WP. 6
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (i): Selective and macro editing

An automated comparison of statistics

Prepared by Elmar Wein, Federal Statistical Office, Germany

I. Introduction

1. Data editing on the basis of soft edit rules enables subject matter statisticians to accept suspicious values. As these decisions may lead to biased data a comparison of actual statistics with plausible reference statistics should serve as a final quality control at the end of the data editing process. In German practice experienced subject matter statisticians perform the comparison manually. If an actual statistic is suspect a search for the erroneous records starts which looks like "searching a needle in the hay" because there are no indicators of suspect values available.

2. The Federal Statistical Office of Germany modernises the processing of its enterprise statistics[i]. The recommendations as regards the data editing process contain the introduction of an automated comparison of actual statistics with plausible reference statistics, e.g. statistics of the previous period. The new approach was realised for the data editing process of the structural business surveys in the German domestic trade. The method shall detect outliers. A table with (flagged) outlying statistics and indicators of the respective (suspicious) records is the main result of this new method.

3. Firstly the contribution will describe the methodology and realisation of the multivariate comparison followed by the experience made with the structural business survey in the German wholesale trade 2010. A judgement of the new approach and an outlook as regards its advancement will be given at the end of this contribution.

II. Details of the automated comparison

A. Overview of the principal component analysis - excursion

record	X1	X2
1	0,5	0,7
2	1,0	1,1
3	1,1	0,9
4	1,5	1,6
5	1,9	2,2
6	2,0	1,6
7	2,2	2,9
8	2,3	2,7
9	2,5	2,4
10	3,1	3,0

	eigenvectors / loadings		eigenvalues	
	PC1	PC2	PC1	PC2
x	0,67787	-0,73518	1,13315	0,22155
y	0,73518	0,67787		
		(eigenvalues)^2	1,28403	0,04908
		totalvar	1,33311	

4. As the comparison uses the principal component analysis (PCA) for reducing the dimensional problem as regards detecting multivariate outliers this method will be explained at first. Let $\underline{X}$ be a dataset with the structure of an n×m matrix that means n records with m variables. The example on the left shows a dataset with ten records and two variables.[1]

[1] The majority of the computations were made with the German version of Excel 2010 which requires values with a comma instead of a dot. The PCA was performed with R.

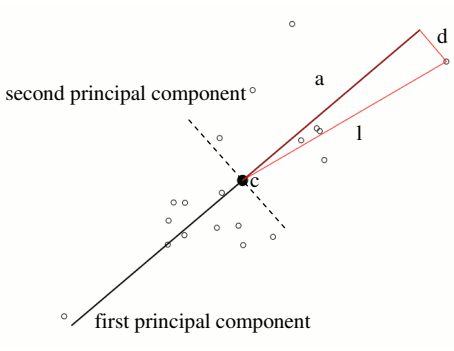
5. A measure of dispersion of a multivariate dataset is the *total variance V*. It is an enhancement of the ordinary variance and is defined as the sum of the squared Euclidean distances between records of a dataset and its mean vector:[ii]

$$V = \sum_{j=1}^{m} Var(X_j) = \sum_{j=1}^{m} \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - \bar{x}_j)^2 \qquad (1)$$

x_{ij} denotes the ith observation of variable j, $\bar{x}_j$ the arithmetic mean of the observations of variable j and $\sum_{j=1}^{m} (x_{ij} - \bar{x}_j)^2$ the squared Euclidean distance between the multivariate observation $\underline{X}_i = (x_{i1},...,x_{im})$ and the center of a multivariate dataset $\overline{\underline{X}} := (\bar{x}_1,...,\bar{x}_m)$. An important property of the total variance is its invariance that means a total variance of a rotated dataset is equal to the total variance of an original dataset. The total variance V of the example above is 1.33311:

