STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 4
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic I: Management and evaluation of editing and imputation procedures

CHARACTERISATION OF QUALITY IN SAMPLE SURVEYS USING PRINCIPAL COMPONENTS ANALYSIS

Submitted by the Office for National Statistics, United Kingdom and
the US Department of Energy, United States[1]

Invited paper

I. INTRODUCTION

1. There are several components of quality, and classifications to describe all of these components have been devised over the last 15 years (Groves 1989, Davies & Smith 1999). In order to get some overall measure of quality, Groves (1989) also uses the concept of *total survey error*, where the biases and variances are all evaluated and put together to give an estimate of the overall mean squared error (mse) of the survey. This captures the quality in overall accuracy terms, but is extremely expensive to evaluate.

2. Instead of taking such a minute approach, we can think of quality as a multivariate measure for any dataset. Each of the quality indicators developed by Groves (1989), Davies & Smith (1999) then becomes one dimension of quality. However these dimensions are often interrelated – a low quality response from a particular sample unit will affect several indicators – so the amount of additional information obtained from a new indicator may be relatively small.

3. In order to make the assessment of quality as straightforward (and inexpensive) as possible, it would be good to concentrate on a small number of indicators which provide most of the information about the data quality. We aim to achieve this from a relatively wide-ranging set of the most easily calculated indicators by using principal components analysis to find the measures which best capture the underlying variation in the data quality measures.

II. METHODS

Survey Data

4. The survey from the UK used here is the Monthly Inquiry into the Distribution and Services Sector (MIDSS), a monthly survey collecting turnover and employment information (employment from a sub-sample only) from a stratified sample of businesses in wholesaling and the services (but excluding retailing). Six variables are used in the analysis here:

- turnover (question code 40)
- employees (50)

[1] Prepared by Paul Smith (Office for National Statistics) and Paula Weir (US Department of Energy).

- male full-time employees (51)
- male part-time employees (52)
- female full-time employees (53)
- female part-time employees (54)

The last four variables should sum to total employees. In this example, data from September 1999 are used to illustrate the approach.

Indicators

5. Several measures are used for different purposes:

- Sampling

Sampling is the most easily measured component of quality from a design-based perspective, because appropriate theory is available. We use two measures:

i) sampling fraction
ii) sampling error

- Response rate

An indicator of the possible size of non-response error. Several versions are used here

iii) "basic" response rate – good responses/sample size;
iv) return rate – all returned questionnaires/sample size;

(the difference between the two indicators above will contain some information about the quality of the sampling frame, as it will capture the number of dead or out of scope sample members)

v) weighted response rate – response by an auxiliary variable (register turnover)
vi) weighted return rate

Data editing indicators

Indicators both of the quality of the data contained in the responses and the quality of the data editing process. Here we use indicators of a number of different aspects

vii) number of warnings (non-fatal edits) subsequently cleared without change
viii) number of warnings leading to changes
ix) number of errors (fatal edits) subsequently corrected

The actual set of edit rules (not optimised in this case) is given in annex 1.

6. These 9 indicators give 46 variables for use in analysis – one for each of 6 questions except for 1 (all employment questions have the same sampling fraction) and 9 (no fatal edit failures for the four subdivisions of employment).

Principal Components

7. The aim of principal components is to find the linear combinations of groups of related variables which have the largest variation (see for example Mardia, Kent & Bibby 19??). In order for the procedure to be used, a reasonably large number of observations is needed. In this case this shows that we need to calculate the indicators on a range of domains. There is undoubtedly a number of ways in which these domains could be devised, but the choice made here is to calculate them by stratum. There are 144 strata containing usable information in MIDSS.

8. In order to avoid the variability of measures with large units swamping the measures with small units, the columns (variables) must be standardised before principal components analysis is undertaken, by subtracting the mean and dividing by the standard deviation.

III. RESULTS

9. The 46 variables resolve into 32 distinct principal components. The remaining 14 variables are collinear (or very nearly so) with other variables in the dataset; for example, a number of the response rates for different questions show high collinearity as there is little item non-response in this survey. The first 5 principal components explain >80% of the variability in the data; the actual proportions of the variability are shown in the table below.

