WP. 31

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iii): Data editing processes within survey processing

IMPECT: RECENT DEVELOPMENTS IN HARMONIZED PROCESSING AND SELECTIVE EDITING

Supporting Paper

Submitted by Statistics Netherlands[1]

Abstract: *Following the implementation of a uniform system for processing structural business statistics (IMPECT1), Statistics Netherlands has applied a similar approach for dealing with short-term statistics (IMPECT2). Again, harmonization and efficiency (economies of scale and selective editing) have been the key words. Although the number of variables is much smaller than in structural business statistics, the complexity of developing a new system for the short-term statistics has proven to be nonetheless substantial. Calculating a reliable growth rate and processing data as efficiently as possible requires not only a well thought-out methodological concept but also a well-oiled data-handling process.*
In this paper, the IMPECT2-system is described, with a focus on some of the more interesting parts of the process.

I. INTRODUCTION

1. The new IMPECT2 system is designed for short-term enterprise-based statistics and focuses on turnover. Depending on industry and size-class, one or more variables are collected on a monthly or quarterly basis. For most industries, if possible, VAT-data is used for small enterprises (less than 10 persons employed) instead of questionnaires. Furthermore, enterprises are given the opportunity to report by e-mail using a specific electronic questionnaire. In this way, the survey sample has been reduced, lowering the response burden and the editing costs. Applying selective editing has reduced the editing costs even more.

2. Contrary to annual business statistics, harmonization of variables for short-term statistics was not very difficult. However, some differences remained and led to different questionnaires. For instance, there is a difference in registration of turnover. In most industries the turnover is registered excluding VAT; in retail trade however it is custom to register including VAT. We decided to take this difference into account and have adjusted our questionnaires accordingly. Furthermore, for some industries questions have been added regarding the value of orders received and regarding the breakdown of turnover into domestic and international turnover. Although there are different variables involved, in the following description of the process we focus on the key variable: total turnover.

3. This paper gives an overview of the entire logistical and statistical process in chapter II. In chapter III some interesting parts of the process are discussed in more detail. Chapter IV contains a brief

[1] Prepared by Arjan de Jong (gjog@cbs.nl).

description of the work ahead. The organisational dimensions, which go hand in hand with the new approach, are discussed in chapter V and finally, in chapter VI some conclusions are drawn. Explanatory notes regarding indications used for statistical periods and different types of cells can be found in appendix A.

II. OVERVIEW OF THE PROCESS

4. In this chapter, the different steps of the new process will be described chronologically in general terms. Starting with the survey design, some characteristics of the data collection process, the editing and the statistical analysis will be discussed.

A. Survey design

5. As stated before, one of the main goals of the redesign is to improve efficiency. This has been obtained in several parts of the process, starting with the survey design.

6. For each reference period (month or quarter), a sample, stratified on core cell and size class (see appendix A), is drawn from the business register. The sampling frame consists of all companies in the business register, excluding non-market producers. The allocation has been optimized on minimal variance of the key-variable: total turnover.

7. Over the last couple of years, extensive research has been carried out at Statistics Netherlands into the possible use of VAT-data for turnover statistics. However, in some industries the differences between the fiscal and statistical definition of turnover has prevented the use of VAT-data as a substitution for individual data collection. Furthermore, joining a fiscal and statistical unit is more difficult for medium size and large companies. Therefore, the use of VAT-data had to be limited to smaller companies (less than 10 persons employed) and could not be applied in some industries. Nonetheless, this has led to an economization of several thousands of sample-units and a substantial reduction of response burden.

B. Data collection

8. Questionnaires are sent out at the end of the reference period. In this way, companies receive the questionnaire at the closing of the accounts, which has proven to be the most appropriate moment for completion. After 10 working days, non-respondents are rated automatically based on their individual contribution to the estimate and the contribution of the cell (see section III B). Depending on the available capacity, the most important non-respondents are contacted by telephone, while the other non-respondents receive a letter of recall.

9. Incoming questionnaires are registered and entered without editing. In addition to a page with questions on turnover, the questionnaire contains a page for comments from the data supplier. These comments are processed separately and may lead to changes in the logistical properties of the unit. An increasing number of companies is switching over to the use of electronic questionnaires. At present, about 15% of the response consists of electronic questionnaires. The incoming data is transformed automatically to the right format.