variable	X1	X2
variance	0,61656	0,71656
total variance V		1,33311

6. Often the high dimension m of a multivariate dataset complicates the detection of multivariate outliers. Thus reducing the dimensionality to q components with q << m is a useful step. Methods for extracting these components are provided by the principal component analysis. The computation of principal components requires the transformation of a dataset into a set of values of linearly uncorrelated variables. The transformation is defined in such a way that the first principal component accounts for as much of the variability in the dataset as possible. Its computation is an iterative process because a line / vector is searched that minimizes the squared differences to all values that means the vector tries to describe the dotted points of the left picture in the best way.[iii] As the distance from the data center (c) is independent from the position of a dataset (line l) minimizing the squared difference (d) to the point shown in the picture and all other dotted points leads to the black line (first principal component). A minimization of the distance d at right angles to the straight line while maintaining the distance l to the data center thus means a maximization of the distances in the direction of the black lines (a = first principal component). This can be achieved by using and keeping the Pythagorean Theorem (here: $a^2 + d^2 = l^2$). The summed squares of the distances between the data center c and all other dotted points in the direction of the black line form the variance of the data in this direction. As the first principal component covers only a part of the scatter computing a second principal component will be the next step. This component corresponds to a line which forms a right angle to the first line in the case of a two-dimensional dataset that means the second component of a multivariate dataset is orthogonal and thus uncorrelated to the first one.[iv] It covers the second highest variance under the restriction of being uncorrelated to the first component. For an m-dimensional dataset m components can be computed which cover the total variance.



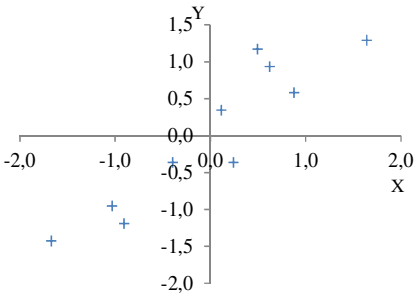
7. The components represent a new coordinate system which can be interpreted as a rotation of the axes of the variables m. As the computation of the principal components consists of minimizing squared distances identical scaling of the variables is one important precondition of the PCA. Additional preconditions of the PCA are the linearity of the variables which are normally distributed and have to be centered to their means. Centering is a necessary prerequisite of PCA to ensure that the first principal component describes the direction of maximum variance. If mean subtraction is not performed, the first principal component might instead correspond more or less to the mean of the data. Linearity is necessary because the idea is to describe the total variance of a dataset by linear combinations of the original observations and the principal components.[v]

8. The computation of principal components comprises the following steps:

(a) Prepare the dataset for the PCA

Let $\underline{X} = (X_{11},..., X_{nm})$ be an n×m matrix and $\underline{\overline{X}} = (\overline{X}_1,..., \overline{X}_m)$ the vector of the corresponding means.

X1	X2
-1,7	-1,4
-1,0	-1,0
-0,9	-1,2
-0,4	-0,4
0,1	0,3
0,2	-0,4
0,5	1,2
0,6	0,9
0,9	0,6
1,6	1,3



Then $\underline{X}^T = (X_{11}^T,..., X_{nm}^T)$ is a transformed matrix with empirical zero means and unit scale. It is recommended to standardise a dataset before performing a PCA to ensure that the analysis will not prefer the variable with the biggest variance. The picture on the left shows the standardized dataset.

(b) Compute the eigenvectors and eigenvalues of the correlation matrix of $\underline{X}^T$ [vi]

	X1	X2
X1	1,00000	0,92593
X2	0,92593	1,00000

Firstly calculate the correlation matrix $\underline{C}$ from the matrix $\underline{X}^T$ which is a symmetric positive definite m×m matrix with the variances in the diagonal. The transformation of a multidimensional dataset into a set of nearly uncorrelated variables is performed via an m×m matrix $\underline{E}$ of eigenvectors from the diagonal of the correlation matrix

E				D	
eigenvectors / loadings				eigenvalues	
	Y1	Y2		Y1	Y2
X1	0,70711	-0,70711		1,92593	0,07407
X2	0,70711	0,70711			
	total variance V			2,00000	

$\underline{C}$. This computation leads also to a matrix $\underline{D}$ with the corresponding eigenvalues.[2] Each eigenvalue is proportional to the portion of the sum of the squared distances of the points from their multidimensional mean. The PCA of the example leads to two principal components. The eigenvalue of the first component is significantly higher than the eigenvalue of the second one. The first principal component is calculated by the loadings (0.70711, 0.70711) and the total variance V of the eigenvalues is 2. This result indicates that the principal components cover the same variance than the original dataset with the original values.[vii]

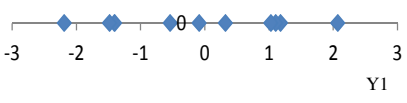
(c) Choose components and form a feature vector

	eigenvalues	portion of variance [%]
Y1	1,92593	96,3
Y2	0,07407	3,7

The component with the biggest eigenvalue explains the biggest part of the variability of $\underline{X}^T$ and thus will be the first principal component. If there are principal components with a marginal contribution to the total variance it is up to the statistician to decide upon their use. In many cases only the principal components with the biggest eigenvalues are used for further data processing because they cover a high proportion of the total variance of a dataset. This decision assumes that the remaining x % of the total variance represent white noise and thus does not contribute to the *natural* variability of a multivariate dataset. The choice of the respective eigenvectors leads to the matrix / vector $\underline{L}$ which is not shown.

