Issues in optical character recognition for statistical data capture

Arij Dekker
Consultant in statistical computing

Email: arijdekker@gmail.com

Summary

Optical character recognition (OCR) is a popular method to capture data for censuses or large surveys. Since statistical offices often do not have enough equipment and skilled staff for an autonomous application of this method, they tend to call upon outside contractors to take responsibility for it. Such contractors, interested in doing business, will not fail to highlight the advantages of OCR in terms of cost and speed.

This paper makes an attempt to point out some of the pitfalls – mostly affecting final microdata quality – that are also present. Census and survey organizers should be aware of those, and make informed decisions, carefully weighing pros and cons. Sound questionnaire design is especially important in order to avoid or mitigate the drawbacks of OCR.

Keywords

Data capture, optical character recognition, OCR, ICR, census, survey, questionnaire

1. Introduction

Data capture for large surveys and censuses represents an important cost component and a major source of data errors. It can also take much time, thereby reducing the timeliness of statistical reporting. Over the years a range of techniques has evolved to improve the process of converting verbal responses into computer variables ready for electronic processing. The principal methods currently available are:

- keyboard data-entry from completed paper questionnaires collected centrally;
- keyboard data-entry in portable devices at the time and place of the interview, discarding the use of printed questionnaires;
- processing machine-readable questionnaire forms, usually via optical character recognition (OCR), which may include intelligent character recognition (ICR);
- on-line enumeration or self-enumeration.

Each of these methods has specific advantages and drawbacks. Centralized keyboarding, the traditional method, allows good control over the capturing process. It is, however, labor-intensive as well as slow, and increases error rates due to the dual transcription involved.

Off-line data capture by interviewers, using portable devices, is fast and can reduce error rates since the data are recorded in a single step and at the source. But in large censuses with tens of thousands of enumerators the logistics can be daunting. It involves training numerous workers who may have little earlier exposure to the kind of equipment and the type of programs being used, the need to synchronize the software in case of updates, and issues of collecting the recorded data from the numerous and physically dispersed interviewers. In some countries there may even be problems in keeping the equipment's batteries loaded. The decentralized nature of the process reduces the options for efficient management. Some experts feel that this method, known as CAPI (computer-assisted person interviewing) currently is better suited to relatively small specialized surveys than to censuses. But, as technological and software development proceed, CAPI is becoming a valid option in a wider range of environments. Brazil and other countries have demonstrated that this holds even for census [1].

Using on-line enumeration, either via an enumerator or by the respondents themselves, removes many problems of synchronization and data management. It is, however, dependent on complete, reliable, and sufficiently fast Internet. The enumerators still need to carry portable computer devices or the respondents themselves should be able to provide the necessary technical resources. Even in most developed countries this is still a far way off, even if self-enumeration may be offered as an option to respondents willing and equipped to take that road.

This leaves OCR as an approach apparently not suffering from several problems noted elsewhere. Data capture takes place at one or more central locations, allowing efficient control and management. Since the conversion process from written characters into computer code can be made extremely accurate, error rates at this stage will be low. Given sufficient scanners and software licenses, data capture can be up to ten times faster than with manual data-entry. The availability of paper questionnaires and their electronic images provides a back-up in case of data loss or doubts concerning certain responses. Now that the underlying technology has become mature, the cost of OCR is usually competitive as compared to other methods. All these positive aspects have led many organizers of censuses and large surveys to choose OCR as their data capturing method of choice.

However, OCR for statistical purposes confronts software application developers with some typical difficulties one should be aware of. There is also a tendency to outsource the data

capturing process to a vendor of OCR solutions, who is not necessarily aware of particular dangers where statistical data collection is concerned. The present article makes an attempt to clarify the issues involved and suggests approaches. The paper is based on the author's recent experiences as data processing adviser for population censuses in several developing countries.

2. Overview of OCR for censuses and surveys

The application of OCR usually involves enumerators carrying specially-designed questionnaires into respondents' homes for these documents to be completed there. It requires special care with the forms, since folding or smutching may disrupt the automatic reading process. The forms also need to be carefully printed, usually on high-quality and thus expensive paper, although technological developments have made this requirement less critical than before [2]. While self-enumeration is possible, it is usually the enumerator who completes the form according to the responses received, question after question. This is preferable since the enumerator is supposed to be fully familiar with the form(s), and to have received instructions in the proper way of writing.

Some questions can be answered by placing a mark in one of the options provided, for example "yes" or "no", making it multiple choice in a closed setting (Fig. 1). In this case OCR drops down into optical mark reading (OMR). Applying OMR removes the need for character recognition, thereby improving the error rate in the scanning process.

