

**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

AUTOMATED EDITING OF ROAD TRANSPORT DATA

Supporting Paper

Prepared by Peter Kruiskamp and Saskia Ossen, Statistics Netherlands¹

I. INTRODUCTION

1. The process of editing road transport data at Statistics Netherlands is very laborious. This is firstly due to the complexity of Road Freight Statistics. Road Freight Statistics consists of many variables with various dependencies. Secondly the questionnaires are presently large and complex, which makes it difficult for respondents to comprehend what should be reported, and subsequently respondents have difficulty to respond correctly and completely. Therefore many of the obtained questionnaires need to be edited. Furthermore, effort can be saved by improving the efficiency of the editing process. Data editing currently takes place at several different points in the statistical process and requires many manual operations.

2. Statistics Netherlands has therefore started a project to redesign the processes to create Road Freight Statistics. This among others includes a new and more efficient design of the data editing process, in which many of the editing procedures will be automated. The questionnaires will also be redesigned in order to reduce the amount of editing, which is however out of the scope of this paper. Also the development of a macro-editing tool, which will be used to check plausibility of the aggregated data is out of the scope of this paper.

3. In this contribution the proposed automated micro-editing procedures will be discussed. First, some background information about Road Freight Statistics is given in section II.A, and the present editing procedures are described in section II.B. Then a lay-out of the redesigned editing process is given. Section III.A deals with procedures for automatically selecting the records that are suitable for automated editing and procedures for automatically checking the plausibility of the micro-data. In section III.B the micro-editing itself is described in detail consisting of procedures for automatically correcting implausible values, imputing missing values and deducing variables that are not part of the questionnaire, but are still needed to produce the necessary output.

II. PRESENT SITUATION

A. Background

4. An important goal of Road Freight Statistics is to report the transport of goods by road. The questionnaire consists of three components, namely data on the enterprise and the vehicle, data on the journey and data on shipments within the journey. For a vehicle in the sample, up to a maximum of 36

¹ The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

variables are requested per shipment in a specific week. Many of the required variables are not straightforward for the respondents. Analysis of received questionnaires reveals that questions often are misinterpreted, leading to false information. Another common practice appears to be that respondents report journeys with single shipments even if they actually carried several shipments. Furthermore, questionnaires are often returned incomplete or with implausible data.

5. The problems of misinterpretation and incomplete data should be prevented at the source, the respondent. As mentioned before, a project to redesign the questionnaires is already in progress. This so called 'e-questionnaire' will be taken in production in 2010. In another project, which is in an advanced stage, variables for Road Freight Statistics are automatically extracted from accounting systems of transporters by means of XML (Extendible Markup Language).

6. Even with these developments, inconsistencies will still occur frequently, and editing procedures will continue to play an important role in order to generate plausible output from the received data.

B. Present editing procedures

7. The output of Road Freight Statistics is used by a number of clients, which all have their own specific needs. Eurostat is one of the most important users of the Road Freight Statistics data. Eurostat not only uses aggregated data, but also receives micro-data, which they use to generate their own statistics on European level. Therefore, it is important to Eurostat that the micro-data are complete and of good quality. To ensure this, Eurostat enforces many constraints on the micro-data (Eurostat, 2008). Besides the so called 'Eurostat variables', a large number of other variables are being used to produce output for the other clients. All these variables need to be edited, and presently this is being done in a manner that has historically grown. This is described in detail in Kruiskamp and Ossen (2008), and is schematically presented in figure 1:

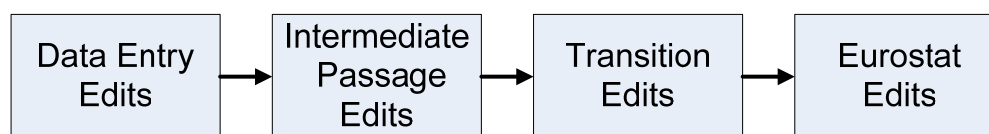


Figure 1. Actual editing scheme for the micro-data of Road Freight Statistics.

8. Each of these stages contains its own checks and editing rules, specific for the purpose of the edits. This will be described in further detail in the next subsections.

B.1 Data Entry Edits

9. Presently many of the questionnaires are still being submitted on paper. Especially the transporters on own account still mainly use paper forms (about 75% of the received forms are submitted on paper), but also 25% of the respondents that transport on hire or reward submit their forms on paper. Typically, respondents omit questions that they feel do not apply to them, or that they think that are not important.