Y1
-2,19
-1,41
-1,48
-0,54
0,32
-0,09
1,18
1,10
1,03
2,07

(d) Derive the new dataset of principal components



A new dataset with principal components $\underline{Y}$ is computed by multiplying the eigenvectors / loadings of the chosen principal components of the matrix $\underline{L}$ with the matrix $\underline{X}$. Referring to the example the vector Y consists of the eigenvectors of the first principal component (Y1).

[2] The eigenvectors of a square matrix are the non-zero vectors that, after being multiplied by the matrix, remain parallel to the original vector. For each eigenvector, the corresponding eigenvalue is the factor by which the eigenvector is scaled when multiplied by the matrix. An important property of the eigenvectors is that they are orthogonal. For further explanations of eigenvectors and eigenvalues see "en.wikipedia.org/wiki/Eigenvector".

B. Methodology of the automated comparison

9. The automated comparison consists of six steps:
   (a) Step 1: After checking the parameter set for the comparison survey specific transformations of the data to be compared (actual dataset) may take place.
   (b) Step 2: A robust dataset is computed from the actual dataset by replacing suspicious univariate outliers with permissible values (winsorising).
   (c) Step 3: The matrix of the loadings is computed on the basis of the robust dataset and principal components are computed on the basis of the robust eigenvectors and the actual dataset.
   (d) Step 4: For the detection of outliers the Euclidean distance is calculated on the basis of the principal components per record and a Nalimov test is applied to the Euclidean distances of the records which belong to a statistic. The steps 1 to 4 are also applied to a reference dataset.
   (e) Step 5: The principal components and scores of the actual and reference dataset are (weighted and) added up to statistics. Differences are computed between actual and reference statistics and outliers are flagged on the basis of the Nalimov test.
   (f) Step 6: A table with the (flagged) statistics on the left and the respective robust principal components and (flagged) scores of the actual dataset including the identifiers of the records on the right is produced. It is the result of the automated comparison.

10. Step 1: Optional survey specific transformations

As the actual dataset will be compared with a plausible reference dataset – preferably the data of the last reporting year – structural developments like the (weighted) number of enterprises should be eliminated because they complicate the comparison. Another suitable transformation may be in the case of structural business statistics to divide turnover and costs of an enterprise by its number of working places because this transformation compensate different sizes of enterprises among the data to be checked.
Let $\underline{X}$ be a dataset with the structure of an n×m-matrix that means n enterprises with m variables then $\underline{X}^T$ represents the transformed dataset. The following example shows the original dataset on the left part. It consists of 10 records which belong to one statistic, e.g. of an economic branch. The records are weighted, records 01 and 10 contain outliers. The original dataset is transformed by adjusting the weights to the value one and the variables “turnover” and “cost trade goods” of every record are divided by the number of working places. The transformed dataset $\underline{X}^T$ is presented on the right part:

	Actual original dataset $\underline{X}$					Actual transformed dataset $\underline{X}^T$			
ID	weight	working places	turnover	cost trade goods	ID	weight	turnover	cost trade goods
01	10	1	45794	65	1	0,2222	45794	65
02	8	1	58918	30553	2	0,1778	58918	30553
03	6	1	60829	36621	3	0,1333	60829	36621
04	5	1	63685	37013	4	0,1111	63685	37013
05	5	1	64505	36960	5	0,1111	64505	36960
06	4	1	65892	35667	6	0,0889	65892	35667
07	3	1	68227	35621	7	0,0667	68227	35621
08	2	1	71355	38714	8	0,0444	71355	38714
09	1	2	84394	55941	9	0,0222	42197	27970
10	1	2	1247730	66973	10	0,0222	623865	33486
sum	45					1,000		