But for some questions the number of options is so large that it would become impractical to preprint all, for example for the variable age or for administrative area codes. In such cases the response needs to be written as a number, the handwriting later to be interpreted by the scanning process. This could be called an open question (Fig. 2). If not all codes are existing, range errors can possibly be made, such as writing a village code that does not in fact occur. Another form of open question is where responses can be answered in text, for example writing down the name of an occupation or industry. It is possible to automatically decipher such verbal responses and for the computer system to attempt allocating the pertinent code by applying algorithms supported by lists of applicable terms and their associated codes.

More common is to insert an image of the text in the scanning output. Coding can then be undertaken by a human operator, possibly with the support of a dedicated computer application. Where computers actively participate in such processes, often the acronym ICR (intelligent character recognition) is used, instead of OCR.

Most census questionnaires apply a combination of OCR, OMR and operator-assisted image processing, but short forms may be limited to OMR or OCR/OMR only.

The OCR process is commonly subdivided in four phases, as follows:

1. Scanning proper. Electronic images of the questionnaires are taken, which are stored in a data base for later reference whenever required;
2. Interpretation. The software interprets to the extent possible every response given on the (image of the) questionnaire.
3. Verification. During interpretation some answers may be found missing (while required), uninterpretable (poorly written), out of the pre-programmed range, or inconsistent with other information already captured. These cases will usually be shown in context on operator's screens and these humans will then attempt to resolve the issues. Thus, OCR cannot function without a certain amount of keyboard data-entry.

4. Transfer. The resulting data files are exported from the scanning system for further processing for editing/imputation, tabulation and so forth.

The OCR process requires prior instructions in order to function with a particular set of input forms. This includes a technical description of each individual form, specifying the locations where information is to be found. For every question the type has to be specified, as well as any logical requirements such as the allowable range of the variable concerned and relationships with other variables that need to be satisfied. Providing such instructions requires specialized skills which may include programming ability in certain computer languages such as Visual Basic. Where such skills are not available in the statistical office, one may have no choice other than to out-contract the entire operation. But this means also a loss of hands-on control.

If human coding of imaged terms is involved, this may be included as part of the verification process. But unless a large number of staff are assigned, it may hold back to a significant degree the progress of the entire data capturing process. Alternatively, one may assign this operation to a separate process whereby the special codes are inserted into the data file at a later stage. In the meantime the basic file will be available more rapidly for generating and publishing results not involving the information collected through text. A pilot census could provide an estimate of the human resources required and point the way to the preferable strategy here.

3. Issues and policy choices

3.1. Batch identification and size

The batch size throughout census processing is usually the collection of questionnaires pertaining to one enumeration area (EA). This is normally equivalent to a set of some 100 to 300 households. One household may require one or several questionnaires, but one questionnaire usually provides place for more members than are included in the average household. Thus, in most practical cases the number of questionnaires being used in a census is not very much higher than the number of households.

The scanning process preferably should use the same processing unit, with each batch being announced to the system with its identification codes, often something like Province – District – Ward – EA. This can be keyed in by the scanning operator, or recorded on a scannable batch cover form prepared by either the enumerator or the scanning operator. If keying is chosen, the scanning process needs to be halted each time for the batch information to be entered into the system. This means that batch sizes should not be too small, otherwise much time will be lost in between batches.

Questionnaires should always be identifiable based on their own characteristics, since - for whatever reason - the order of the forms may become distorted, requiring rebuilding and re-ordering the EA batches. Therefore the household questionnaires also contain Province – District – Ward – EA codes, perhaps – by mistake - different from those in the batch identification. This would mean that either a questionnaire has been misplaced, or that the enumerator made an error when writing those codes for the particular household. This mistake will not be immediately noted when the scanning process is entirely separate from the later phase of interpretation. In other words, the identification of individual households is at this stage unknown to the system. The inconsistency concerned needs to be rectified later on.

This situation is important, since it illustrates the fact that OCR allows inconsistencies in the data file that would not occur with keyboard data-entry. In the latter case processing usually is also by EA, but the data-entry application will not allow households within the same EA to carry different batch identifications.

For the OCR process inconsistencies of this type can be dealt with by operator intervention in the verification stage – requiring a programmed routine. They can also be relegated to the editing and imputation operation following data capture.

3.2 Range control

Open questions answerable by writing numerals can receive a response that is out of the permissible range of values. Even closed questions can be out-of-range if they require a response and none has been provided. Again – and as opposed to the situation for manual keying – such mistakes are not caught immediately in the OCR process, since the scanning operation is not equipped for that. And as before, the problems can be relegated to the verification stage, but the more often this happens, the larger becomes the number of staff, computers and software licenses required. One may end up in a situation where verification becomes a bottleneck and much of the potential gains in speed and reduced workforce of the OCR process are absorbed by greater resources required at the verification stage.

