

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (i): Automated editing and imputation and software applications

AUTOMATIC IMPUTATION FOR SHORT TERM STATISTICS

Invited Paper

Prepared by Elmar Wein, Federal Statistical Office, Germany⁽¹⁾

I. INTRODUCTION

1. The European regulation 1165/98 of short term statistics imposes high demands on the timeliness of statistics in retail trade because first results have to be disseminated only 30 days after a month under observation. As a consequence of this fact there is only a limited period for enterprises to report monthly data on turnover and persons employed.
2. In general German retailers transmit their data to the statistical offices when they fill out the forms of the tax authorities. During the last years more and more small enterprises were permitted to report only at the end of a quarter or after 6 months. So especially smaller enterprises do not meet more and more the deadlines and thus the respective statistics suffer from unit non response. In the case of short term statistics in retail trade between 30% and 40% of the amount of turnover has to be estimated for first results.
3. Missing values for turnover and employees are imputed by two different methods. The current imputation method for turnover causes on average an absolute revision of the month to month of the last year change rate of around 0.9 percent points if the first result is compared with the one after 6 months. This revision is unacceptable because the month to month change rate of retail turnover lay between -2 and 4 percent and it was often not far away from 0. During the last 3 years changes of sign of the change rates took place and affected the statistics' reliability.
4. The revisions of the monthly index on retail trade were criticised by national and European users because the index belongs to the Primary European Economic Indicators. Providing reliable statistics is an important goal of the German Federal Statistical Office. As a consequence the unit in charge of short term statistics in domestic trade began with the development of efficient estimation methods for turnover and persons employed in 2006. The advancement was funded by Eurostat and the results are documented in an English report on the CIRCA server.²
5. The aim of this contribution is to describe the new imputation approach in the next chapter and continue with a presentation of important results derived from simulations. At the end some considerations will be provided as regards its implementation and possible advancements.

⁽¹⁾ elmar.wein@destatis.de; the author thanks Sascha Kless for developing two imputation methods.

² Document "Reducing the need for revisions in German retail trade statistics by means of improved estimations" on http://circa.europa.eu/Members/irc/dsis/ebt/library?l=/task_forces/tf_retail_trade/retail_quality_2009/meetings/1_erste_sitzung/documents&vm=detailed&sb=Title; the author thanks Eurostat for funding the advancement.

II. A NEW IMPUTATION APPROACH FOR GERMAN SHORT TERM STATISTICS ON RETAIL TRADE

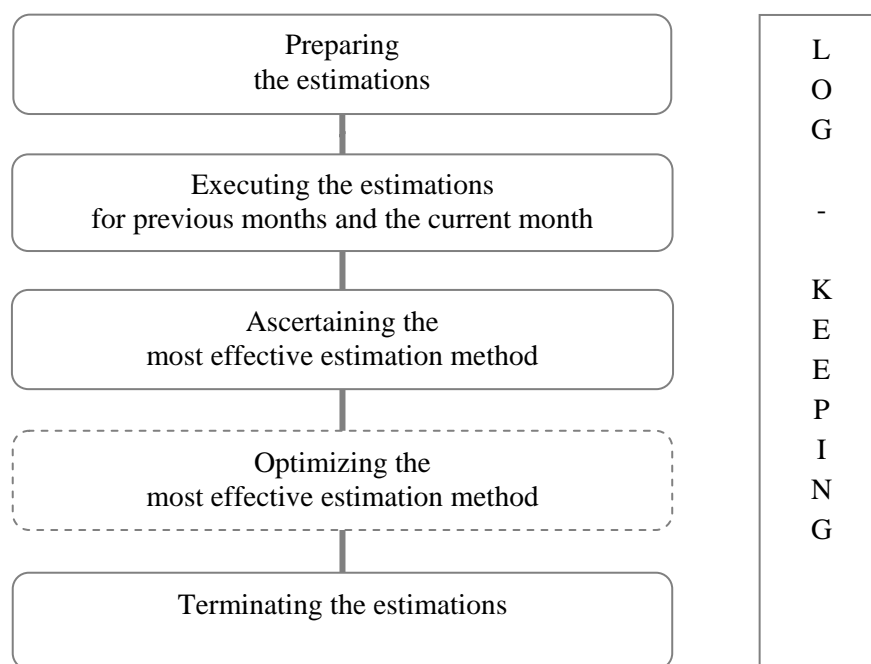
A. Preconditions for advancing the imputations