11. Step 2: Create a robust dataset

The purpose of this step is to obtain a robust dataset so that outliers are not masked as regards their detection. A robust method based on the first and third quantile subtracted / added by the 3.5 times the interquartile range is used to detect as much suspicious values as possible.[viii] They are replaced by a winsorising method with the last permissible values. The winsorising approach is used because it ensures that all the records of a dataset remain available for further analysis.[ix]
In the example the suspicious turnover of enterprise 10 was replaced by the upper bound and the suspicious costs for trading goods of enterprise 1 by the lower bound (right part of the following table):

Actual transformed dataset $\underline{X}^T$

ID	weight	turnover	cost trade goods
01	0,222	45794	65
02	0,178	58918	30553
03	0,133	60829	36621
04	0,111	63685	37013
05	0,111	64505	36960
06	0,089	65892	35667
07	0,067	68227	35621
08	0,044	71355	38714
09	0,022	42197	27971
10	0,022	623865	33487
Q1		59396	31286
Q3		67643	36875
IQR		8248	5589
LowBound		30530	11725
UpBound		96510	56436

Actual robust dataset $\underline{X}^T$

ID	weight	turnover	cost trade goods
01	0,222	45794	11725
02	0,178	58918	30553
03	0,133	60829	36621
04	0,111	63685	37013
05	0,111	64505	36960
06	0,089	65892	35667
07	0,067	68227	35621
08	0,044	71355	38714
09	0,022	42197	27971
10	0,022	96510	33487
mean		63791	32433
stdev		14814	7975

12. Step 3: Perform a robust PCA and compute principal components

Firstly the robust dataset is standardized as described in section 8a and the correlation matrix is computed as shown in the following table:

Robust standardised dataset $\underline{X}^T$

ID	weight	turnover	cost trade goods
01	0,222	-1,215	-2,596
02	0,178	-0,329	-0,236
03	0,133	-0,200	0,525
04	0,111	-0,007	0,574
05	0,111	0,048	0,568
06	0,089	0,142	0,406
07	0,067	0,299	0,400
08	0,044	0,511	0,788
09	0,022	-1,458	-0,560
10	0,022	2,209	0,132

Robust PCA

	eigenvalues	proportion of V
Y1	1,53744	76,9
Y2	0,46256	23,1

Loadings

	turnover	cost trade goods
Y1	0,70711	0,70711
Y2	0,70711	-0,70711

correlation matrix

	turnover	cost trade goods
turnover	1,00	0,54
cost trade goods	0,54	1,00

The matrix of the loadings is computed on the basis of the robust standardized dataset and documented on the right part of the table above. The first principal component accounts for nearly 77 per cent of the total variance. As the second principal component accounts for a relatively big proportion of the total variance it will be used for the error detection too.

At the end of this step the principal components are computed on the basis of the original dataset and the robust loadings. The principal components represent the new dataset $\underline{Y}$:

Actual transformed dataset $\underline{X}^T$				Prinicipal Components $\underline{Y}$			
ID	weight	turnover	cost trade goods	ID	weight	Y1	Y2
01	0,222	45794	65	01	0,222	32382	32336
02	0,178	58918	30553	02	0,178	41662	20057
03	0,133	60829	36621	03	0,133	43013	17118
04	0,111	63685	37013	04	0,111	45032	18860
05	0,111	64505	36960	05	0,111	45612	19477
06	0,089	65892	35667	06	0,089	46592	21372
07	0,067	68227	35621	07	0,067	48244	23056
08	0,044	71355	38714	08	0,044	50456	23081
09	0,022	42197	27970	09	0,022	29838	10060
10	0,022	623865	33486	10	0,022	441139	417461

Loadings		
	turnover	cost trade goods
PC1	0,70711	0,70711
PC2	0,70711	-0,70711

Y of robust mean $\underline{C}$	68040,9	22173,5

13. Step 4: Compute the Euclidean distances per record and flag outliers

To determine multivariate outliers the Euclidean distances (ED) are computed for each record. For this step the mean vector of the robust dataset is used, which is transformed by the loadings to the vector $\overline{\underline{Y}}$. Then the sum of the Euclidean distances on the basis of principal components is an error indicator per record. It is defined as:

$$d = \sum_{k=1}^{m} \sqrt{(y_k - \bar{y}_k)^2} \qquad (2)$$

Multiplying the error indicator d with the weight of a record leads to the relevance indicator r:

Prinicipal Components $\underline{Y}$				Euclidean Distances, error indicator d, and relevance indicator r						
ID	weight	Y1	Y2	ID	weight	ED1	ED2	d	r	significance of d
01	0,222	32427	32336	01	0,222	35614	10162	45776	10172	
02	0,178	63266	20057	02	0,178	4775	2116	6891	1225	
03	0,133	68908	17118	03	0,133	867	5056	5922	790	
04	0,111	71204	18860	04	0,111	3163	3314	6477	720	
05	0,111	71747	19477	05	0,111	3706	2696	6402	711	
06	0,089	71813	21372	06	0,089	3772	802	4574	407	
07	0,067	73432	23056	07	0,067	5391	883	6274	418	
08	0,044	77831	23081	08	0,044	9790	907	10697	475	
09	0,022	49616	10060	09	0,022	18425	12114	30539	679	
10	0,022	464818	417461	10	0,022	396777	395287	792064	17601	*

Y of robust mean	68041	22173

Nalimov test	mean	91562
for d = 792064	variance	246509
	test statistic	3,00
	crit. value	1,56

Finally the Nalimov Test is applied to the error indicator d to flag records with outlying distances.[x] The example shows that record 10 possess the biggest error indicator which was flagged by the Nalimov test. Record 1 possesses the second highest but insignificant error indicator and record 9 received also a high error indicator. The reason may be the high distance of the record from the mean which was caused by dividing the values through the working places. According to the relevance indicator r the records 10 and 1 keep their priorities as regards a deeper review opposed to record 9.

The table shows that the priority setting among the records based on ED1 / Y1 is identical with the setting based on the final error indicator d. As Y1 accounts only for nearly 77 per cent of the total variance this fact indicates that principal components with marginal contributions to the total variance can be neglected for outlier detection.

The steps 1 to 4 are also applied to the plausible reference dataset with the exception that there is no labelling of records with outlying Euclidean distances.

14. Step 5: Aggregate principal components to statistics, compare statistics and flag outlying differences

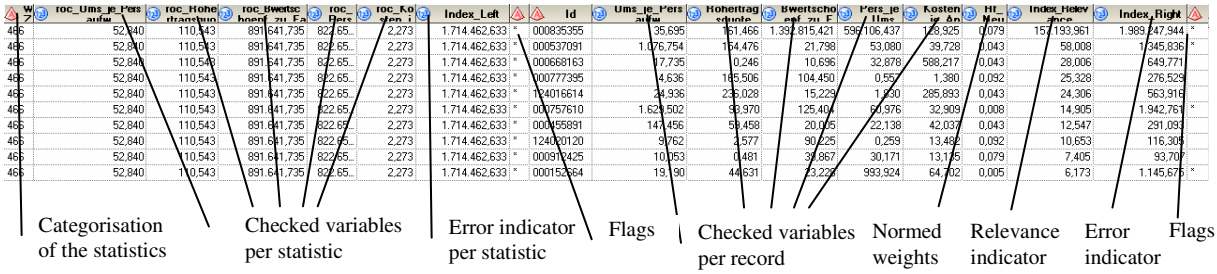
The principal components for the actual dataset y_k^a and the plausible reference dataset y_k^r are weighted and summarized to statistics. Secondly the squared differences between the results are computed per component in per cent of the plausible reference results and summed to one indicator. Let w_l be a weight of a record l an indicator per statistic i_s is defined as:

$$i_s = \sum_{k=1}^{m} \left(\frac{\sum_{l=1}^{n} w_l y_k^a - \sum_{g=1}^{h} w_g y_k^r}{\sum_{g=1}^{h} w_g y_k^r} \right)^2 \quad (3)$$

At last the Nalimov Test is performed among all indicators i_s to flag statistics with outlying differences.

15. Step 6: Produce the table of the results

The result of the automated comparison shall inform subject matter statisticians about the priority setting as regards manual reviews and possible errors in the actual dataset. Consequently it was decided to produce a table with information about the deviations per statistic on the left and the corresponding records on the left. The table is sorted downwards on the basis of the indicators i_s and for each statistics the respective records are sorted downwards depending on the relevance indicator r.
Information about errors is provided by the squared Euclidean distance to the (weighted) mean per variable that was checked – for the statistics as well as for the contributing records. The following picture shows the result of the automated comparison in practice:



The right column of the table above shows the flagged records that need to be checked. Some of them have a significant error indicator but a low relevance indicator. Consequently the subject matter statistician can decide to check all flagged records or only flagged records with a high relevance indicator.