10. These paper questionnaires are all entered manually. During this data entry procedure, many errors and omissions are detected and corrected and imputed 'on the fly'. This is not only very time-consuming but it also requires specialist insight, and is difficult to control. Data editors are needed to enter the paper questionnaires, and their edits are not logged, other than on the paper questionnaires themselves. Also, due to the complex structure of processing the data, one tends to over-edit the data to ensure that fewer conflicts will arise further on in one of the editing procedures. The emphasis of editing in this stage lies on obtaining complete, logical and consistent questionnaires.

B.2 Intermediate editing passage edits

11. At this stage the electronic questionnaires enter the editing process. The editing rules at this stage are mainly focussed on providing valid values of the variables. The following checks are done:

- Are the values known?
- Do known values belong to the specified domain?
- Are values of variables consistent with each other?
- Are the values valid?
- Are the values realistic?

12. Most of the electronically submitted questionnaires do not satisfy all checks, and these questionnaires are edited manually. In practice, not only the so called ‘hard’ errors are edited, but also warnings about unusual high or low values (‘soft’ errors) are normally corrected, to prevent possible problems at further stages.

B.3 Transition edits

13. At this stage the records are transformed into a ‘generic’ record. The editing rules are specific for this ‘generic’ record. However, many of the checks at this stage are very similar to the checks at the previous stage. At this stage however, suggestions are provided automatically when checks are not satisfied. Due to slight differences between the checks at this stage and the previous stage, errors are sometimes detected on values that passed earlier checks. This is the reason that one tends to over-edit the data at previous stages. In practice the amount of errors that are detected at this stage is relatively small.

14. When all checks are satisfied, missing empty journeys are imputed. An empty journey means that the vehicle has travelled without any goods, for instance to a next place of loading. In many cases, respondents do not report these journeys since they do not think that they are important for Road Freight Statistics. However, these data are necessary to calculate the amount of traffic of transporters on the Dutch roads. For domestic journeys, empty journeys can be derived automatically, but for international journeys the necessary empty journeys are created manually.

B.4 Eurostat edits

15. The micro-data that is meant for Eurostat are transformed into a specific format, imposed by Eurostat. The checks that are performed at this stage are identical to the checks that Eurostat itself also performs on the data. Many of these checks are technical checks, and are related to data-processing by Eurostat. The other checks are again very similar to the checks at the previous two stages, and usually do not cause ‘hard’ errors (with the exception of a few specific checks). The ‘soft’ errors that occur at this stage are generally not edited.

B.5 Overview

16. Overall the present process of editing the submitted questionnaires is elaborate and requires a lot of specialist insight from data editors. Furthermore, a number of checks are performed more than once, and sometimes with slightly different constraints. This is not only time consuming, but also is a irritation factor for data editors, who as a consequence tend to over-edit the data, in order to prevent checks to fail further on in the editing process. In addition, because most of the editing procedures are performed manually, it is hard to guarantee that the editing process is always performed in a consistent manner. For all these reasons it was of vital importance that the editing process would be redesigned.

III. REDESIGN OF THE DATA-EDITING PROCESS

17. The main purpose of the redesign of the editing process is to create a process where editing is done in a consistent, reproducible and efficient manner. To ensure this, as much as possible should be edited automatically. The redesigned process is described in detail by Kruiskamp and Ossen (2009a) and Kruiskamp and Ossen (2009b). The schematics of the procedures of this process is shown in figure 2:

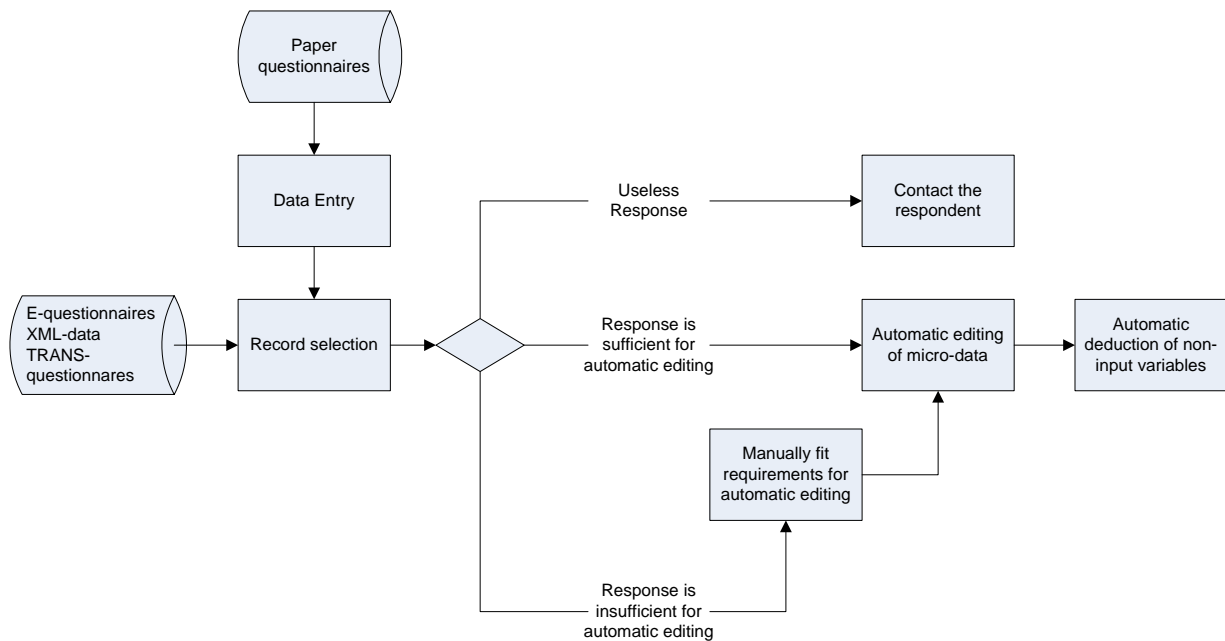


Figure 2. Redesign of the editing procedures for the micro-data of Road Freight Statistics.

18. The automated stages of these redesigned editing procedures will be described in further detail in the next sections. In section III.A it is discussed how records can be automatically selected for automatic editing. Section III.B deals with the automatic editing and deduction of micro-data.

A. Record selection

19. As soon as the unedited data are electronically available, records that are suitable for automatic editing can be selected, based on the quality of the response. Furthermore, records which are not yet suitable for automatic editing, but can be manually edited such that they do fit the requirements for automatic editing will be selected. What remains are records that are not directly usable, and for which contact with the respondent is necessary in order to process these data.

20. In order to be able to perform this selection automatically, the following steps need to be taken. First the quality of the data needs to be defined. How this can be done is described in subsection III.A.1. Furthermore, explicit selection criteria need to be defined to be able to perform the actual record selection. This is discussed in subsection III.A.2.

A.1 Determination of the quality of the variables

21. An extensive set of checks has been derived from the existing data-editing stages (Kruiskamp *et al.*, 2009). Double checks and redundant checks have been deleted, and inconsistencies between checks of the different stages have been resolved. For every check, a corresponding action has been assigned. When values of variables prove to be improbable, these values are removed to be automatically deduced further on in the process. When conflicts arise between variables, it is automatically determined which value is plausible, and which is not. Only when automatic actions are not feasible, variables are marked to be edited manually.

22. For example, several checks are performed on the variable ‘distance travelled for the journey’. The distance travelled for a journey consisting of one shipment is for instance checked with the calculated distance between the variables ‘place of loading of the goods’ and ‘place of unloading of the goods’, travelled on roads suitable for goods transportation. This data is provided by a route planner. When the reported ‘distance travelled for the journey’ deviates significantly from the calculated distance, while ‘place of loading of the goods’ and ‘place of unloading of the goods’ are considered probable, the reported distance is likely to be wrong, and will be deleted. However, when the reported distance is a whole lot higher than the calculated distance, it is more likely that the ‘type of journey’ is improbable. In

that case it is not likely that the journey contained only one shipment, and the value 'single shipment' should be deleted.

23. The result of the checks and corresponding actions is that the variables of the records either contain probable values or are empty. Therefore, the quality of the records can be determined by examining which variables are filled. In the next subsection, it is shown how this information can be used to select the records that qualify for automatic editing.

A.2 Selection criteria

24. A questionnaire is suitable for automated editing when the variables that are important for the output are plausible, or can be imputed automatically. The variables that should at least be available in order to produce reliable output have been determined after deliberation with experts on Road Freight Statistics. As a result the following variables were assigned essential to produce Road Freight Statistics:

- Place of loading of the goods
- Place of unloading of the goods
- Distance travelled for the shipment and/or the journey
- Type of goods that are transported
- Weight of the goods

25. These variables are the essential purpose of Road Freight Statistics: *How much of what goods are transported from where to where, and what distance is travelled?* Therefore, in order to be able to automatically edit a questionnaire, its records should at least contain these so called essential variables, or otherwise it should be possible to deduce these essential variables automatically.