Range errors may also be left for the editing and imputation process, but at that stage referring back to the original questionnaire (which, by the way, will not necessarily resolve the issue) may be considered bothersome. This will lead to computer programmed guess work or replacing range errors by “not stated” codes. In both cases this will have a negative effect on data quality. Where automatic imputation is applied it’s important to pay attention to imputation frequencies, making sure that statistical distributions remain essentially unaffected.

3.3 Inconsistencies

Keyboard data-entry systems will normally not allow events of child birth to be recorded for a male person. An error message will be shown and the operator will be asked to review his work by re-inspecting the questionnaire on his desk or reconsidering his data entry if recording during his visit at the household. In difficult cases a supervisor could be consulted. By contrast, a scanner will readily accept such and other inconsistencies.

Much can be improved by thorough quality control before the scanning process. Nevertheless, as any person dealing with large enquiries knows, plenty of even obvious inconsistencies will slip through. In some cases editors may even introduce new errors.

The inconsistencies brought out by software with an editing component will need to be weeded out in the verification process and the following editing/imputation stage. This may be more time consuming and more error prone than in case manual keyboard capture would have been applied.

4. Forms design

Forms intended for OCR need to satisfy special requirements. Already mentioned are the need for sturdy paper and careful printing. The paper weight should be 90 grams per square meter or more, the printing precise and the cutting accurate as well as uniform throughout the stock of possibly many millions of questionnaires. Not all countries have printing firms that can deliver this

level of quality, so costly imports may be required. The companies that have taken on outsourced job for OCR services will readily blame the quality of the questionnaires if the scanning does not run smooth. Therefore the task of providing the questionnaires may be added into the contract, shifting responsibility to the contractor, but this obviously increases the price tag.

4.1 One or several types of form

A typical census uses separate forms for private and collective households. For the larger households dedicated extension forms may be employed. Furthermore a cover form per EA is usually present.

While OCR can easily recognize and process differently forms of multiple designs, this does complicate the process. Therefore census offices often try to work with a single form for all types of households and extensions. The cover information for an EA will always require a separate form or keying in by the scanning operator.

Not all of the questions that are asked from private households and their members do apply to collective households. Thus, with only a single form many questions will need to remain unanswered for collective households. This tends to be a problem for poorly trained enumerators, who may be inclined to collect responses for every question, no matter the household type. If this mistake happens, so in case the household type is marked as “collective” but many variables applicable only to private households have been completed, the question arises as to what is wrong. Is the household type wrong, or should the superfluous responses be struck? The opposite problem occurs too, with “private” selected as household type but the remainder of the form completed as if the household were collective. Again, the problem comes out clearly with an operator keying, but remains hidden during the first phases of the scanning process.

The persons enumerated in a household should be uniquely numbered in order to be able to individually identify and sort them. If there is a dedicated opening form per household, there the persons can be pre-numbered 1 to n, n being the number of available person lines. Only on extension forms the person numbers need to be handwritten. With a single multi-purpose form all person numbers must be provided by the enumerator.

In the case of a single multi-purpose form there may also be a risk of collecting the same information repeatedly. Most census questionnaires have a separate section for dwelling information (type of dwelling, ownership, water supply …). In a larger household this section will return on each form that is used, creating the risk that enumerators will provide the same information several times, maybe inconsistently. Of course there should be guidelines to prevent this, for example the instruction to complete housing information on the first form only. But in practice every imaginable mistake will occur in the case of large census files.

A final, but quite important, consideration is enumerator comfort. Working with a single universal questionnaire type, these questionnaires possibly bound together in a book, will be more comfortable than carrying five different types of questionnaires, each of which may run out.

Summarizing, it should be noted that the choice between one or several data collection forms has important consequences that affect the quality of response collection as well as logistics and operator comfort. In an outsourced situation the contractor will likely prefer the single form, since being more concerned with efficiency than with data quality. But the census organization should be aware of the issues and make informed choices. Alternatives can be compared in pilot exercises.

4.2 Line or column diversion and similar problems

In relatively light censuses it is common to record the attributes of a household member into a single line or column, but the face of one census form may not be large enough to hold all. Then the line or column needs to be continued, on the back side or on a second form. The important thing for enumerators is not to lose track and start recording data for one individual in an area actually pertaining to someone else. That risk is exacerbated by the fact that young children will not need to provide information about education, occupation and industry, while males obviously don't report about fertility. In some cases information for older persons or females needs to be recorded next to blank attribute entries for those not reporting. Going astray to the wrong block is a mistake easily made. The result is: males with associated fertility while adult females report "not stated", young children with academic degrees, and so on. Again, manual data capture systems can catch much of this at the source, but OCR will first accept the erroneous data and depends on editing routines at a later stage for remedial action.