10. VAT-data is joined with statistical units and filed in BaseLine, a central database at Statistics Netherlands to store secondary data. The relevant data for this survey is derived from this database and, after a specific editing procedure (see section III A), added to the IMPECT2-database.

C. Editing

11. In the editing process, records are checked and corrected if necessary. This process can be subdivided into 3 phases: 1 automatic correction, 2 establishing plausibility, and 3 either interactive or computerized editing.

12. ***Automatic correction***. Since the incoming questionnaires are entered without editing, the data will contain a large number of errors. Some errors are obvious and are corrected automatically in this stage of the process.

13. Firstly, about 10% of all records contain values in euro instead of in 1000 euro. This type of error is detected using the respondents' previously reported turnover or the average turnover of the stratum. If this turnover ratio is over 300, the financial variables are divided by 1000.

14. Secondly, records with more than one variable may have missing values. If the missing value or values can be determined with almost absolute certainty, they are filled in.

15. ***Plausibility and selection.*** In order to reduce editing costs and improve timeliness, the attention of the editing staff should be concentrated on important records: records that are implausible and have a noticeable impact on the estimates. To identify important records, two aspects have been taken into consideration: contribution to the weighted estimate and growth rate (for more details see section III C). If a record scores above preset limits on both aspects, it is selected for interactive editing. All other records are edited by computer.

16. ***Interactive editing***. The editing staff edits the record by resolving the (indicated) errors and implausibilities. If necessary, the respondent can be approached by telephone. On the editing screen, a lot of information is at the editors' disposal to facilitate taking the right editing decision.

17. ***Computerized editing***. As the questionnaires contain only a small number of variables, computerized editing is kept rather simple. Only errors in variables that are involved in an addition can be resolved. Remaining item-nonresponse is dealt with in the imputation process.

D. Statistical analysis

18. After finishing the editing, the data processing continues with a number of activities that take place under the responsibility of the statistical analysis department.
First, the clean data is transformed from reporting period to statistical period. Then, automatic imputation for unit- and item non-response takes place, outliers are detected followed by the weighting process. This results in indices and growth rates per publication cell that, with reference information, are presented to the analyst. The analyst can view the micro data and in case of remaining implausibilities, decide to correct the record or to contact the editing department.

19. ***Transformation***. Depending on industry, a certain percentage of companies report on four-week basis, some use alternating four- or five-week periods. As the editing might involve contact with the reporting company but the publication is on a monthly basis, these records are transformed in this stage of the process. First, the number of working days of the reported period is established (working days may differ between industries!), after which an average turnover per working day is calculated. Then, this average is multiplied by the number of working days in the calendar month. If there are two reporting periods available, together covering the calendar month, the weighed average of both periods is used to calculate the monthly turnover. If less than 30% of the working days of a specific month are covered, the unit is considered as non-response.

20. ***Establishing population.*** The first step in the imputation process, is establishing the population to be used for imputation and weighting. For period [t] we need the population of [t] (N_t) and the population of the previous period, [t-1] (N_{t-1}). These two populations differ as a result of new companies

("birth") and companies that have ceased to exist ("mortality"), and as a result of administrative changes (e.g. merging or splitting up of companies). Since we do not want administrative changes to influence the estimates, the populations are corrected for these changes. Birth and mortality are regarded as "real" phenomena and are consequently accepted.

21. ***Imputation***. In the imputation phase, unit- and item non-response is treated in the same way. The imputation only takes place for survey-samples; mass-imputation (imputation of all records in the population) is not applied.

22. In general terms, the turnover in [t] of a non-respondent is calculated by multiplying the (imputed or reported) turnover of the non-respondent from the previous period [t-1] with the growth rate of respondents in the same cell. If there are insufficient respondents in the cell to calculate a reliable growth rate, the cell is joined with neighbouring cells (see section III D for a more detailed description).

23. ***Weighting.*** Following the imputation of item- as well as unit non-response, the weighting process starts. First the weighting method is selected: with or without use of VAT-data as auxiliary information. Then outliers are detected and dealt with, after which the actual weighting takes place.

24. Preferably, VAT-data is used in the weighting scheme. However, experience has shown that VAT-data is not in time for calculation of preliminary results (within 30 days after the reporting period) for a sufficient number of units. Furthermore, VAT-data cannot be used in parts of some industries due to the difference between fiscal and statistical turnover. Finally, VAT-data is ignored if a weighting cell contains less than 20 records with VAT-scores. In all these situations, direct weighting is applied.

