in table 6. Further options (covariance and correlation matrix, graphical output, ...) are gradually being implemented. Previously, a routine in SAS-IML was used, which limited accessibility.

Conclusions

25. Although all observations that have been selected for manual inspection in the above exercise show deviant behaviour, in a majority of cases, no correction is made after inspection. Large swings in aggregates from one year to another are mainly due to companies changing category rather than erroneous extrapolation. As a result of the experience gathered with the Mahalanobis-distance, the national accounts division is looking at different ways to keep the classification variable “category” more stable over time.

26. Experimentation with the above technique has shown that using the Mahalanobis-distance is very efficient in identifying outliers. It is a very versatile technique so its use is not limited to the above case. Other potential uses are comparing two or more variables which supposedly measure the same phenomenon (e.g. unemployment) or comparing two versions of certain variables to identify major revisions.

27. There are some limitations. Most importantly, variables included in the analysis should be symmetrically distributed and unimodal. Correlations between variables should be very high (preferably $|r|>90\%$). Relations between the variables ought to be linear, although this problem can be solved using a linearising transformation. Finally, it is recommended that the number of variables included in the analysis remain limited (<10) so as not to complicate interpretation of the transformed y-variables.

TABLES

Table 1. Basic statistics

	P1_99	P2_99	P1_98	P2_98
average	7,83	7,64	7,80	7,63
standard deviation	1,05	1,08	1,04	1,07
skewness	-0,22	-0,19	-0,19	-0,19
kurtosis	0,15	0,06	0,08	0,06

Table 2. Correlation matrix

	P1_99	P2_99	P1_98	P2_98
P1_99	100%			
P2_99	98,8%	100%		
P1_98	92,4%	91,6%	100%	
P2_98	91,7%	92,6%	98,8%	100%

Table 3. Principal components analysis

Components	Eigenvalues	Proportion	Cumulative
PRIN1	3,829	95,7%	95,7%
PRIN2	0,147	3,7%	99,4%
PRIN3	0,021	0,5%	99,9%
PRIN4	0,004	0,1%	100,0%

Table 4. Eigenvectors

	PRIN1	PRIN2	PRIN3	PRIN4
P1_99	0,500	-0,500	0,497	-0,503
P2_99	0,500	-0,500	-0,492	0,508
P1_98	0,500	0,505	0,504	0,492
P2_98	0,500	0,495	-0,508	-0,497

Table 5. Frequency table

Bin	D		D_{large}		D_{small}	
	Freq	cum%	Freq	cum%	Freq	cum%
0	0	0%	0	0%	0	0%
1	8676	37%	569	22%	8107	39%
2	10109	80%	1613	85%	8496	80%
3	2757	92%	240	95%	2517	92%
4	961	96%	100	99%	861	96%
5	409	98%	8	99%	401	98%
6	199	99%	14	99%	185	99%
7	121	99%	4	100%	117	99%
8	64	99%	10	100%	54	99%
9	40	100%	0	100%	40	100%
10 and more	96	100%	0	100%	96	100%
□	23432		2558		20874	

Table 6. Extract from an output table (first 10 lines)

P1_98	P1_99	P2_98	P2_99	lgP1_98	lgP1_99	lgP2_98	lgP2_99	Y1	Y2	Y3	Y4	Dist
5.000	115.492.000	10.001.000	121.220.000	3,70	8,06	7,00	8,08	-22,76	9,63	14,79	-3,93	29,07
17.312	1.527	5.330	818.299	4,24	3,18	3,73	5,91	5,76	-18,38	-10,36	16,03	27,12
20.998	12.389.988	87.354	49.274	4,32	7,09	4,94	4,69	-12,38	14,43	6,80	-13,46	24,27
295.365.000	185.706.000	30.417.000	241.000	8,47	8,27	7,48	5,38	-0,31	11,62	2,90	-14,50	18,81
43.112.339	162.874.745	20.907.989	589.571	7,63	8,21	7,32	5,77	-3,97	11,41	4,70	-12,86	18,26
27.863.551	5.350.146	117.642	2.799.390	7,45	6,73	5,07	6,45	10,76	-4,75	-12,58	4,24	17,73
300.828	809.454	1.392	430.287	5,48	5,91	3,14	5,63	8,32	-4,02	-14,28	4,96	17,72
905.405	13.127.648	188.538.380	39.897.241	5,96	7,12	8,28	7,60	-12,81	1,89	11,58	-1,86	17,47
54.000	7.000	2.299.000	2.463.000	4,73	3,85	6,36	6,39	-4,57	-13,22	3,26	9,50	17,22
17.569.618	1.249.355	15.723.584	15.563	7,24	6,10	7,20	4,19	-2,77	6,32	6,39	-12,75	15,85

Figure I.a
Histogram log(P1_98)

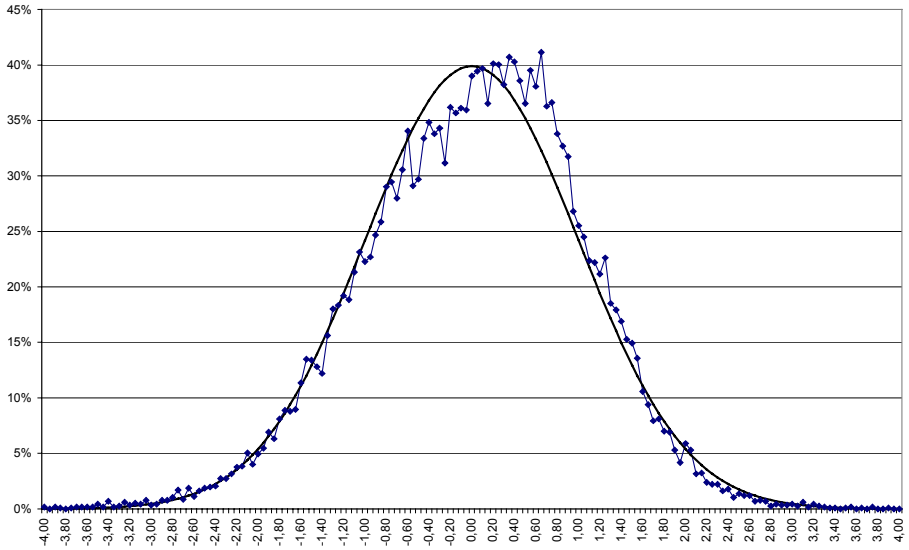

Figure I.b
Histogram log(P1_99)