6. The advancement of the imputation methods have to be made under the following general conditions:

- Improvements have to be achieved in a medium-term that means that there is no time for a fundamental modernisation of statistics' processes.
- The demands as regards the timeliness can not be changed.
- The existing data processing environment has to be utilized. This means:
 - There are only 24 data for an enterprise for estimations available.
 - More powerful imputation methods are not available because of an old fashioned data base system with missing statistical functions. Only simple univariate descriptive statistics can be calculated.
 - Due to the in part small extent of random samples in some Länder, the number of the donor data sets is limited so that the estimations must be based on the NACE four-digit level.
 - Because of the time pressure, the estimations must, as always, be made fully automatically. Manual intervention should be restricted to the process control.
 - The new estimation methods must ensure reliable estimations even if only few enterprises reported just in time and data are very volatile.

B. Algorithm of the new imputation approach

7. The estimation for missing turnover and employees of enterprises with available data for former months consists of the following phases:



Preparing the estimations

- Sort the data sets of active enterprises by the *Land*, NACE four-digit number and identifying number of an enterprise in ascending order.
- Ascertain data sets with missing data for turnover and document them in a log file.
- Calculate and store the coefficient of variation for each data set to be estimated.

Executing the estimations

- Estimations are carried out for each data set with a missing value in the current reporting month using different estimation methods for available months and for the current reporting month.
- The estimations are stored in a temporary database table so that they can be compared with the available data of the retrograde months.

Ascertaining the most effective estimation method

- For each data set with a missing value the differences between the estimated values and the available data are calculated for retrograde months.
- The most effective estimation method in the past is that with the lowest sum of absolute differences.
- The estimated value of the optimal method is noted in the temporary data basis table, and the identifying number of the estimation method is placed in the estimation log, so that the estimations can be reconstructed in the course of a possible error search.

Adjusting the most effective estimation method

- The differences from the estimated values of the most effective estimation method and the available data of the retrograde months are compiled to form the mean estimated difference.
- The estimated value for the current reporting month is adjusted by the mean estimated difference if the mean is smaller than the half of the mean estimated value.
- The adjustment is documented in the estimation log.

Terminating the estimations

- Add the estimated value per enterprise to the current data stock.
- Document the termination of the estimation per data set in the estimation log.
- Delete temporary database tables.

C. Methods used for imputing

8. As mentioned above the algorithm performs test estimations for retrograde months on the basis of several methods. In detail 8 different estimate methods were developed - 7 methods for turnover and persons employed and 1 method only for persons employed. The methods used for imputing turnover can be classified in methods for enterprises with a ...

- seasonal turnover development,
- turnover development influenced by the preceding month and
- turnover development at random.

9. For enterprises with a seasonal turnover development the existing method was used:

Abbreviation	M30
Designation	Same month of the previous year + trend of the enterprise
Brief description	Turnover of the same month of the previous year u_{m-12} (base value of the estimation) is updated with a trend component $\frac{u_{m-3} + u_{m-2} + u_{m-1}}{u_{m-12-3} + u_{m-12-2} + u_{m-12-1}}$ from three previous months and three months of the previous year of the relevant enterprise. The method is suited for enterprises whose turnover exhibits seasonal patterns.
Formula	$\hat{U}_m = \frac{u_{m-3} + u_{m-2} + u_{m-1}}{u_{m-12-3} + u_{m-12-2} + u_{m-12-1}} \cdot u_{m-12}$

If

$$u_{m-12} < \text{median}(u_{m-12-1}, u_{m-12}, u_{m-11}) - \text{stddev}(u_{m-12-1}, u_{m-12}, u_{m-11}) \text{ or}$$

$$u_{m-12} > \text{median}(u_{m-12-1}, u_{m-12}, u_{m-11}) + \text{stddev}(u_{m-12-1}, u_{m-12}, u_{m-11})$$

then

$$U_{m-12}^{glan} = 0,25 \cdot u_{m-12-1} + 0,5 \cdot u_{m-12} + 0,25 \cdot u_{m-11}$$

Prerequisites for use	$u_{m-12} \neq 0 \wedge u_{m-1} \neq 0$
Procedure	Availability of data from 24 preceding months. <ul style="list-style-type: none"> • Outlier trend/seasonal components are recognized and corrected using the following check: If trend > 4 then trend = 4 • Round the result.