Principal component	Proportion of variation explained (%)	Cumulative proportion of variance explained (%)
1	40.49	40.49
2	19.63	60.13
3	8.23	68.35
4	6.89	75.24
5	3.13	78.37
6	2.41	80.78
7	2.30	83.08
8	2.16	85.24
9	1.85	87.09
10	1.64	88.74
11-32	11.26	100.00
33-46	0.00	100.00

10. The loadings (coefficients premultiplying the input variables to derive the principal components) give some information on what the main determinants of the principal components are (the larger coefficients show which variables are most important in each principal component). Only the first five pc's are shown in the table. These seem to divide naturally between the main types of indicators – the first and third principal components are derived mostly from the response rates, the first picking out the information in the various response rates to give an "overall response indicator" (since all the coefficients have the same sign). The third, in contrast, picks out the difference between the weighted and unweighted response rates (as the coefficients for these have different signs).

11. The similarity of the coefficients between questions and the way in which the coefficients divide into blocks by the indicator, and not by the question, indicates that (very) little additional information is obtained by calculating the same indicators for another question in the survey, and that more information is obtained from an additional indicator. (A few coefficients do not agree with the blocks within which they are placed – these are more heavily shaded to pick them out.)

12. The second principal component (pc) picks out sampling errors and, to a smaller degree, the (non-fatal) edit failures which are identified by editing but which are unchanged after follow-up. Both of these contribute to the underlying variability, one through the sampling variability, one through the estimated population variance (which is increased with more variable response data). Hence we can interpret this pc as a "sampling variability" indicator. The number of unchanged non-fatal edits does not contribute substantially to any of the other first five pc's, and hence could legitimately be left unmeasured with little loss of information on this component.

Indicator	Question code	Loadings in				
		pc 1	pc 2	pc 3	pc 4	pc 5
3. response rate (number)	40	0.20	-0.04	-0.16	-0.05	-0.02
	50	0.21	-0.03	-0.18	-0.04	0.00
	51	0.21	-0.02	-0.19	-0.03	0.03
	52	0.21	-0.02	-0.19	-0.03	0.03
	53	0.21	-0.02	-0.19	-0.03	0.03
	54	0.21	-0.02	-0.19	-0.03	0.03
5. response rate (turnover)	40	0.17	0.07	0.06	-0.03	-0.13
	50	0.19	0.10	0.21	0.03	-0.04
	51	0.20	0.11	0.20	0.04	-0.01
	52	0.20	0.11	0.20	0.04	-0.01
	53	0.20	0.11	0.20	0.04	-0.01
	54	0.20	0.11	0.20	0.04	-0.01
4. return rate (number)	40	0.20	-0.03	-0.18	-0.03	-0.01
	50	0.21	-0.03	-0.19	-0.01	0.00
	51	0.21	-0.02	-0.20	-0.00	0.03
	52	0.21	-0.02	-0.20	-0.00	0.03
	53	0.21	-0.02	-0.20	-0.00	0.03
	54	0.21	-0.02	-0.20	-0.00	0.03
6. return rate (turnover)	40	0.16	0.08	0.06	-0.02	-0.12
	50	0.19	0.10	0.21	0.04	-0.03
	51	0.20	0.11	0.20	0.05	-0.00
	52	0.20	0.11	0.20	0.05	-0.00
	53	0.20	0.11	0.20	0.05	-0.00
	54	0.20	0.11	0.20	0.05	-0.00
7. warnings cleared w/o amendment	40	-0.03	0.14	-0.06	0.20	0.03
	50	-0.01	0.20	-0.04	0.03	-0.03
	51	0.07	-0.17	0.03	-0.04	0.09
	52	0.04	-0.12	0.10	0.02	0.15
	53	0.06	-0.20	0.11	-0.06	0.09
	54	0.04	-0.22	0.16	-0.08	-0.10
8. warnings leading to amendments	40	-0.04	0.02	-0.07	0.44	-0.06
	50	-0.04	0.03	-0.08	0.42	-0.24
	51	0.02	-0.07	0.02	0.22	0.33
	52	0.02	-0.07	0.05	0.02	-0.22
	53	0.05	-0.19	0.06	0.17	0.43
	54	0.04	-0.18	0.05	0.11	0.43
2. coefficient of variation	40	-0.03	0.23	-0.10	-0.01	-0.02
	50	-0.05	0.27	-0.00	-0.09	0.26
	51	-0.04	0.27	0.00	-0.08	0.24
	52	-0.04	0.28	-0.10	-0.07	0.17
	53	-0.07	0.27	0.00	-0.08	0.23
	54	-0.05	0.28	-0.05	-0.08	0.20
1. samp fraction	40	0.06	-0.28	0.15	-0.11	0.06
	50	0.05	-0.28	0.17	-0.10	0.03
9. no of errors	40	-0.02	0.04	-0.08	0.45	-0.15
	50	0.01	-0.10	0.00	0.44	0.21

13. The fourth pc picks out the fatal and non-fatal errors which *are* amended during the course of survey processing, a "proportion of errors" indicator.