25. Outliers are distinguished into two categories. The first category comprises records with a large difference between VAT-score and reported score. The second category is formed by records with a reported score that differs much from the median score of the weighting cell. Eventually, all outliers are assigned a weighting factor of 1.0.

26. If a weighting cell encloses less than 10 records, the cell is joined with the neighbouring weighting cells within the same size class group. If the minimal number of 10 records is still not met, this group of weighting cells is joined with the corresponding size class group of a core cell that is part of the same publication cell for the preliminary results.

27. ***Calculation of indices***. The weighted values are aggregated at the appropriate level for both [t] and [t-1], after which the growth rate is calculated. The index for period [t] is defined as the product of growth rate and index of the previous period. For each variable in the survey, an index is calculated at the most detailed level of publication.

28. For part of the survey, VAT-data is used as substitution for data derived from questionnaires. As stated above, the VAT-information is not in time for the preliminary results. For this reason, the growth rates for the strata involved need to be estimated. The growth rate for these smaller size classes (SBS size class 10-30: up to 10 persons employed) G_s is derived from the established growth rate in the medium size classes (SBS size class 40-60: 10 to 100 persons employed) G_m, corrected for the difference in population development in the size class groups: $G_s = G_m \,.\, \frac{N_{m,t-1}}{N_{m,t}} \,.\, \frac{N_{s,t}}{N_{s,t-1}}$. In this situation the aggregated turnover is calculated as a result of the growth rate instead of the other way round. This estimation is only used for the preliminary results; the indices of following releases are based on VAT-data.

29. ***Macro checking and validation***. A final but vital step in the analysis process includes human intervention. The results of the preceding steps are presented to analysts. The indices and growth rates can be shown at a chosen level, accompanied by expected values, response percentages, etc. If the figures cause suspicion, the analyst is able to view all micro data of the cell involved, with indication of

automatically detected outliers. The analyst can overrule or add outlier indications, correct records or contact the editing department for further information. After the validation, publications can be generated.

30. ***Publication strategy.*** For the monthly statistics, three releases are published. The first is disseminated 30 days after the reporting period and contains preliminary results for the key variable (total turnover) on a high level of aggregation of publication cells. One month later, the next publication is brought out, with the key variable for all publication cells. Finally, three months after the reporting period, the third release comprises all variables and publication cells. For quarterly statistics, two releases are published. Filters are used to prevent disclosure of confidential or unreliable data.

III. DETAILED DISCUSSION

31. As referred to in the introduction, some parts of the general process might be interesting enough to elaborate on. First, we will focus on two specific parts of the data collection process: the editing of VAT-data and the prioritization of follow up. Selective editing is the essence of the editing process and deserves a closer look as well. Finally, the inns and outs of the imputation process will be discussed.

A. Editing of VAT-data

32. The VAT-data that Statistics Netherlands receives, are based on fiscal units. These units may differ from the statistical units we want to describe. For that reason, not all VAT-records are usable for statistical purposes. Our department of the business register is responsible for matching fiscal with statistical units. In our process, we only use records that match completely. However, the remaining records may still contain implausible values. These records need to be detected and eliminated from the files we use in our survey. Note that only one procedure is used to edit VAT-data, regardless the use (as substitute for individual data collection or as auxiliary information for the weighting process).

33. First, records with a turnover smaller than or equal to zero are deleted. Then, the median turnover is assessed for all combinations of core cell and size class. For each record the so called "factor score" is calculated as the ratio between the turnover of the record and the median turnover of the corresponding core cell and size class. This is executed per period (month or quarter, depending on the frequency of the survey) for [t], [t-1] and [t-2].

34. The factor scores are now used to detect records with an extreme growth rate. If the ratio of the factor scores of a record for two successive months (highest factor score divided by the smallest) is more than 40, we regard the growth rate as extreme. As we presume that the monthly fluctuation can be larger than the quarterly fluctuation, a limit of 20 is applied for quarterly statistics.