10. Besides the method M30 4 other versions were developed. Three of them are characterised by specific trend components. M10 adds the current season component from a similar enterprise to the enterprise's specific previous months and thus follows the idea of the nearest neighbor approach. The trend component of M20 consists of 4 months, an estimate of the current month and 3 previous months with the corresponding months of the preceding year. Method M60 uses the current observed trend from other enterprises and the enterprise's turnover from the corresponding month of the previous year. Opposed to the previously described methods M80 possess an adapted turnover value that takes into account the difference of the number of working days between the current month and the corresponding one of the preceding year.

11. For enterprises with a turnover development influenced by the previous month the following method was developed:

Abbreviation	M40
Designation	Previous month + current information
Brief description	Estimation with the previous month The method imputes partial autocorrelations between the turnover of a month and of the previous month as well as the corresponding months of the previous year. It is suitable for enterprises that estimate the turnover in the current month with the data from the previous month.

Formula

$$\hat{U}_m = \hat{u}_s \cdot u_{m-1} \text{ with}$$

$$\hat{U}_s = \left(1 + \frac{u_{m-12} - u_{m-12-1}}{u_{m-12-1}} \right)$$

and

general limitation of the seasonal component:

$$0,25 \leq \hat{U}_s \leq 4$$

Prerequisites for use	<ul style="list-style-type: none"> • $u_{m-1} \neq 0 \wedge u_{m-12-1} \neq 0$ • Availability of data from 14 preceding months.
Procedure	<ul style="list-style-type: none"> • If coefficient of variation ≥ 0.6: <p>with $\hat{U}_s = 1 + 0,25 \cdot \left(\frac{u_{m-13} - u_{m-14}}{u_{m-14}} \right) + 0,5 \cdot \left(\frac{u_{m-12} - u_{m-13}}{u_{m-13}} \right) + 0,25 \cdot \left(\frac{u_{m-11} - u_{m-12}}{u_{m-12}} \right)$</p> <ul style="list-style-type: none"> • Round the result.

12. For enterprises with a turnover development at random the following method was developed:

Abbreviation	M70
Designation	Median/mean from the available reports of the previous months
Brief description	Estimation using the median of turnover from one or more previous months. The method is suitable for enterprises whose turnover does not exhibits seasonal patterns. Obtaining the mean size of the enterprise's turnover the mean/median of the previous month is used.

Formula	$\hat{U}_{i,m} = \tilde{U}_{i,m}, m = m - 1, \dots, m - 12$
	with
	i ... enterprise to be estimated
	m ... period (month)
Prerequisites for use	At least 1 previous month's report must be available.
Procedure	<ul style="list-style-type: none"> • The median is used when the coefficient of variation ≥ 0.6 or number of available months < 3 <li style="padding-left: 20px;">The mean is used when the coefficient of variation < 0.6 and number of months > 3 • Round the result.

13. For estimating missing values on employees the following method was developed:

Abbreviation	M90
Designation	Previous month
Brief description	Estimating employees (b) with data of the previous month The method assumes partial autocorrelations between the number of employees of an current and preceding month. The method is applicable for enterprises with constant numbers of employees.
Formula	$\hat{B}_m = b_{m-1}$
Prerequisites for use	Availability of data from five preceding months
Procedure	<ul style="list-style-type: none"> • If $\left(\tilde{b}_{m-5,\dots,m-1} + \sigma_{m-5,\dots,m-1}^b\right) < b_{m-1} < \left(\tilde{b}_{m-5,\dots,m-1} - \sigma_{m-5,\dots,m-1}^b\right)$ then $\hat{B}_m = 0,5 \cdot b_{m-1} + 0,3 \cdot b_{m-2} + 0,2 \cdot b_{m-3}$ • Round the result.

III. SIMULATING THE NEW IMPUTATION APPROACH

A. The concept of the simulations

14. The new imputation approach was extensively tested for domestic trade (retail and wholesale), hotels and restaurant industry on the basis of data from 25 months because the approach will be used for these four NACE sections. To evaluate the imputation methods, the retail trade data from five Länder with the biggest amount of turnover as per December 2006 were used. This made it possible to assess the effects of more effective estimation methods on the first results of the retail trade statistics. The estimation methods were tested for the months January until December 2006.

15. With regard to the evaluation of estimation methods, in the retail trade the data sets of those enterprises were chosen that reported over all months. The enterprises' data were extrapolated in order to take into account the differing influence of the enterprises on the results. Since the applicable enterprises attained higher turnover a random sample was taken from this subset of around 2,000 data sets in order to gain a similar turnover distribution as for the enterprises to be estimated in practice.

