

Preparation for Egypt's Population and Housing Census. Data Processing Challenges

Bahy El- Din Abdel- Hamid Mortagy¹ and Nevine Naguib Hegazy²

Summary

The Central Agency for Public Mobilization and Statistics (CAPMAS) utilizes all its capabilities to provide statistical information covering all economics, social, health and cultural areas in Egypt by data gathering, processing and analysis.

The Latest Population Census was under taken in the 1996, The next Census will be held in the year 2006.

To ensure proper arrangements for conducting the Census and to deal with any concerns, Preparation should start three years before the census date.

This paper discussed that automatic data capture is available for Latin numbers, CAPMAS developed the software to capture Arabic (Hindi) numerals and separate alphabet to facilitate the process of filling out forms for the enumerators.

This paper illustrated the software and Hardware requirements that will be needed for the census which will use nearly 24 million questionnaires.

This paper presented the challenges that will face by CAPMAS during the preparation phases of the ICR / OCR Implementation and using scanning technology.

CAPMAS has adopted a progressive statistical program that is based on advanced techniques and international recommendation to facilitate and reduce data collection and processing time, The data can be captured on computers automatically by scanning and using ICR / OCR technology , This will reduce the data processing time and achieve more accurate results.

Key words

Population Census , ICR / OCR Technology, Scanning using Arabic numbers, Decentralization.

1. Introduction

The Census is one of the most important national statistical projects. Its results influence the formulation of many political, economic, and social policies, in addition to the implementation of essential programs for raising the standard of living.

The Central Agency for Public Mobilization and Statistics (CAPMAS) has adopted a progressive statistical programme that is based on advanced techniques and international recommendations. This programme should facilitate and reduce data collection and processing time which will, as a result, improve the decision making process through the provision of accurate and timely demographic data.

1: Head of IT section, CAPMAS, Salah Salema Road, Nasr City, Cairo, Egypt: Email: mortagy@capmas.gov.eg

2: Head of Data Entry Department, CAPMAS, Salah Salema Road, Nasr City, Cairo, Egypt. Email: nevinehegazy@hotmail.com

The programme is characterized by a number of improvements over the previous census carried out in 1996. Some improvements deal with the process, and others deal with techniques. The main improvements are:

1. Preparation should start early to ensure that proper arrangements are firmly in place or conducting the census. In our case we started three years before the census date.
2. Data processing time should be less than the time taken for the previous census.
3. The questionnaires should be specially designed so that the data can be captured on computers automatically by scanning and ICR/OCR technology. This will reduce the data processing time and achieve more accurate results.
4. The decentralization of data processing activities to four centers in the main regions of Egypt (Cairo- Alexandria - Ismailia and Upper Egypt). Each region is responsible for between 5 to 7 governorates, with local offices in several towns. The local offices are capable of performing limited data processing functions.
5. Census results will be presented and disseminated in a visual format to improve data user understanding and thus improve decisions by policy makers. In this regard the data from the 2006 Census will be presented spatially using a geographical information system (GIS).

In order to achieve these improvements, the 2006 population census will utilise advanced technologies for data collection and input. Intelligent Character Recognition (ICR) and Optical Mark Reader (OMR) will be used in capturing input data from the census questionnaires/survey instruments for accelerating the processing of census data. It is estimated that it will take about 9 months to process nearly 24 million (24,000,000) questionnaires with 22 characteristics (fields) using scanning and OCR/ICR technologies. This is less than half the time that would be required for manual data entry which is estimated to take more than a year and a half using single data entry. The remainder of this paper will discuss the technologies in greater detail.

The quantity of equipment that is required depends on the number of questionnaires that will be used, Estimation of household Population, Buildings and, Establishments in 2006:

No. of population	73.1 million
No. of household	15.9 million
No. of building	14 million
No. of housing unit	23.5 million
Country area	1.000.000 Km ²

Referring to our previous estimate of household population, buildings and establishments, we would need 12 servers, 1100 Pc's, for traditional data entry. It is estimated that the equipment requirement when using ICR techniques will be 16 servers, 600 Pc's, 11 scanners and the estimated cost will be reduced by 20 % compared to manual data entry.

2. Technology Overview

2.1 Intelligent Character Recognition and Optical Character Recognition

Intelligent Character Recognition (ICR) and Optical Character Recognition (OCR) recognize and capture alphanumeric characters on computer at a very high speed.

ICR/OCR, the two terms are often used interchangeably, provide complete form processing and documents capture solution. They use modular architecture that is open, scaleable and workflow controlled, and the modules include forms definition, scanning, image, pre-processing, and recognition capabilities.

ICR/OCR captures data from forms, thus reducing keystroke errors, reducing data entry time, error and cost and as a result brings data entry one step closer to complete automation, while maintaining the high level of accuracy required in form processing applications.

Forms containing character images can be scanned through scanner and then recognition the engine of the OCR system interprets the images and turns images of machine printed characters into ASCII data while ICR has the ability to turn images of hand written or printed characters into ASCII data.

2.2 Optical Mark Reader

OMR works with specialized forms. Each form contains data in pre-coded format, ID marks, which look like black boxes on the top or bottom of a form, to identify the form, and timing tracks along one edge of the form indicate to the scanner where to read the marks. OMR is a data collection technology that does not require a recognition engine since it cannot recognize hand – printed or machine printed characters. The use of Optical Mark Reader was one of the first attempts to optically process statistical information. With OMR the image of the document is not scanned or stored, hence it is considered a simpler technology than OCR / ICR and, if the forms and the system are properly designed, OMR is more accurate than either ICR or OCR. The main limitation of OMR is that it occupies a lot of space on a form and so it can only be used on short uncomplicated questionnaires.