35. If a record is detected in the previous step, it has to be eliminated for one of the two periods, either [t] or [t-1]. The selection is based on the comparison with the factor score of [t-2]. If the absolute value of the difference between the factor scores of [t] and [t-2] is larger than that of [t-1] and [t-2], the record is eliminated from period [t]. Else, the record is eliminated from [t-1].

36. The result of the procedure described above are VAT-files that only contain usable records. These records are used both individually (as a substitute for individual data collection and for the weighting process) as well as aggregated (as auxiliary information for the weighting process).

37. The use of these factor scores instead of the absolute difference between the turnover of the record and the median turnover, has two advantages. First, a record that has a relatively large turnover will not be eliminated unjustly, because its factor scores will be high in both [t] and [t-1]. Second, seasonal effects are automatically taken into consideration, as the median turnover will be affected in the same way as the turnover of the individual unit.

B. Prioritizing follow up

38. The critical factor with regard to improved timeliness is mainly the response rate. To be able to produce results as soon as possible, it is vital to direct the available capacity to the most important non-respondents. Experience has shown that contact by telephone is the fastest way to collect data from non-respondents. However, this is very time consuming and there is always a shortage of capacity. Thus, a methodology has been worked out to prioritize the following up.

39. For the rating of non-respondents, two aspects are taken into consideration: 1. the individual contribution of the non-respondents' estimated turnover to the estimated turnover of the core cell (*CR*) and 2. the weighted response rate (sum of turnover of respondents as a percentage of the total estimated turnover) of the core cell (*RR*).

40. As the turnover of non-respondents for period [t] is still unknown, and the turnover for [t-1] is not yet stable, the turnover of [t-2] is used as a substitute for the calculation of *CR* and *RR*. However, not all sample elements in [t] were also in the sample of [t-2]. For these units, we will have to estimate the (weighted) turnover. For that reason we have defined four collections of units at the level of the core cell:
A: units in sample of both [t-2] and [t] and responding in [t]
B: units only in sample [t] and responding in [t]
C: units in sample of both [t-2] and [t] (both response and non-response)
D: units only in sample [t] (both response and non-response)

For the contribution-rate *CR* for units in collection C, the following formula can be written:

$$CR_C = \frac{Turnover_{unit,t-2} \text{ x } Weighting\ factor_{unit,t-2}}{\sum_C (Turnover_{unit,t-2} \text{ x } Weighting\ factor_{unit,t-2}) + \sum_D (\overline{Turnover_{stratum,t-2}} \text{ x } Inclusionweight_{stratum,t})}$$

Likewise, for the contribution-rate for units in collection D:

$$CR_D = \frac{P_{new} \,.\, (\overline{Turnover_{stratum,t-2}} \text{ x } Inclusionweight_{unit,t})}{\sum_C (Turnover_{unit,t-2} \text{ x } Weighting\ factor_{unit,t-2}) + \sum_D (\overline{Turnover_{stratum,t-2}} \text{ x } Inclusionweight_{stratum,t})}$$

For the response-rate of the core cell *RR*:

$$RR = \frac{\sum_A (Turnover_{unit,t-2} \text{ x } Weighting\ factor_{unit,t-2}) + \sum_B (\overline{Turnover_{stratum,t-2}} \text{ x } Inclusionweight_{stratum,t})}{\sum_C (Turnover_{unit,t-2} \text{ x } Weighting\ factor_{unit,t-2}) + \sum_D (\overline{Turnover_{stratum,t-2}} \text{ x } Inclusionweight_{stratum,t})}$$

where
- $Turnover_{unit,t-2} \text{ x } Weighting\ factor_{unit,t-2}$: the weighted turnover of a unit in [t-2]
- $\sum_C (Turnover_{unit,t-2} \text{ x } Weighting\ \ factor_{unit,t-2})$: the turnover of the core cell in [t-2] for part C
- $\sum_D (\overline{Turnover_{stratum,t-2}} \text{ x } Inclusionweight_{stratum,t})$: the estimated turnover of the core cell in [t] for part D
- *stratum*: core cell x size class
- P_{new}: a parameter to assign a different weight to new units. As we do not have historical data for new units, we consider them to be more important. That is why P_{new} is normally larger than 1. We have started with default value 3.

41. If we consider all statistics to be equally important, the prioritization has to be executed at the highest level. After calculating *CR* and *RR* for all units and core cells, the actual rating is assessed as follows.