C. The automated comparison in practice

16. Pascal Avieny realised a prototype of an automated comparison by a collection of macros with the base system of SAS 9.2 and the package stats. The macros require only a unique record identifier, numerical variables to be compared, an actual dataset and a reference dataset as SAS files with identical variables, and a configuration file in English. The automated comparison can be extended for applying survey specific transformations. The code of the macros is documented in English and the processing of the automated comparison is documented in an English log file. The current version of the macros is considered as a prototype.

17. The automated comparison was used for the first time in June 2012 for comparing the structural business statistics 2010 in the German domestic trade with the reference data of 2009. The structural business statistics deliver information for each member state of Germany classified by the NACE level 3.[3] They are based on a survey with 47 variables which belong to the categories “personnel”, “working

[3] NACE: Nomenclature generale des Activites economiques dans les Communautes europeennes - Classification of Economic Activities in the European Community

places"; "expenses", "revenues", and "groups of products". The variables are checked by the traditional data editing processing which consists of around 80 checks and approximately 30 signals. Besides the collected variables aggregates on record level like the gross operating surplus are also checked - but only with signals. The domestic trade of Germany is characterised by a concentration of enterprises. This fact complicates a detection of errors by measuring the distance of a value from its mean. Consequently relations between variables were computed per record and distances were measured among the relations of the records.

18. Based on subject matter experience as regards the presence of suspect data it was decided to restrict the automated comparison on the variable turnover, gross profit, costs for personnel employment per employee, value added minus gross profit and working places. To facilitate the detection of outliers in the presence of big plausible value the variables were divided by the working persons. This step is only meaningful if there is a significant relationship between revenues, expenses and working persons. The comparisons were made for the member states of Germany and NACE level 3. The decision as regards the variables was tested by comparisons between raw and plausible data of 2010. The signals obtained by the comparison were checked by manual inspections of the corresponding erroneous records. The test turned out that the working persons should be used instead of employees in fulltime equivalents – this variable was used at first. The test indicated also that a relation between turnover and working persons may lead to misleading errors but it was not clear the amount of wrong signals. So it was decided to test this relation. As there was no experience with the comparison available it was decided to use always all principal components.

19. The subject matter statisticians had to export the actual and reference data out of the new software used for data editing and perform the comparison with SAS. The comparison took place at the end of the traditional data editing process and served as a final check. As there was a significant time delay it was recommended to correct at least records with flagged Euclidean distances that belong to flagged statistics. If there were enough time available all flagged actual records should be checked because an anonymous dataset has to be disseminated.

20. Statistical offices of six German states and the Federal Statistical Office tested the prototype of the automated comparison. The test with the survey in wholesale trade revealed that 945 of approximately 12,900 records were flagged. 122 of the 945 records with the biggest indicators were checked manually. The reviews revealed that 71 flagged records could be accepted from a subject matter point of view. Around 50 % of these flags were set due to a missing relationship between working persons and the turnover. Although it was known that some wholesale trading companies each of them with one or two working persons deal with goods with a value of billion Euros the rather big number of flagged enterprises was a little bit surprising. After the removal of the relations between the variables and the working persons the number of the flagged records decreased by nearly 700.
The suspicious records which were flagged correctly contained signals related to the suspicious variables but these values were accepted wrongly by the subject matter statisticians. In these cases the automated comparison was a valuable extension of the traditional data editing process.

21. The test by one statistical office of a German state revealed that the comparison did not detect a change in the area of hotels which was caused by a big enterprise with an erroneous big weight. The enterprise reported a wrong turnover in 2009 and consequently obtained a big weight of 50 by the following annual sample rotation. In 2010 it reported a turnover of 7 million Euro instead of 300,000 Euro for the year before. The automated comparison failed because the methodology doesn't take account for this type of error. To solve this problem an indicator was developed on the basis of the ranks of the variable turnover and the weight. In general enterprises with big turnovers have weights around 1 and enterprises with small turnovers weights up to 70. Consequently the difference of the rank of the turnover and the rank of the weight is different from zero. If an enterprise with a big turnover has a big erroneous weight the difference of the ranks will be around zero and thus far away from the mean of this indicator.