14. The fifth pc is a less easily interpretable combination of the sampling error and the non-fatal edits which are subsequently amended, mostly with opposite signs, suggesting some further aspect of underlying sampling variability.

15. The amount of extra information within each variable can also be interpreted with reference to the correlation matrix for the input variables (not shown), which indicates when variables are highly correlated and where one such variable may be sufficient.

IV. RESULTS, SECOND EXAMPLE

16. The same method was also applied to survey data from the U.S. Energy Information Administration's Annual Fuel Oil and Kerosene Sales Report. This survey collects sales volume of major petroleum products by end-use sector by state from a cross-stratified sample of retailers and resellers. Six variables were also used in this analysis which were:

- No. 2 fuel oil sales for commercial use (product-line 11)
- Low sulfur No. 2 diesel sales for commercial use (product-line 12)
- High sulfur No. 2 diesel sales for commercial use (product-line 13)
- No. 2 fuel oil sales for industrial use (product-line 16)
- Low sulfur No. 2 diesel sales for industrial use (product-line 17)
- High sulfur No. 2 diesel sales for industrial (product-line 18)

17. The same indicators were used to the extent that the data allowed. One indicator, number of fatal errors subsequently corrected, was not possible because fatal failures are immediately rejected and not even written to the database. For the sampling fraction, the same sampling rate applied to all variables. The same four response rates were also used; however, because there is no partial non-response in this survey, the response rate/return rate did not vary among the variables. The weighted rates did vary because of the different weights used. As a result, the 8 indicators give 33 variables for the analysis performed on the stratum as the observation.

18. In this analysis, the first five principal components explained less of the variability than the first survey, only 65 %. The 80% level was not achieved until tenth principal component as shown in the table below.

Principal component	Proportion of variation explained (%)	Cumulative proportion of variance explained (%)
1	31.02	31.02
2	17.92	48.94
3	5.79	54.73
4	5.52	60.26
5	4.32	64.58
6	4.00	68.58
7	3.41	71.99
8	3.15	75.13
9	2.78	77.91
10	2.74	80.65
11-32	19.34	99.99
33-	.01	100.00

The loadings in the first five principal components are shown in the table below.

19. The first component picks out the weighted rates, both the response and return rates, and the sampling fraction, with very similar coefficients. The second and third principal components 1) both pick the (unweighted) response and return rates (with similar coefficients in size), and warnings cleared without change, but have opposite signs between the components, and 2) coefficients of variation. The second principal component also picks up the sampling fraction, but with an opposite sign from the first component. The fourth and fifth components pick warnings leading to amendments, but vary in size and sign, and the fourth principal component also picks the weighted return rates and the fifth principal component picks the unweighted response rates. Also highlighted in the table are the larger coefficients noticeable in the fifth component for product-lines eleven (11) and sixteen (16), No. 2 fuel oil commercial and industrial, respectively, for the weighted response and return rates, as well as warnings changed or not changed. These higher coefficients are also evident in the fourth component for all but the warnings leading to change.

V. FURTHER ANALYSIS

20. The principal component analysis was repeated using Proc Factor instead of Princomp and rotating the first six components, as suggested by the scree.

21. As compared to the princomp that yields eigenvectors, factor unrotated gives the eigenvector times the square root of the eigenvalue. The six factor solutions are interpretable. The rotated factors with the focus on varimax factor pattern corresponds to:

Indicator	Product-line code	Loadings in				
		pc 1	pc 2	pc 3	pc 4	pc 5
3. response rate	ALL	.09	.24	-.25	-.09	.20
5. response rate (weighted)	Line 11	.23	.06	-.12	.29	-.15
	Line 12	.26	.11	-.06	-.14	.04
	Line 13	.26	.05	-.03	-.14	-.08
	Line 16	.26	-.02	-.14	.19	-.21
	Line 17	.26	.01	-.01	-.20	.06
	Line 18	.27	.01	.00	-.17	-.07
4. return rate	ALL	.08	.24	-.25	-.09	.20
6. return rate (weighted)	Line 11	.23	.08	-.11	.32	-.13
	Line 12	.25	.14	-.05	-.10	.04
	Line 13	.26	.07	-.01	-.12	-.07
	Line 16	.25	.01	-.12	.23	-.21
	Line 17	.26	.04	.00	-.20	.08
	Line 18	.27	.04	.02	-.15	-.06
7. warnings cleared w/o amendment	Line 11	.14	-.19	.22	.18	.21
	Line 12	.13	-.14	.35	.01	.01
	Line 13	.11	-.18	.34	-.04	-.11
	Line 16	.13	-.20	.08	.22	.28
	Line 17	.15	-.14	.24	.08	.07
	Line 18	.13	-.21	.34	-.08	-.14
8. warnings leading to amendments	Line 11	.09	-.15	-.06	.17	.46
	Line 12	.10	-.08	-.10	.04	.26
	Line 13	.08	-.08	-.17	.23	-.08
	Line 16	.10	-.16	-.09	.15	.43
	Line 17	.07	.01	-.08	-.24	.15
	Line 18	.06	-.13	.02	-.14	-.18
2. coefficient of variation	Line 11	.02	.26	.20	.31	.03
	Line 12	-.01	.31	.23	.10	.13
	Line 13	.03	.31	.22	.05	.05
	Line 16	.05	.27	.13	.33	-.14
	Line 17	.04	.28	.24	-.09	.18
	Line 18	.06	.29	.26	-.04	.05
1. sampling fraction	ALL	.21	-.26	-.06	-.02	-.03

1) Weighted response and return rate (with product-lines 11 and 16 having lower loadings)
2) Coefficients of Variation
3) Warnings resulting in no change
4) Product-lines 11 and 16 weighted return and response rate.
5) Product-lines 11 and 16 warnings resulting in change or no change (with product-line 12 warnings resulting in change also included)
6) Unweighted return and response rate.

22. The sampling fraction and the warnings resulting in change for product-lines 13, 17 and 18 do not follow a simple structure. The varimax prerotation method factors are presented in the table below with the factor pattern highlighted.

Principal Components Prerotation Method: Varimax
Rotated Factor Pattern

	factor1	factor2	factor3	factor4	factor5	factor6
Response rate	0.31350	0.17518	-0.15353	0.00852	-0.09675	0.85595
Wtd response rate 11	0.50538	0.10789	0.18418	0.67506	0.09210	0.22382
Wtd response rate 12	0.84050	0.15275	0.07702	0.11443	0.08828	0.21714
Wtd response rate 13	0.83567	0.03821	0.15862	0.16076	0.04545	0.04349
Wtd response rate 16	0.63699	-0.06196	0.16119	0.61718	0.14271	-0.03928
Wtd response rate 17	0.84272	0.00157	0.18444	0.01633	0.18666	0.04814
Wtd response rate 18	0.87898	-0.00026	0.20628	0.12765	0.09706	0.04196
Return rate	0.30009	0.18378	-0.16997	0.00750	-0.10390	0.85825
Wtd return rate 11	0.48320	0.17791	0.17354	0.68424	0.10282	0.23449
Wtd return rate 12	0.81786	0.22386	0.05534	0.14811	0.09075	0.22670
Wtd return rate 13	0.82637	0.10601	0.13876	0.16993	0.04017	0.05212
Wtd return rate 16	0.61829	0.03236	0.13021	0.66163	0.13347	-0.02931
Wtd return rate 17	0.84851	0.06701	0.15596	0.00419	0.18765	0.05564
Wtd return rate 18	0.87694	0.08004	0.18185	0.12857	0.09468	-0.03978
Warnings – no change 11	0.11963	-0.08087	0.60868	0.13582	0.47829	-0.07708
Warnings – no change 12	0.18927	-0.01109	0.72275	-0.00335	0.10512	-0.08544
Warnings – no change 13	0.14603	-0.16474	0.78083	-0.01287	-0.04602	-0.07210
Warnings – no change 16	0.11098	-0.14659	0.43160	0.17829	0.60827	-0.05320
Warnings – no change 17	0.19643	-0.06924	0.67404	0.10031	0.19728	0.05961
Warnings – no change 18	0.24453	-0.21020	0.75454	-0.01881	-0.01101	-0.23531
Warnings – change 11	0.07231	-0.11108	0.10644	0.04902	0.75949	-0.01371
Warnings – change 12	0.27854	-0.06822	-0.08024	0.04280	0.52419	-0.14209
Warnings – change 13	0.09573	-0.15927	-0.01524	0.42621	0.20273	-0.08123
Warnings – change 16	0.12066	-0.17451	0.13216	0.06902	0.71873	0.04623
Warnings – change 17	0.28579	0.01223	0.11632	-0.29995	0.05984	0.07951
Warnings – change 18	0.25266	-0.25883	0.08485	-0.00716	-0.02503	-0.36051
Coefficient var. 11	-0.08012	0.72967	0.03918	0.25389	-0.08497	0.23090
Coefficient var. 12	-0.00871	0.81003	-0.12663	-0.07678	-0.09608	0.15654
Coefficient var. 13	0.16006	0.78450	-0.14448	-0.03659	-0.13152	0.06802
Coefficient var. 16	0.07906	0.72685	-0.10465	0.42448	-0.14087	0.03299
Coefficient var. 17	0.25446	0.73443	-0.11635	-0.25603	-0.04309	0.08790
Coefficient var. 18	0.29917	0.76478	-0.08961	-0.11538	-0.11995	0.00973
Sampling fraction	0.46365	-0.51114	0.45388	0.22427	0.30035	-0.09130