1. Assess for each core cell the non-response unit with the highest *CR*

2. Calculate the priority-score of all units from step 1 as follows: priority-score = (1-*RR*) + *f* .*CR* (where *f* is a parameter, which has a default value of 1).
3. Compare the priority-scores found in step 2, and assign rating 1 to the unit with the highest priority-score (let us call this unit *1* from core cell *a*)
4. Now, calculate the priority-score of the second important non-response unit from core cell *a* (unit *2*), regarding unit *1* as response.
5. Compare the priority-score of unit *2* with the priority-scores of all other units from step 2, and assign rating 2 to the unit with the highest priority-score.
6. Repeat steps 3 to 5 until all units are rated.

42. Depending on available capacity, a proportion of the units with the highest ratings are contacted by telephone. All other non-respondents will first receive a reminder by post. After 12 working days, all remaining non-respondents are considered for contacting by telephone.

43. In some other countries, all units are first reminded by post, and after a certain period of time, the most important remaining non-respondents are contacted by telephone. This seems a “cheaper” approach, as some of the non-respondents Statistics Netherlands contacts by telephone, are also likely to react to a written reminder. Seeing that we want to obtain a fairly steady growth rate as soon as possible, Statistics Netherlands has chosen this method. An evaluation study would be very recommendable to assess whether or not the chosen approach leads to higher response rates and more steady index figures.

C. Plausibility indicator and selective editing

44. Application of selective editing can reduce editing costs and contribute to improved timeliness. On the other hand, not editing of important errors is a risk that needs to be taken into consideration. The success of selective editing stands or falls with making the right selection. For this purpose, a plausibility indicator has been developed, taking two aspects into account: contribution and growth rate.

45. ***Contribution***: For each record, the contribution to the estimate (weighted turnover of the record divided by the total turnover of the core cell) is calculated. However, the estimate for the reporting period (referred to as [t]) has not been established yet whereas the estimate of [t-1] is not yet stable. Therefore, the most recent estimate of [t-2] is used as reference, with a correction for the number of working days and –if necessary- for seasonal effects. The contribution is calculated for both [t-2] and [t], after which the maximum value is used for the comparison with a preset threshold. If this value is higher than the threshold, the plausibility indicator on contribution PI_C is assigned the value 1, else it is assigned the value 0.

46. ***Growth rate***: The growth rate is defined as the turnover ratio of [t] and [t-1] of the record. In case of non-response in [t-1], a previous period is used. If there is non-response for more than six months, the median level of the publication cell is used as reference. If the growth rate scores above the upper limit g_{max} , or under the lower limit g_{min} the plausibility indicator on growth rate PI_G is assigned the value 1, else it is assigned the value 0.

47. ***Selection***: The selection is based on the plausibility indicator PI: $PI = PI_C \; . \; PI_G$. If $PI = 1$, the record has to be edited interactively, else ($PI = 0$) the record is edited by computer. This implies that a record has to score both on contribution and on growth rate. There are however some exceptions. Firstly, all records that contribute more than 10% to the estimate are selected for interactive editing, regardless the growth rate. Secondly, records from companies with more than 250 persons employed, are considered to score above the limit on contribution, which means that in these cases only the growth rate determines the selection.

48. ***Threshold for contribution*** : The scores of respondents for period [t-j] of the core cell are sorted by their contribution. The limit for contribution is set at the median level, which means that records that represent a contribution above the contribution of the middle-most record, are assigned $PI_C = 1$.

49. ***Threshold for growth rate***: The growth rates of respondents for [t-j] / [t-j-1] for each core cell and size class group are established and the records are sorted accordingly. Then, g_{max} is set at the 95th percentile score and g_{min} at the 5th percentile score. Note that the threshold for growth rate is defined per core cell and size class group whereas the threshold for contribution is assessed per core cell. In this way we try to include smaller companies with implausible growth rates in the selection.

50. As there is little experience with selective editing of short-term statistics at Statistics Netherlands, the thresholds for the contribution and growth rate as described above have been set in such a way that a relatively large proportion of records (about 40 %) is currently being selected for interactive editing. Since the fourth quartile of the records ordered by contribution already represents about 80% of the total turnover, the interactively edited records will probably represent at least 90% of the total turnover.