26. This means that the quality of a record is directly related to the availability of these essential variables. The selection criteria for automatic editing therefore focus on these essential variables. More specific, to select records that are of sufficient quality for automatic editing, checks are performed on the essential variables. When one or more essential variables prove to be improbable, checks are performed on the variables that are necessary to automatically deduce the above-mentioned improbable essential variables, to establish whether deduction of these variables is possible. In this manner the possible availability of the essential variables can be determined. When it is uncertain if the essential variables can be deduced, the quality of the questionnaire will have to be judged manually. If possible, the essential variables will be imputed manually. When all essential variables are available, the questionnaire is actually suitable to be edited automatically. This process of record selection is schematically described in figure 3:

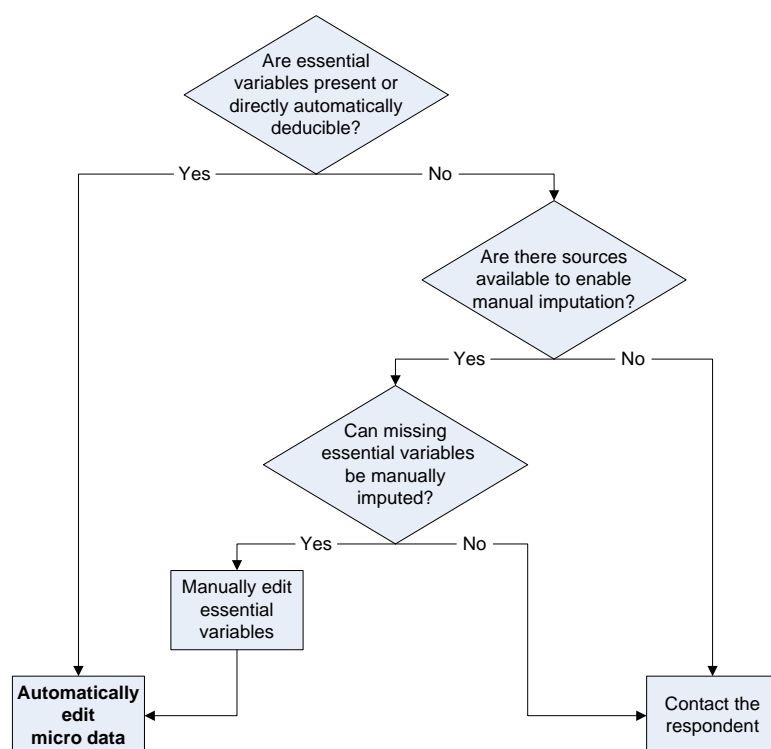


Figure 3. Decision diagram with the necessary actions for record selection.

B. Automatic editing and deduction of micro-data

27. Automatic editing of the questionnaires consists of two separate steps. First, the plausibility of the values of the remaining variables (all variables besides the essential variables) has to be determined. This is already discussed in subsection III.A.1. Again improbable values are deleted. Consequently the variables of the records either contain probable values or they are empty.

28. The next step is to automatically deduce the remaining variables, if possible. This applies to the variables which are used to produce output and that were emptied in the previous step. Furthermore, variables that do not appear in the questionnaire but are nevertheless used to produce output will be deduced in this step. In total 66 variables are used to create all output for Road Freight Statistics. For each of these variables one or more deductions are described in Kruiskamp *et al.* (2009). Input for these deductions are other variables that contain probable values, as well as information from external sources such as the Dutch Vehicle Registration, the Dutch Enterprise Register and a sophisticated route planner.

29. When more than one possible deduction is described for a specific variable, the consideration which deduction is used depends on several conditions. The different deductions are ordered based on their estimated quality. However, it is possible that some of the input information necessary for the most qualitative deduction is not available or is of poor quality. In that case a less preferred deduction can still provide the required information.

30. When none of the described deductions produce satisfying results, a couple of possibilities remain. A number of variables do not need to be filled if deduction of the variable does not lead to a probable value. For other variables, more complex imputation procedures can still provide plausible values. For example, historical data or nearest neighbour information can be used to impute these values. Only when all previous imputation efforts fail, due to the fact that input information for the automatic imputations is not available, variables will have to be edited manually in order to obtain the necessary micro-data.