III. Conclusions

22. An automated comparison tries to detect and flag suspicious differences between statistics of an actual edited period and a plausible reference period. In addition to this it provides hints on possible

causes of the differences – especially by establishing relations between the differences on the level of the statistics and the corresponding records ("drill-down-approach"). In principal an automated comparison may be more powerful than the traditional manual check because structural differences can be eliminated as well as specific properties of variables like the location and variance. In addition to this the automated comparison disburdens statisticians from a strenuous work that requires solid and extensive subject matter experience.

23. The Federal Statistical Office of Germany realised a prototype of an automated comparison on the SAS platform that first computes a dataset without univariate outliers. It is used for performing a principal component analysis. The eigenvectors of the PCA are used to compute principal components on the basis of the original dataset. Outliers are detected by using the Euclidean distances among principal components. Finally statistical significant outlying distances are flagged on the basis of the Nalimov test. This processing needs some improvements. First the removal of univariate outliers doesn't lead to robust covariances. They may be estimated on the basis of the MCD-algorithm. To facilitate the detection of outliers the diagnostics offered by the PCA should be used to reduce the dimensionality of a dataset.

24. The requirements of the automated comparison are relatively low: a unique record identifier and two datasets that contain identical numerical variables. The low requirements, the opportunity to expand the macros by survey specific modules, and the chosen methodology which enables to reduce the dimensionality problem in the context of detecting multivariate outliers creates better preconditions for using the automated comparison as a generalised tool. It is expected that the comparison can be used at least for all structural business statistics.

25. The test of the prototype by the structural business statistics in domestic trade revealed that the adequate choice of the variables to be compared clearly influences the hit rate of the comparison and provide helpful information as regards its optimisation.

26. The comparison detected suspicious records that were flagged by signals which were rejected by subject matter statisticians. From that point of view the comparison can be considered as a useful addition to the traditional data editing process. The first positive experience indicates that an optimized version would be a valuable tool for German structural business statistics. In spite of the relative low hit rate the respective subject matter statistician pleads for the advancement of the tool.

27. Anonymous datasets of the structural business statistics are disseminated and on the other hand there are only traditional data editing procedures available. These preconditions require a correction of all records. Consequently the prototype has to support the final check. If automated error detection and correction methods are available it is planned to use it as a macro editing tool already from the beginning of a data editing process.

---

[i] Federal Statistical Office: "Reformation of enterprise statistics", Final report, p. 58, Wiesbaden March 2012
[ii] Wikipedia, total variance: de.wikipedia.org/wiki/Totale_Varianz
[iii] Wikipedia, PCA: de.wikipedia.org/wiki/Hauptkomponentenanalyse
[iv] Wikipedia, PCA: en.wikipedia.org/wiki/Principal_component_analysis.
[v] Jonathon Shlens: A Tutorial on Principal Component Analysis, snl.salk.edu/~shlens/pca.pdf.
[vi] The explanation of the PCA in this contribution focusses on the eigenvalue decomposition. An alternative approach for performing a PCA is the singular value decomposition. For further details please see "en.wikipedia.org/wiki/Singular_value_decomposition".
[vii] It can be proofed that the principal components describe the complete variability of a dataset (see http://www.statoek.wiso.uni-goettingen.de/veranstaltungen/Multivariate/Daten/mvsec4.pdf).
[viii] Wikipedia, outlier: en.wikipedia.org/wiki/Outlier#Identifying_outliers; a robust method is used because the aim of this step is to obtain rough estimates for the mean and the variance.
[ix] Wikipedia, winsorising: en.wikipedia.org/wiki/Winsorising
[x] Wikipedia, Grubbs' test for outliers: en.wikipedia.org/wiki/Grubbs%27_test_for_outliers. The Nalimov test requires data which are approximately normal distributed. It computes a test statistic for a potential outlier which is the absolute difference to the mean divided by the variance of the sample and compares it with a critical value which is approximately T-distributed. A value is flagged as outlier if its test statistic is bigger than the critical value. The Nalimov test is a slight modification of Grubbs' test for outliers because the test statistic is corrected by the sample size. The modification shall improve the performance of the test in the case of a sample size $n < 30$.