D. Imputation

51. ***Creation of panel pairs***. The imputation for non-response is based on the growth rate of so called "panel pairs". A panel pair is formed by the scores of a sample unit that has responded in both [t] and [t-1]. It is important for the stability of the estimate, to have a sufficient number of panel pairs. The difference in growth rate is expected to vary more with small companies than with large companies. For this reason, more panel pairs are needed in smaller than in larger size classes. Furthermore, the number of panel pairs needed is related to the variance of a cell, which is indicated by the number of samples and fraction. Therefore, a threshold value V_t for the minimum number of panel pairs is defined: $V_t = 1 + \sqrt{S_k}$, where S_k is the number of samples in imputation cell k. If there are insufficient panel pairs in cell k for the calculation of the growth rate, cell k is joined with neighbouring cells in the same size class-group. If even the size class-group contains insufficient panel pairs, this group is joined with the corresponding size class group of a core cell that is part of the same publication cell for the preliminary results.

52. ***Outlier detection***. Panel pairs with extremely high or extremely low growth rates, disturb the imputation and need to be detected. For the calculation of the growth rate that is used for the imputation, these outliers will be eliminated. The detection of outliers consists of the following steps.

53. First, the growth rate per panel pair is assessed for the individual or joined imputation cell. Second, the median growth rate for the cell is calculated. Third, for each panel pair the residual of the individual growth rate and the median is calculated as $res_i = g_i - median(g_j \mid \forall_{j \in c})$ where res_i represents the residual for panel pair i, g_i its growth rate and $median(g_j \mid \forall_{j \in c})$ the median value of the growth rates of all panel pairs in cell c. Fourth, the cut-off value z is established as $z = upper\_f + a.(upper\_f - lower\_f)$ with $upper\_f$ the third quartile score and $lower\_f$ the first quartile score and a a parameter (initially we have used 0.5). Finally, a panel pair is considered an outlier if the absolute value of the residual is larger than the cut-off value z. Note that these outliers are only eliminated for imputation purposes; there is a separate outlier detection in the weighting process.

54. ***Imputation***. Three different ways of imputation for non-response can be applied, depending on available information.

55. First, if there is a (reported or imputed) score for the same unit and variable from the previous period ($V_{i,t-1}$), the imputed score for a variable V for non-respondent i in period t in imputation cell c with growth rate G_c is calculated as $V_{i,t} = G_c . V_{i,t-1}$. The growth rate G_c is the average growth rate of the panel pairs in imputation cell c, after elimination of outliers.

56. Second, if unit I has not been in the sample in a previous period, $V_{i,t-1}$ will not be available. If available, VAT-information might be used in this situation. In some industries, there is a difference between fiscal and statistical turnover. When using VAT-data, this difference has to be taken into

consideration. The imputation in this situation can be written as $V_{i,t} = B_{i,t} \cdot \frac{\sum V_{c,t}}{\sum B_{c,t}}$ with $B_{i,t}$ the VAT-score for unit *i* in period *t* and $\frac{\sum V_{c,t}}{\sum B_{c,t}}$ the ratio between the sum of scores of variable V in cell *c* and the sum of scores in the same cell of the corresponding VAT-variable. As there are still some methodological difficulties to be solved, this imputation method has not been put into practice yet.

57. Third, if neither reported or imputed scores for the previous period nor VAT-information is available, the average score of the imputation cell is imputed.

58. ***Editing imputed values***. After imputation, inconsistencies may occur in records where more than one variable is involved. The final step of the imputation phase is consequently the automatic editing of imputed records.

IV. WORK AHEAD

59. The implementation of the new system and process has started in the second half of 2002 with the statistics on commercial services. Every successive month and quarter, new industries were added, until the last short-term statistic (quarterly statistics on wholesale trade and transport) was added in the first quarter of 2003. Although the production process is now in use, some aspects of the system need to be evaluated.

60. First, the data collection process. The results of the prioritization of the follow up, indicated by the algorithm that has been developed for this purpose, need to be evaluated. Does the algorithm select the most important non-respondents? Is the chosen approach (important non-respondents are contacted by telephone instead of by post) effective and efficient; does it lead to higher response rates and to more steady growth rates shortly after the reporting period?

61. Second, the editing process. As stated above, a rather large proportion of all records is edited interactively. A further enquiry into the selection, the calibration of the thresholds of the plausibility indicator and the effects of selective editing on the estimates is vital to assess whether goals have been achieved and where improvements can be made.

