Notes on Census Editing

Griffith Feeney
September 1999

1 Introduction

Census editing may be broadly defined as the process of scrutinizing of census data at various stages of processing for omissions, inconsistencies and anomalies and taking appropriate action when any of these conditions is found. Editing serves numerous purposes, including quality control, making data more convenient for users, and avoiding the dissemination of obviously incorrect data.

Editing occurs at several key stages of census processing. Edit checks will generally be made by field supervisors during enumeration to provide quality control over the work of enumerators, by central office staff involved in coding and data capture, and by computer programs toward the final stages of processing.

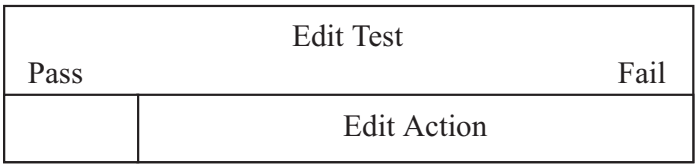
There are important differences in the nature of editing at different stages of the process. These notes focus on computer editing procedures implemented during and after data capture. To provide perspective and to clarify the nature of the editing operations, however, it is useful on occasion to consider editing at earlier stages.

Census editing and survey editing are similar in many respects. Censuses generally contain fewer questions, often far fewer, and vastly more records, often by several orders of magnitude. Census edits tend therefore to be simpler, and speed of processing is a more important concern. Much of what is said below applies to survey as well as census editing, but these notes focus on census editing only.

2 Edits

The editing operations carried out during the course of census processing divide into several stages. With occasional exceptions, each stage consists of a sequence of individual edits, executed one after the other on each record processed.

An *edit* may be defined as consisting of two components, a precisely specified, undesirable condition that may be observed in the data, and an action to be executed whenever this condition is found. The first of these two elements is called an *edit test*, or "check", or "rule". The second element is the corresponding *edit action*. Thus an edit may be represented diagrammatically in the following way (Chapin 1974).



Edit tests and actions may be executed manually or by a computer program. Editing procedures tend to be manually executed in the early stages of processing and computer executed in the later stages.

This definition of "edit" departs from some prior usage. Sunder, Patrick and Binder (1975), for example, explicitly define "editing" as "the identification of data fields judged to have a high probability of error" and "imputation" as "the adjust-

ment of these fields to values judged to contain (probably) smaller errors". Felligi and Holt (1976) follow this usage.

Usage is not consistent, however, for the Sunder *et. al.* paper, which is concerned with imputation as much as with editing by these definitions, is titled "On the editing of survey data."

This example illustrates a common bifurcation of meaning, in which "editing" sometimes refers only to testing for problems and sometimes to testing and (attempting to) fix problems.

Several considerations support the terminology proposed here.

- The usage of the word "edit" in connection with data editing derives from "copy editing" of manuscripts for publication, which includes correcting as well as identifying errors.

- The purpose of performing an edit test is usually to decide whether some corresponding edit action should be performed. It is natural and appropriate to consider the action together with the test, subsuming both under a single designation.

- Imputation is only one type of edit action. Edit tests at one stage of processing may correct errors by referring to the previous stage of processing, as when data entry errors are corrected by reference to the original questionnaires or code sheets.

- The more general definition of "edit" used here is useful in discussing the process of editing as a whole, including editing at all stages of census processing.

3. Edit Tests

With few if any exceptions, every edit test specifies, first, one or more items of information contained in the census records, and second, values or combinations of values for these items that are invalid, undesirable or suspect.

- *Missing value tests* look for missing information that should be present. Missing values should not be confused with information that is "not applicable" because the entity to which the record refers lies outside the domain of the attribute (*e.g.*, age at first marriage for a never married woman. Values are literally "missing" in both cases, but in the latter case they are *supposed* to be missing, whereas in the former case they are *not* supposed to be missing.

- *Invalid code tests* look for invalid codes in computer records, e.g., a code of "7" where there should be a code of "1" representing male, or "2" representing female. Invalid code tests will generally be made as part of the data capture process, in which context it may be possible to consult the original record to find the correct value. If it is not possible to correct an invalid code in this way invalid codes are effectively the same as missing values.