16. First, turnover was estimated for the available reports using the existing method. From the estimations and reports, the estimation errors were formed as differences, weighted and added to the turnover of the enterprises to be estimated, divided by all turnover of a reporting month and multiplied by a value of 100. This produced the weighted average estimation error in percent. It shows the extent to which under- and overestimations can offset one another. Weighting with the turnover of

enterprises ensures that errors for large enterprises have a greater influence on the assessment of an estimation method.

17. As mentioned previously, over- and underestimations can offset one another and thus make the results of new estimation methods appear too positive. Therefore, to assess a new estimation method the weighted, average estimation errors and the absolute, weighted average estimation errors were used equally. This procedure for the present estimation method was also employed for the new estimation methods.

18. In addition the ascertainment of the best estimation method is based on the absolute average estimation error and its optimization using the average estimation errors across all reporting months. Due to the choice of absolute and non-percentage deviations, the differing turnover shares of the retail trade in respective months are accounted for, i.e. estimation errors in December have a greater influence on the choice of an estimation method.

19. Only up to 24 previous months are available in the scope of drawing up the statistics. Since some estimation methods make use of previous years' values, all estimation methods were applied to and assessed for the available 12 previous months. 24 values are not enough to obtain reliable long series of imputed values and thus the simulations might have led to optimised methods for specific parts of time series. Another critical aspect is the fact that the methods had to be optimised by comparisons with reported values and not been checked from a theoretical point of view.

B. Final results for turnover in retail and wholesale trade

20. The methods described above were intensively tested and improved. At the end the following results were obtained for retail trade (52 NACE Rev 1.1):

Estimation error [%]	Present imputation method		New imputation approach	
	Mean	Median	Mean	Median
Original	5.7	5.9	0.1	-0.1
Absolute	18.7	18.1	12.4	11.2

The table shows that the mean absolute estimation error could be reduced from more than 18% to 12%. The mean original estimation error of the new imputation approach is negligible. This result may be at random and it can not be expected that similar results will be achieved for economic branches on deeper NACE digits.

21. As the algorithm uses different methods depending only on the data intensive analysis was carried out as regards the methods' employment:

Method	Number of records		Deviation [%]		Absolute Deviation [%]	
	Absolute	Percent	Mean	Median	Mean	Median
Total	2,078	100.0	0.1	-0.1	12.4	11.2
M10	253	12.2	3.3	0.8	14.1	10.3
M20	148	7.1	2.3	0.1	14.2	13.3
M30	206	9.9	0.9	-0.9	16.8	14.3
M40	153	7.4	2.5	2.0	13.7	12.9
M60	495	23.8	-0.5	-0.9	8.8	7.9
M70	823	39.6	-2.1	-0.3	14.9	13.6

The table shows that M10, M60, and M70 were used by most of the enterprises. One common aspect of M10 and M60 is that they impute on the basis of current information. Opposed to that M70 uses

only available data of previous months. The different signs of the mean and median of the deviation indicate that the results presented in section 19 are to certain degree at random.

22. The table in the previous section shows different degrees of the methods' employment. This fact induces the question whether general principles exist that determine their use. Additional analysis was carried out only for the estimations in wholesale trade (11,000 records):

Mean values of the estimated extrapolated turnover in 1,000 EUR

Method	Months							Extrapolated Turnover	
	09/06	10/06	11/06	12/06	01/07	02/07	03/07	Mean value	Median
M10	6,452	7,266	6,890	6,359	6,141	6,131	7,328	6,653	6,452
M20	8,446	7,049	7,522	9,062	9,465	8,989	7,503	8,291	8,446
M30	8,840	10,653	5,912	6,376	8,270	8,988	11,392	8,633	8,841
M80	10,741	11,460	13,293	11,755	11,086	8,910	10,366	11,087	11,086
M60	8,048	7,356	7,683	8,075	5,791	6,017	7,082	7,150	7,356
M70	5,393	5,179	5,551	5,119	4,625	4,226	3,926	4,860	5,119
Mean values per month									
Mean values	7,987	8,161	7,809	7,791	7,563	7,210	7,933	7,779	7,883
Medians	8,247	7,311	7,206	7,226	7,206	7,520	7,415	7,721	7,901

The table reveals that the method M70 was chiefly selected in the simulations by smaller enterprises. Opposed to that M80 was chosen by the enterprises with higher turnover by contrast, for it exhibits in almost all reporting months the greatest extrapolated estimated median turnover. This indicates merely a *measurable* calendar influence in larger enterprises. M20 and M30 were also used for estimations of bigger enterprise whilst M10 was used from smaller ones.