31. As an example of a variable that is required to be filled, the deduction of ‘distance travelled for the journey’ is shown in figure 4:

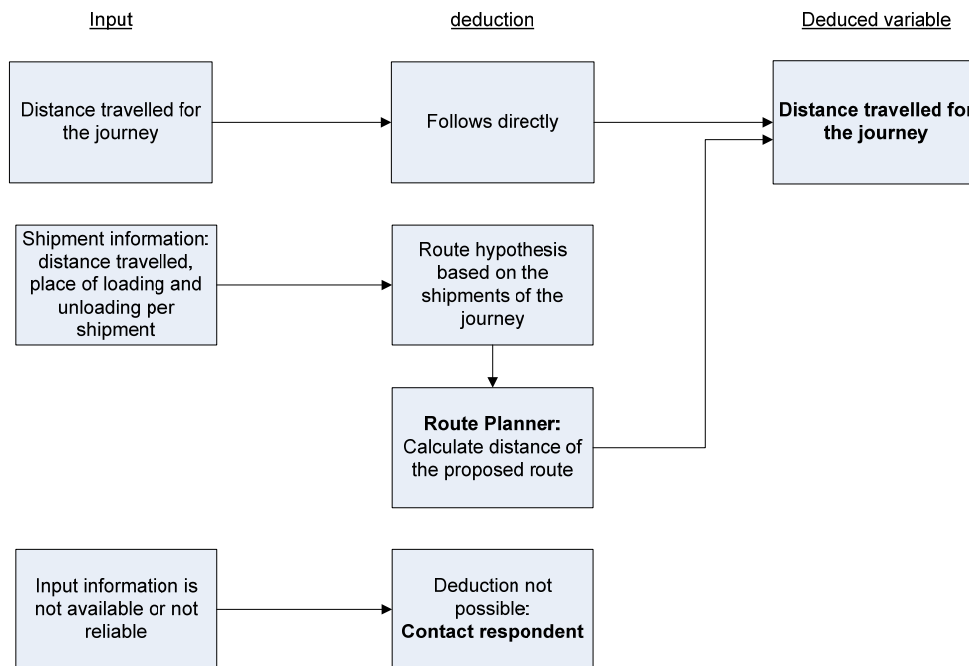


Figure 4. Deduction scheme for the variable 'distance travelled for the journey'.

32. When the input variable 'distance travelled for the journey' is considered probable, it is clear that no deduction is necessary. Otherwise, shipment information is used to obtain the value of the variable. A hypothesis is formulated in what order the shipments were transported, and what the journey's most likely route must have been. When the journey consists of only one shipment, the solution is of course straightforward. When several shipments are involved, a sophisticated algorithm calculates the most probable route. In the next step a route planner calculates the distance travelled for this route. When both the input variable 'distance travelled for the journey' and shipment information are missing, it will be checked if shipment data can be derived. If this is not the case, the distance travelled for the journey cannot be deduced. Consequently, the respondent has to be contacted to provide additional information.

IV. CONCLUSIONS

33. In this paper the redesigned process of editing reported data on transport of goods by road is described. The editing procedures have been automated as much as possible, which is to provide a considerable efficiency gain, and ensures a high level of consistency and reproducibility of the data-edits. To achieve this, criteria for automatically selecting records that are of sufficient quality for automatic editing have been specified. Furthermore, automatic checks on values of variables have been defined. In case a check contains more than one variable, actions have been formulated specifying exactly which of these variables are more likely, and therefore have to be edited, and which ones will remain unchanged. To derive values for variables failing the checks and to derive values for variables that do not appear in the questionnaire, automatic deductions have been formulated.

34. In general these improvements are expected to significantly enhance the quality and efficiency of Road Freight Statistics. Taking into account other favourable factors such as the development of a better e-questionnaire and that the use of XML data is stimulated, it is expected that incoming data will not only be of better quality, but that also the effort needed for correcting wrong values will decrease.

References

Eurostat (2008), *Road freight transport methodology*. Reference Manual for the implementation of Council Regulation No 1172/98 on statistics on the carriage of goods by road, Luxembourg.

Kruiskamp, Peter, Saskia Ossen (2008), *Inventory of micro-editing stages of road transport data* (in Dutch). Internal Unpublished research paper, BPA-number DMH-2008-11-15-PKRP, Statistics Netherlands, Heerlen.

Kruiskamp, Peter, Saskia Ossen (2009a), *Proposed redesign of data-editing for Road Freight Statistics* (in Dutch). Internal Unpublished research paper, BPA-number DMH-2009-03-10-PKRP, Statistics Netherlands, Heerlen.

Kruiskamp, Peter, Saskia Ossen (2009b), *Automated stages of the redesigned editing strategy for Road Freight Statistics* (in Dutch). Internal Unpublished research paper, BPA-number DMH-2009-05-19-PKRP, Statistics Netherlands, Heerlen.

Kruiskamp, Peter, Saskia Ossen, Rogier Hellenbrand (2009), *Rules for automated checks and deduction of variables for Road Freight Statistics* (in Dutch). Internal Unpublished research paper, Statistics Netherlands, Heerlen.