- *Consistency or conflict tests* look for inconsistent values, e.g., persons whose records indicate that they are, on the one hand, three

years old and, on the other hand, are grandmothers of the head of the household in which they are enumerated.

- *anomaly tests*, which look for combinations of values which are not strictly inconsistent, but which are very improbable, e.g., 13 year old widows.

Edit tests that involve one or more items from a single record are called *intra-record tests*. Tests that involve items from two or more records are called *inter-record tests*. Inter-record tests may involve (i) items from records of the same type (*e.g.*, all referring to persons in a household) or (ii) records of different types (*e.g.*, a household record and the corresponding person records).

One may for example check all person records in a given household to ascertain that one and only one person has been identified as household head, which requires looking only at person records; or one may ascertain whether the number of person records present following a household equals the number of persons specified in the household record, in which case one requires the household record as well as the person records.

Most persons enumerated in a census are located within households, which are in turn located in enumeration districts, which are in turn parts of lower level units of a more or less well defined administrative hierarchy. Census taking involves the processing of records for entities at each level of this hierarchy, person, household, enumeration district, and so on. Deciding how to handle exceptions to the general rule, such as homeless persons, or geographical entities such as national parks, is one of the tasks of the census taker.

Structure tests are inter-record checks designed to insure the integrity of the records of this hierarchy created by census definition and by national administrative geography. There are several types of structure tests.

- *Counting tests* ascertain whether the number of records of a given type is correct, e.g., by comparing the number of persons in a household indicated by the household record with the number of corresponding person records.
- *Duplication tests* look for multiple copies of the same record.
- *Existence tests* check for the existence of a record of a given type, e.g., test whether every set of person records comprising a household includes a record for the household head.
- *Uniqueness tests* check for the existence of more than one record of a given type when there should be only one, e.g., a test for more than one person coded "head of household" in a household.
- *Sequence tests* check that stipulated sets of records occur in the prescribed order, e.g., that the record for the head of household occurs first, the record for spouse of head of household second, etc.

4. Edit Actions

Edit actions are of two principal types. The first type aims to correct, adapt or improve the process that generated record that failed the test. A radical example would be the reassignment or dismissal of an enumerator who consistently botches

the completion of questionnaires. Actions of this kind tend to occur in the earlier stages of processing and to be executed by people rather than by computers.

The second type of edit action aims to change the record that failed the corresponding test in such a way that the record would pass the test if the test were immediately re-executed. Actions of this kind tend to occur in later stages of processing and are more likely to be executed by computers than by people.

A third, subsidiary type of edit action, that be executed together with either the first two types of action, is to record the failure of a test in a log file.

Edit actions that change the record that failed the test divide into two groups, corrections and imputations. These are discussed in the following two sections.

5. Back checking and Correction

To "correct" an item means to replace an incorrect or missing value with a correct value.

Correction in census editing occurs as a result of a process that may be called "back checking". Census processing involves the movement of information from one medium to another, *e.g.*, from the spoken word of the respondent to the written marks on the census questionnaire, from the marks on the questionnaire to marks on central office coding forms, and from questionnaires and coding forms to digital media.

Back checking may be defined as the process of checking output information against input information. When an edit test detects an invalid code immediately following data entry, for example, the value in the computer record may be compared with the value on the form from which it originated. If the values do not agree, the output value on computer record will be changed to agree with the input value on the form. This is a correction in the sense that it corrects an error introduced by the process of data entry, though it is of course possible that the value on the input form is wrong.

6. Imputation

Edit tests will occasionally fail, however, even if all census processing operation are executed perfectly, for despite the best efforts of census takers, respondents will not always supply all the information requested, or will not supply it accurately, whether because they are unable or unwilling to do so.

To eliminate missing values and resolve contractions then requires assigning values that we have no way of knowing. *Imputation* is defined in this context as the assignment of values according to rules that provide some assurance that the aggregate statistical properties of the census data are preserved. To "impute" an item means to assign a value in such a way that the statistical integrity of the data set is preserved, but without the expectation that the value imputed for any particular record will be, except by chance, the true value.

7. Rationale for Imputation

A generally recognized and fundamental principle of census taking is that census data should be presented to users as it is collected, without adjustment of any kind. There are several reasons for this self-imposed discipline.