23. As mentioned above the employment of the methods depends only on the data. This aspect leads to the question of changes as regards the frequency of changes. The following table indicates a certain stability of the employment in wholesale trade:

Estimation method frequencies in percent of the estimations (100 = 10,188)

Method	Reporting months							Mean value	Median
	09/06	10/06	11/06	12/06	01/07	02/07	03/07		
M10	23	23	21	22	24	26	25	24	23
M20	10	9	10	10	11	11	11	10	10
M30	4	5	4	4	5	5	6	5	5
M80	8	9	10	10	11	10	11	10	10
M60	16	16	16	16	16	19	19	17	16
M70	39	38	39	38	33	29	28	35	38
Total	100	100	100	100	100	100	100	100	13

The table shows relative small frequency of changes of the application rates. Detailed analysis of some records confirms this impression. Some of the checked wholesale enterprises possess changes of the method one or two times over 12 months.

C. Final results for employees in wholesale trade

24. The advancement of the imputation methods for turnover was also used for improving the imputation methods for employees. For that reason the imputation approach was used and the collection of methods was supplemented by method M90. It is the current imputation method which imputes by an enterprise's specific number of employees of the previous month. The use of the new imputation approach led to following results:

Estimation error [%]	Present method		New imputation approach	
	Mean	Median	Mean	Median
Original	0.4	-1.7	0.5	0.3
Absolute	9.9	11.0	4.9	4.2

With the exception of the original mean estimation error the new imputation approach led to significant improvements of the estimations for employees. The results were a little bit surprising because the methods were developed primarily for the variable turnover. They are reasonable because the indices on employees indicate a lower degree of volatility.

IV. FINAL REMARKS

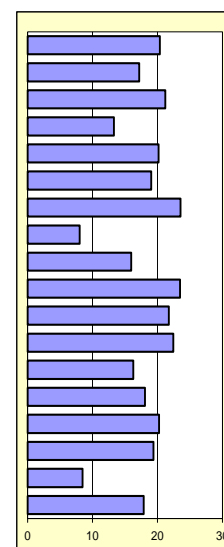
25. The simulations indicate better estimations for short term statistics in domestic trade, hotels, and restaurants in the future. The concept is very flexible and thus enables the inclusion of more powerful methods. The approach can be used for different characteristics as proved by the simulations.

26. Although method M80 bears in mind the influence of calendar effects it is expected that particularly Easter effects will influence the estimates in the future as well. So further analysis may be necessary to solve this problem if big revisions will occur. As the current database system offers only limited functionalities it should be checked whether seasonal and calendar factors of X-12-Arima may serve as workaround.

27. The new imputation approach will be realised until March 2010. As the available IT capacities are restricted it is planned to implement only M30, M60, M70, and M90. The approach will be introduced first in statistics on wholesale trade to check how it works. After that it's use will be expanded to the other statistics.

28. The new approach was planned as a "one button solution". This fact may promote an extensive use of it. Opposed to that the simulations show very clearly that the most used methods need current data of respondents. As a consequence it is recommended to monitor the amount of estimated values as indicated by the following image that shows the amount of estimated turnover per Land and month for retail trade:

Land	Reporting month												Mean
	01	02	03	04	05	06	07	08	09	10	11	12	
A	23,1	21,7	22,7	18,8	17,0	18,7	22,1	20,8	20,0	18,9	20,9	19,8	20,4
B	14,1	21,3	14,0	12,1	13,9	30,6	18,0	13,5	12,8	19,3	21,5	15,1	17,2
C	27,8	14,4	11,0	23,1	21,9	19,5	22,6	22,9	21,0	18,3	23,9	28,1	21,2
D	19,8	12,8	13,5	13,9	12,7	20,5	9,6	12,0	8,2	13,8	13,4	9,1	13,3
E	25,2	18,8	22,1	21,8	27,2	30,2	16,2	19,7	14,2	15,6	15,4	16,0	20,2
F	21,9	19,1	17,8	18,6	19,2	20,1	21,6	14,4	19,5	14,0	19,0	23,6	19,1
G	24,7	24,8	19,4	24,7	23,5	18,3	28,8	24,9	20,9	21,1	24,3	26,9	23,5
H	11,1	8,4	8,2	6,5	10,7	8,8	6,2	5,6	5,8	6,0	13,5	5,6	8,0
I	21,7	16,3	16,0	15,2	16,8	16,2	12,6	14,9	14,7	13,1	17,6	16,3	16,0
J	32,0	26,1	19,4	32,8	26,6	16,6	25,5	24,4	18,8	14,9	16,3	28,1	23,5
K	31,6	19,5	19,1	29,7	25,5	29,1	16,5	20,2	16,3	19,1	18,9	15,6	21,8
L	28,0	22,3	25,1	22,0	22,6	22,2	20,2	19,0	18,9	19,6	30,3	19,0	22,4
M	21,2	16,1	17,8	18,6	15,4	16,4	13,2	17,4	13,0	14,6	18,2	13,5	16,3
N	25,7	19,5	16,1	18,9	17,1	17,6	22,0	15,2	15,0	16,2	18,9	14,6	18,1
O	30,6	24,4	16,9	18,7	19,0	18,3	22,7	17,7	16,0	22,3	20,7	15,6	20,2
P	28,8	19,3	17,8	20,9	22,7	17,4	21,4	18,1	13,9	13,2	18,2	20,8	19,4
Q	13,8	11,7	6,5	11,7	7,9	5,3	9,3	8,0	4,0	6,6	6,6	10,2	8,5
Total	22,8	17,5	17,1	18,3	19,9	20,5	16,8	16,7	14,9	14,7	18,3	17,2	17,9
Minimum (Länder)	11,1	8,4	8,2	6,5	7,9	8,8	6,2	5,6	4,0	6,0	13,4	5,6	8,0
Maximum (Länder)	32,0	26,1	25,1	32,8	27,2	30,6	28,8	24,9	21,0	22,3	30,3	28,1	23,5
Median (Länder)	25,0	19,4	17,8	18,9	19,1	18,5	20,8	17,9	15,5	15,9	18,9	16,2	19,8



References

- Federal Republic of Germany: Trade Statistics Law (HdlStatG) of 10 December 2001 (Federal Law Gazette I p. 3438), last amended by Article 17 of the Law of 7 September 2007 (Federal Law Gazette I p. 2246), in conjunction with the Federal Statistics Law (BStatG) of 22 January 1987 (Federal Law Gazette I p. 462, 565), last amended by Article 3 of the Law of 7 September 2007 (Federal Law Gazette I p. 2246).
- European Community: Council Regulation (EC) No 1165/98 of 19 May 1998 concerning short-term statistics, Official Journal EC No L 162 p. 1, amended by Annex III No 78 of Regulation (EC) No 1882/2003 of the European Parliament and of the Council of 29 September 2003 (Official Journal EU No L 284 p. 1), amended by Regulation (EC) No 1158/2005 of the European Parliament and the Council of 6 July 2005 (Official Journal EU No L 191 p. 1), amended by Article 2 of Commission Regulation (EC) No 1503/2006 of 28 September 2006 (Official Journal EU No L 281 p. 15), amended by Article 12 of Regulation (EC) No 1893/2006 of the European Parliament and the Council of 20. December 2006 (Official Journal EU No L 393 p. 1).
- Eurostat: "Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys", http://edimbus.istat.it/dokeos/document/document.php?openDir=%2FRPM_EDIMBUS
- Eurostat: "Verhaltenskodex Europäische Statistiken", http://epp.eurostat.ec.europa.eu/pls/portal/docs/PAGE/PGP_DS_QUALITY/TAB47141301/VERSIONE_TEDESCO_WEB.PDF
- Statistical Offices of the Federal Government and the *Länder*: PL-Fachkonzept (stanet-web.stba.testa-de.net/jetspeed/portal/cms/Sites/destatis/StaNet/Navigation/EVAS__Navi/0__Statistikuebergreifen d/Arbeitshilfen/PL__Fachkonzept/PL__Fachkonzept.psm1)
- Statistical Offices of the Federal Government and the *Länder*: Qualitätsstandards in der deutschen amtlichen Statistik, [Document on the website of the Federal Statistical Office](#), Wiesbaden 2003
- Douglas C. Montgomery; Cheryl L. Jennings; Murat Kulahci: "Time series analysis", Hoboken 2008
- Joachim Hartung: "Statistik", 14th edition, Munich 2005
- Klaus Neusser: "Zeitreihenanalyse in den Wirtschaftswissenschaften," Wiesbaden 2006
- NIST/SEMATECH: "e-Handbook of Statistical Methods", www.itl.nist.gov/div898/handbook/, 2008
- UNECE: "Statistical Data Editing", Volume 1, www.unece.org/stats/publications/editing/SDE1.htm, 1994