VI. CONCLUSIONS

- Most of the variation in this set of indicators is explained by the response rates.
- Weighted and unweighted forms contain different information, but using different questions or alternative response rate definitions adds very little.
- Sampling errors and sampling fractions contain similar information.
- More complete "sampling error" information is supplemented by the number of non-fatal edits which are not amended.
- Fatal edit failures and non-fatal edits which do lead to amendment also contain similar information – one of these may be sufficient as an indicator.

References

Davies, P. & Smith, P. (eds.) (1999) *Model Quality Reports in Business Statistics*. ONS, UK.

Groves, R. (1989) *Survey errors and survey costs*. Wiley, New York

Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979) *Multivariate Analysis*. Academic Press, London.

ANNEX 1: EDIT TESTS

MIDSS current turnover tests

30011 - Credibility check (Only if 6 returns in past year)
Value must lie between the average value over the last year ± the larger of:

(i) 2·57 × the standard deviation; or
(ii) 0·1 × the average

If the value lies outside of this range generate a warning.

30012 - Large value change
If the check in (i) fails, check (ii)

(i) Check that the current value lies in the following interval,

$$0.5 < \frac{\text{returned turnover}}{\text{previous turnover}} < 2.0$$

(ii) Check that the current value lies between

$$0.5 < \frac{r_1}{r_2} < 2.0$$

Where r_1 = value this period / value last period
r_2 = same as r_1 but for previous year

If last years ratio is unavailable, use the imputation link from the period a year prior to the current period.
Generate a warning if r_1 / r_2 lies outside of the range.

30013 - Large turnover ratio (Large sph change)
If the check in (i) fails, check (ii)

(i) Check that the current sph value lies in the following interval,

$$0.5 < \frac{\text{current sph}}{\text{previous sph}} < 2.0$$

Where, sales per head (sph) = total turnover / selection employment

(ii) check that the current value lies between

$$0.5 < \frac{r_1}{r_2} < 2.0$$

Where, r_1 = value this period / value last period
r_2 = same as r_1 but for previous year

If last years ratio is unavailable, use the imputation link from the period a year prior to the current period. Generate a warning if r_1 / r_2 lies outside of the range.

30014 - High turnover (High value check)

Value must be ≤ 150% of the previous highest value for the contributor. If the current turnover exceeds this value, a warning should be issued.

30015 - Low turnover (Low value check)
Value must be ≥ 50% of the previous lowest value for the contributor. If the current turnover is lower than this value, a warning should be issued.

30044 - Turnover per head out of range
If a contributor is responding for the first time and the value lies outside of the following range, generate a warning.

120% of the industry turnover per head ≤ turnover per head ≤ 80% of the industry turnover per head

Additional total turnover tests added by DVU

Test 13
If total turnover lies outside the following range, generate a warning.

$$0.333 < \text{total turnover} < 9{,}999{,}999$$

Test 18
Check that the month on registered turnover comparison lies within the following interval,

$$0.033 < \frac{\text{total turnover}}{\text{register t urnover}} < 0.5$$

If the value lies outside of this range generate a warning.

Test 36
If the cost of sales is greater than the total turnover, i.e. Question 36 > Question 40, generate a warning.

Test 47
If the total turnover is zero generate a warning.

Test 48
If turnover > £500k and is equal to the previous period, generate a warning.