62. Third, the analysis process. The imputation of non-response has proven to be somewhat questionable in some cases for new sample units. Further, changes in the business register forced us to make some adaptations in the weighting process. Both these aspects need to be evaluated.

V. ORGANISATIONAL DIMENSIONS

63. At Statistics Netherlands, two departments and four sub-departments are involved in processing the short-term statistics. The data collection and editing is carried out by the business surveys department whereas the analysis and publication is the responsibility of the statistical analysis department. Prior to the implementation of IMPECT2, the entire process was carried out by a single department. The new organisational structure has been formed to improve efficiency through process specialization: the knowledge of different processes is concentrated in separate departments and sub-departments. However, this implies not only advantages. The communication between the departments needs to be excellent. Besides, as only implausible records are presented to the data editor, he or she lacks the overview of the entire data set. It is difficult to take the right edit decision, if the effects on the estimates cannot be seen. These disadvantages have been recognized and might lead to a reconsideration of the division of responsibilities between the departments involved.

VI. CONCLUSIONS

64. As stated before, the key objectives of the new system for processing the short-term statistics are harmonization and improvement of efficiency. Harmonization of variables, methodology and processing has led to more transparency. Results have to be reproducible at any time. Furthermore, the manageability has been improved and the maintenance costs have been reduced. Selective editing and use of VAT-data have contributed largely to the improved efficiency of the entire process. Nonetheless, some comments need to be made.

65. The implementation was accompanied by some difficulties that could partly have been avoided. Due to the limited amount of time for development and testing, the system was implemented although it had not been tested completely with realistic data. As a result, the start up period overran its time. Lack of capacity, the introduction of the SBS-size class and a changed survey design also prevented the processing of data with both the new and the previous system. Therefore, the accuracy of the new results could not be ascertained by comparing them with the results according to the previous system, so other sources needed to be used for this purpose.

66. As described in chapter IV, some aspects of the system need to be evaluated and some improvements need to be made. Mistakes are inevitable and were indeed made at some point. However, a leap forward has been made by implementing a uniform methodology, system and way of processing the short-term statistics. We are convinced to be on the right track; what we need for the future is staying power.

REFERENCES

This paper is based on documents in the Dutch language, produced within the framework of the project. Therefore only a general reference can be made.

Kuurstra, D.A., 2002, *IMPECT2: interactief of automatisch gaafmaken, de selectie* (interactive or computerized editing: the selection), Statistics Netherlands, Voorburg.

Valkenburg, J.J.M., 2002, *IMPECT2 FO Ophogen* (functional design weighting process), Statistics Netherlands, Heerlen.

Valkenburg, J.J.M., 2002, *IMPECT2 FO Imputeren* (functional design imputation process), Statistics Netherlands, Heerlen.

Velzen, J.H. van and W.H. Vosselman and others, 2002, *Handboek productie KS-en* (companion short-term statistics), Statistics Netherlands, Voorburg.

Visschers, J.W.C.H., 2003, *IMPECT2 FO Rapelleren-Appendix A* (prioritizing follow up), Statistics Netherlands, Heerlen.

APPENDIX A

Statistical period

In this paper [t] is used to indicate the most recent statistical period. Depending on the survey, the period is either a month or a quarter. For example: in July, we refer to the data of June and the second quarter as [t], and to May and the first quarter as [t-1]. The indication [t-j] is used to refer to the same period of the previous year.

Different types of cells

In the description of the statistical process, the following cells are mentioned.

- Core cell: aggregate of one or more NACE-classes (4-digit groups) or sub-classes. The editing process has been optimized on this level.
- Publication cell: aggregate of one or more core cells.
- Size class: company size, based on the number of persons employed (SBS-size classes 10 to 93)
- Size class group: aggregate of successive size classes. We distinguish three size class groups: less than 10, between 10 and 100, 100 and more persons employed.
- Stratum: core cell x size class.
- Imputation cell: initially stratum. If necessary extended to core cell x size class group or eventually to publication cell x size class group.
- Weighting cell: aggregate of one or more core cells x size class group. If necessary extended to publication cell x size class group. If VAT-data is used in the weighting process, a weighting cell is split in two parts. One part consists of records for which VAT-data is available, the other part comprises all other records.