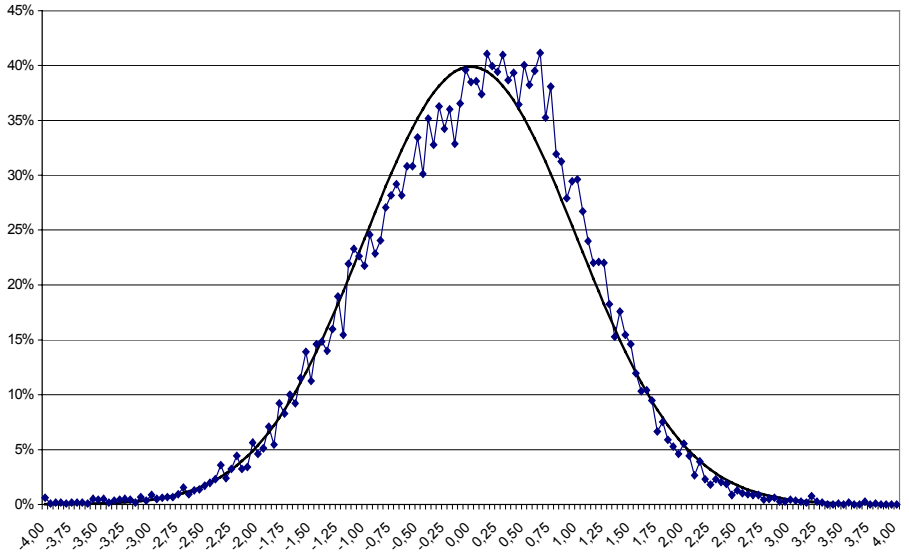

Figure I.c
Historgram log(P2_98)

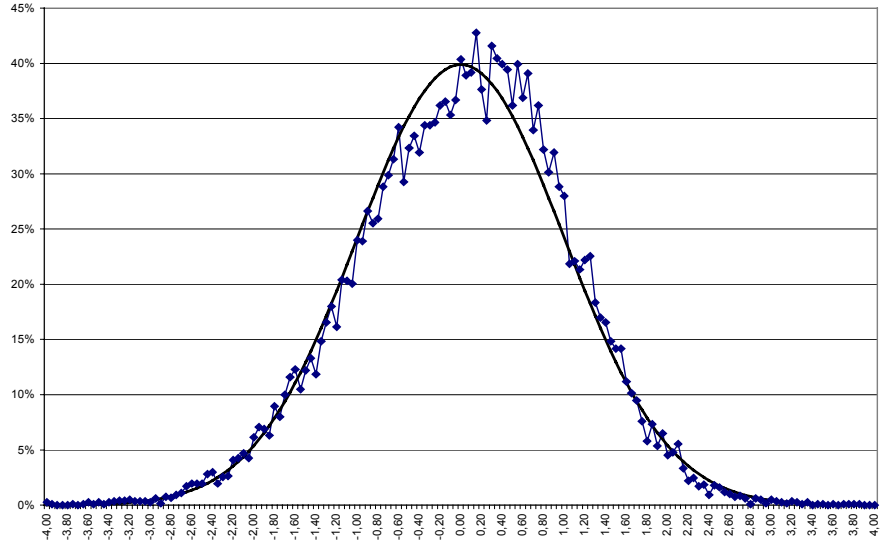

Figure I.c
Historgram log(P2_98)

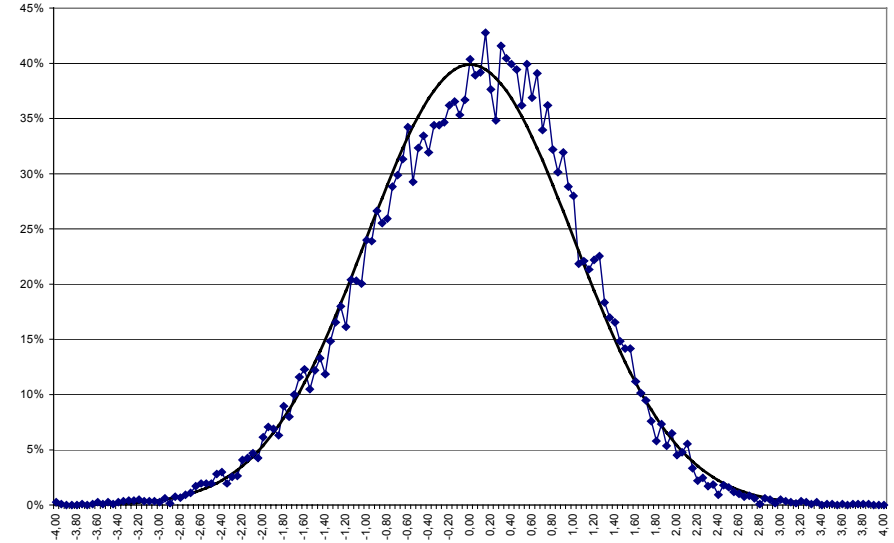

Figure 2: logP1_99 vs. logP2_99