- Adjustment of any kind creates a risk that what begins as a well

intentioned service to users devolves into something less benign. Even if the reality of this is remote, suspicion of the census taker by policy makers and the general public may be as damaging as actual misconduct. Anything that resembles tampering with the data incurs a risk.

- Adjustments may be done poorly and/or may be inadequately documented. Even if an adjustment is done well by the standards of the time, new and better methods are likely to be developed in the future. Adjustment prior to release of the census data is likely to foreclose the option of applying these new and better methods.

- When census data are adjusted, the original unadjusted data tend to be lost to the general census user, even if available within the government office that conducted the census.

There are nonetheless compelling arguments for some imputation in census data processing. The simple example of what to do with census records for which sex is not stated will serve to illustrate.

If the value of sex for these records is not imputed, every table produced that shows population disaggregated by sex will require three instead of two rows or columns for that purpose. This will increase the volume of such tables, which will comprise a substantial proportion of all tables, by one half.

This will increase printing costs very substantially. Yet the information supplied by the "sex not stated" numbers will almost certainly be negligible, and in fact the values will be a positive irritant to most users.

When it is further considered that the redundancy of census data usually makes possible a good inference about the sex of a respondent based on other information, such as whether or not the respondent replied to questions addressed only to women, the case for imputing sex when it is not stated becomes very strong.

The case is however particular to the specific item. Age is probably left unreported in census records more often than sex, yet it is very common to see census publications that include "age not stated" numbers in tables involving an age dimension. The more numerous possible values of age make it more difficult to impute than sex, and at the same time greatly reduce the extra burden imposed by including them in census publications.

8. Danger of Imputation

Manual processing of census data, such as was carried out in India until very recently, puts severe practical limits on how much imputation can be carried out on the census returns. With the advent of computer editing programs, this restraint is greatly reduced, and there is a danger that imputation routines will be over used.

The danger of insufficiently discriminating imputation may be illustrated by the case of labor force information collected in the 1971 census of Indonesia, documented by Gavin Jones (1974). Series C of the 1971 census reports showed 0.9 million persons unemployed, whereas Series E of the reports, based on precisely the same data, showed 3.7 million unemployed. The difference was due solely to different imputation procedures used in the two reports.

The imputation procedure used in Series C gave an unemployment rate for

females in West Java of only 2.6 percent, but the procedure used in Series E gave 20.8 percent. The differential was only slightly less pronounced for males.

Imputed data of this kind is obviously seriously misleading. Especially in the absence of detailed documentation, it implies knowledge of reality that does not in fact exist. The information conveyed derives primarily from the imputation procedure, rather than from the replies of the respondents.

It is important not to use the word "correction" when we mean "imputation", a habit that appears to be regrettably easy to fall into. One can never object to census data being "corrected", but the case for *imputing* values that are missing or judged incorrect must be made case by case.

9. Resolving Inconsistencies

To resolve an inconsistency, such as a record indicating that the age of the respondent is three years and that the relation to head of household is "grandmother", it is necessary to change either or both values. Formal resolution may always be accomplished by replacing one or the other reported values by a "missing" code, but is not particularly satisfactory and seems to be rarely done.

Resolution is accomplished, rather, by (i) deciding which of the conflicting values is more likely to be wrong and (ii) imputing this value in such a way as to resolve the inconsistency. Resolving inconsistencies is, on account of (i), a more difficult task than imputing missing values.

10. The Importance of Redundancy

Census information is partially redundant. Some redundancy is inherent in the items included. Knowing that a person is the child of the head of household, for example, tells us something about their relative ages.

Other redundancy is introduced by questionnaire design, in particular by any skip instructions in the questionnaire. If only women are asked about the number of children they have borne and the number currently surviving, for example, the presence of this information on a record suggests that the record refers to a woman.

Redundancy is critical for editing. Absent redundancy there would be no inconsistency of responses, and thus no necessity for consistency checks. Absent redundancy, no item would provide information about any other item and it would not be possible to improve the imputation of missing values for one item by consulting the values of other items.

11. Edit Processes for Intra-Record Edits

We may define an *edit process* for intra-record edits as a sequence of edits intended to be executed one after the other on each record processed. This may be represented diagrammatically as

Edit 1
Edit 2
. . .
Edit n

When a record is processed, the first edit is executed. The edit test is made, and if it fails, the edit action is taken. Then the second edit is executed in the same way, and so on for all the edits.