Traditional questionnaires can be designed to reduce this problem. This can be done by using a small booklet per questionnaire where a narrow cover flap records the essentials for each individual such as person number, name and relationship in the household. One or more follow-up pages then provide space for the attributes, with the person's basic information remaining visible at the same line level. Clever designs, including folding open additional questionnaire spaces, provide enumerators with good support to avoid line diversion.

Sheets of different sizes and folded pages constitute a problem for scanning equipment. Applying such crafty enumerator-assisting designs is virtually impossible in an OCR environment.

Rather than using one, possibly extended, line per individual, one can also elect to print a separate section for attributes pertaining to a subset of the population. This could be fertility for women of reproductive ages. In that case the information collected must be somehow linked to the individual, probably via a repeating person identification number (Fig. 3).

This can go wrong in many ways, especially in households with large numbers of members and multiple questionnaires. The risk of diversion becomes substantial to the extent that this design option is to be considered ill-advised.

4.3 Duplicate household identifications

Duplicate household identifications are the scourge of census files. They will usually become apparent only after the files have been sorted, and then produce merged households with multiple household records, several heads of household and a large number of household members.

Keyboard data-entry programs will usually prevent the same household number occurring twice within the same EA. The problem may nevertheless surface if two EA's happen to have – by mistake – the same identification. If keyboarded data are stored in a common depository, using the same EA identification twice can also be prohibited. This largely removes the possibility of duplicate household numbers happening.

The OCR system will readily accept in scanning household identification numbers of any kind, or frequent stops may hold back the operation. Remedying duplicates at the verification stage requires specialized programming usually not within the scope of contractor services or census programmers. Thus the problem will come to the surface only at the editing and imputation phase.

Now in the situation where the census has been conducted without specially-designed questionnaire extension sheets, the occurrence of more than one household record presents a puzzle. It can have several causes:

- Duplicate household identifications because of coding mistakes;
- Multiple household records resulting from household information having been collected erroneously from more than one questionnaire in case of large households;
- Worst case scenario: a combination of both.

How to disentangle this situation, including placing members in their proper households, presents a challenge to the programmer developing the editing/imputation application. Fortunately there tend to be some clues. For example, scanning equipment usually adds individual codes to each sheet being processed, which codes would be identical for persons recorded on that sheet. But large households occupy, of course, more than one sheet. Apart from retrieving all questionnaires involved and starting the reconstruction of households from there, there usually is no fail-proof solution that solves every case.

4.4 Correcting writing mistakes

It is inevitable that enumerators sometimes must correct their earlier writings, because of misunderstandings, since respondents changed their answers, or simply because of human error in completing the questionnaire. It also occurs that supervisors, upon checking the work of their staff, find a need for rectification.

On a questionnaire intended for later keyboarding that usually is not much of a problem. Crossing out the mistake and writing the correct response nearby will usually be enough. Scanning systems cannot deal with the issue so easily, since they lack the kind of observational intelligence required.

A solution would be to have the enumerators work with lead pencils and erasers. But under field conditions writing with pencil is often less dependable than ballpoint, and erasing may turn out to be insufficiently thorough, giving rise to false marks or mistaken interpretations. Rewriting the entire questionnaire is also an option, but it is time-consuming and in the process new mistakes may be made.

Fig. 3 demonstrates an approach where a small dedicated mark indicates that an entire line has to be disregarded. This is the mark under Person ID in Q27. While the mistaken line will still be scanned, the presence of the error mark can be used to disregard the line during interpretation. If necessary, a new line containing revised information can be added to the form.

One should note that this will usually interfere with the ordering of persons on the form, which would normally list first the head of household, followed by spouse, children, and so on. Special care is required in case of repeating person ID’s, since obviously any number change for a person should be applied everywhere on the form where this number is written. The issue can become complex for large households using multiple forms that also contain repeating ID’s.

4.5 A need for thorough tests

While careful testing is always a necessity for census methods, it is of paramount importance where OCR is involved. This is since weaknesses in questionnaire design can grave consequences. Several pilot tests, including at least one covering a large sample of the population, with fully developed forms and with all phases of the scanning and interpretation process in place, are absolutely required. It is likely that the tests will turn out to have been of crucial value in eliminating design flaws and processing bottlenecks that had remained unnoticed.