Suppose hypothetically that there were no redundancy in the data being processed (this rarely if ever happens in census practice), so that no inconsistency between different responses is possible and no consistency or anomaly edits are necessary. Then all edit tests will be for missing values or invalid codes. We may assume for the sake of illustration here that invalid codes have already been removed from the records, replacing them by an "unknown" code if it is not possible to correct them.

This leaves only missing values, and the development of the edit process involves (i) deciding which items, if any, should be imputed if values are missing, and (ii) designing appropriate procedures for the imputations decided on. However these decisions are made, the result of executing the process on a batch of records will be to replace missing values by imputed values for all items that are imputed. Re-executing the edit process on the edited records would therefore leave them unchanged.

If redundancy is present, however, there will generally be numerous consistency tests involving two or more items. Some of these tests are likely to involve common items. Suppose for example that edits i and j, $i < j$, both involve age. Then it is possible that (i) the edit i test passes, (ii) the edit j test fails and results in a new, imputed value for age such that (iii) edit i, if re-executed, would fail.

It is possible, in other words, that the action taken as a result of one edit may cause one or more subsequent edits to fail. It is therefore not necessarily the case that subjecting a batch of records to the edit process will result in an edited set of records that will pass all edits tests if the process is re-executed. This is the problem of "cycling", which may required repeated execution of the edit process to obtain a set of edited records that passes all edit tests in the process.

One solution to the problem of cycling is simply to execute the edit process repeatedly until all records pass all tests, but this is clearly unsatisfactory. A better solution is to design an edit process that assures that one execution will result in a set of edited records that will pass all tests, so that re-executing the edit process would not change the value of any item on any record.

The approach to developing such a process is evidently to consider edits involving common items jointly. We are likely to begin our thinking about edit tests in terms

of simple, *atomic* statements about impermissible combinations of values, such as

$$MS = unmarried \quad \text{and} \quad RHH = spouse,$$

where "MS" denotes "marital status" and "RHH" denotes "relation to head of household". If another such edit involves one of the same items, however, *e.g.*

$$RHH = spouse \quad \text{and} \quad AGE = 0 - 9,$$

we may form a *compound* edit test by combining these two tests with logical "or". The edit test for the compound edit will thus look either for a failure of the first edit test, an unmarried spouse, or a failure of the second edit test, a 0-9 year old spouse of head of household. We can then attempt to design an edit action that will insure that both of these tests pass.

Carrying this process to its logical extreme, we might combine *all* the individual edit tests into a single compound edit test. The achievement of Felligi and Hold (1976) was to show that having done so, it is possible to derive *from the tests themselves* a set of edit actions that will insure that the record will pass all tests if these test are re-executed.

12. Request for Comments

Involvement in the 1980 census of Indonesia instilled in me a keen interest in census editing, but as will be evident from the reference list, my knowledge is seriously dated. I would be grateful for any comments on these notes and for references to more recent literature.

The reference list includes two references that are worth citing, but which have not found their way into the text above. Banister (1980) is a useful general discussion on the risks of imputation. Granquist (1986) focuses on survey rather than census editing, but provides valuable insights and proposals.

13. Acknowledgements

Thanks to Dave Dolson for comments on an initial version of this document, dated June 1999.

14. References

Banister, Judith. 1980. Use and abuse of census editing and imputation. *Asian and Pacific Census Forum* **8**(3):1-3, 16-18 and 20.

Chapin, Ned. 1974. A new format for flow charts. *Software Practice and Experience* **4**(4):341-357.

Felligi, I.P., and D. Holt. 1976. A systematic approach to edit and imputation. *Journal of the American Statistical Association* **71**(353), Applications Section.

Granquist, Leopold. 1984. On the role of editing. *Statistisk Tidskrift* **2** 105-118.

Jones, Gavin. 1974. What do we know about the labour force in Indonesia? *Majala Demografi Indonesia* **2**(1):17-36.

Sunter, A.B., C.A. Patrick, and D.A. Binder. 1975. On the editing of survey data. Invited Paper No. 13, Warsaw Meeting, September 1-9, 1975, International Association of Survey Statisticians, International Statistical Institute.