5. Conclusion

Using OCR for census purposes has plenty of advantages, which over the years have been widely proclaimed. Private contractors looking for outsourced business will happily emphasize them. The balance, however, is not only positive. OCR seriously restricts the flexibility of questionnaire design, which can directly affect the quality of data collection. Furthermore, mistakes in household identification, out-of-range variables and failing internal consistency within a set of responses will usually not be noted at the initial data capturing stage, but only later in the process. Obvious mistakes that slip through field control can thus be diagnosed only later, requiring remedies that will be tentative and/or time consuming. Organizers of censuses and large surveys should make their own informed judgements on these drawbacks, rather than leaving it to commercial companies who may be less statistically-conscious than would be desirable and who also have a less immediate interest in data quality.

On-the-spot data recording by enumerators circumvents such problems and therefore can reduce the error load. Even traditional keyboarding at central locations can capture errors earlier and facilitate the correction process.

This paper is not meant to discourage data capture through OCR for censuses and large surveys. The important advantages in cost and speed may well be decisive in many environments. But census designers should be aware of the downside, including the possibility of a larger error load as compared to other data-entry methods. Thorough testing is imperative.

References

[1] A. Vicente S. de Miranda, The use of handheld computers in the 2010 Brazilian Population Census, Instituto Brasileiro de Geografia e Estadistica (2013)

[2] United Nations Statistics Division, Principles and Recommendations for Population and Housing Censuses, Revision 3 (2015). Available from http://unstats.un.org.

Figure 1 Closed questions

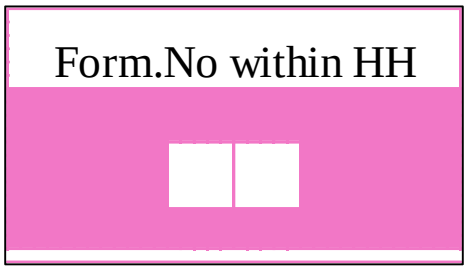

Figure 2 Open question

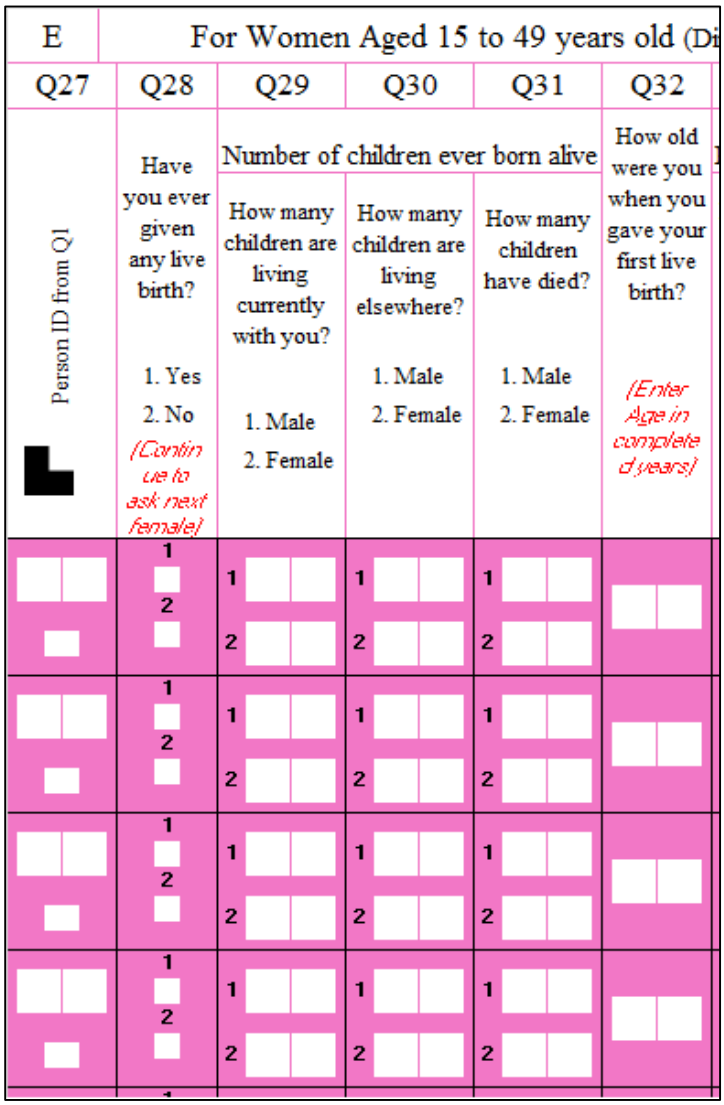

Figure 3 Repeating person ID