2.3 Assessment of the Use of ICR for Census Data processing

The use of ICR has a number of advantages and it presents a number of challenges. This section includes a list of the major beneficial features and limitations. The major beneficial features are:

- It reduces the data entry time and increases its accuracy (compared to the use of manual data entry operators).
- Validation rules may be included in the system to validate and correct the data.
- Errors are identified using different colors that facilitate the review and correction process.
- Recognized data fields are updated automatically in the appropriate database form.
- Scanned forms are stored digitally thus eliminating the need for physical storage of paper forms.
- The system stores data in a database thus facilitating data analysis.
- It reduces the number of data entry persons needed.
- The system is scalable to include more clients when required.
- Forms can be designed easily.
- Drop out color improves the accuracy of recognition.

On the other hand:

- Quality of paper is an important factor. Thin or dirty forms may cause a problem. From our experience during three trial experiments, we faced a problem which led to the decrease in the recognition rate.
- Errors in filling of questionnaires decrease the rate of recognition as well as in manual data entry.
- The speed of gathering data by enumerators is less than the traditional method. An enumerator, in the same time period, may enumerate 250 households using traditional methods, and 200 households using new technology because the filling of ICR forms needs more care to write in definite positions. It needs more care than manual data entry. The enumerator must write in the specified box.
- Variation of enumerator handwriting can cause major problems in form processing and may decrease the recognition rates.
- Printing quality can cause problems if it is too dark or too light. This may reduce the rate of recognition of the forms.
- Defining a paper drop out color is an important factor since different scanners may have different appropriate colors.

Table 1:
Comparison between ICR / OCR and OMR

Points of Comparison	ICR / OCR	OMR
Hand print recognition	Y for ICR	N
Machine print recognition	Y	N
Recognition of checks an "x" s	Y	Y
Requires timing tracks / form IDs	N	Y
Require registration marks	Y	N
Electronic image storage and retrieval	Y	N

Table 2:
Comparison between ICR with Manual Data Capture

Point of Comparison	ICR	Manual Data Capture
Speed	High	low
Accuracy	High	low
No of users	Small	More than ICR
Quality of paper	Needed	Not Needed
Cost	Less than manual	More than ICR
Storage of questionnaire		
keep it	Not needed	Need large space to store it
Storage of image	Needed	Not needed
No of Pc's used	Small	More than ICR
Error during human intervention	Small	More than ICR

Trial runs are often used to determine the appropriate color(s).

- Drop out color, usually red, is the color facility in ICR system that allows the system to pick up only the meaningful information from an ICR form. The system only needs to identify the black parts, and to compare them to specifications to recognize parts that are filled or written.

2.4 Assessment of using scanned images for data entry and checking compared to a paper questionnaire

Images provide a significant step towards a paperless office; no more carrying of questionnaires to and from the work station; clear desks; faster processing; minimal physical storage requirements either near the operator or in long term storage facilities. The scanned questionnaires can be stored in various locations decreasing the chance of them being destroyed. Finally, questionnaires are available on-line and can be displayed within seconds as opposed to searching for documents on shelves which is very time consuming. Electronic filing of questionnaires is prone to less filing errors especially if a file indexing application is used.

Larger and more expensive computer screens 17" are needed for ideal display of the image. 15" screens are the standard for CAPMAS. Larger servers and additional computer memory are required to appropriately store and process images. However, this is a minor extra cost compared to the massive benefit of having the questionnaires in electronic format.

3. How to obtain good results of scanning?

Improvements in the scanning process have a positive influence on the speed and accuracy of the census. Several factors must be considered in order to ensure appropriate and correct scanning. These include quality of the form, appropriate preparation of field personnel and their supplies, and appropriate design of the quality control activities.

3.1 Quality of the Form

In order to increase the quality of the form several steps should be considered:

- A reliable printing press should be considered. Poor print quality may cause problems during the scanning and recognition phases.
- Appropriate ink and careful consideration of the dropout color for the questionnaires.
- The use of paper heavier than 80 gm per square meter may reduce paper crashes and over read the other side of a single page.

3.2 Field Personnel Preparation and Supplies

Careful handling and filing of the ICR / OCR documents are of paramount importance, therefore, survey enumerators should have appropriate supplies such as a documents bag, several black pencils, corrector, etc. Training enumerators on how to

write numeric or alphabetic characters to achieve maximum recognition cannot be understated. For example, each box should contain only one character, characters should not extend outside designated boxes, and unnecessary lines of characters such as points, decorative strokes, etc are prohibited.

3.3 Quality Control Process

To improve accuracy a number of quality control checks are required. For example procedures need to be developed to ensure that all the questionnaires are scanned completely, with no omissions or duplication, the same as manual data entry. Other procedures are needed to ensure the quality of the various processes such as the recognition process. A sample may be used to perform quality assurance checks on these processes. For example, quality assurance tests should be performed on the recognition process to ensure that acceptable recognition rates are maintained.

Reading errors are often due to:

- Poor form condition due to dirt, folds, crumples, etc...
- Forms improperly fed into scanner (at an angle).
- Forms are partially filled (incomplete forms during data collection) this is the same as manual data entry.

Reading errors may be reduced by the following:

- Preparation of coding library (dictionary).
- Learning ICR software on obscure fonts or specific country's handwriting style. Learning the software increases the recognition rates.

4. Census Forms

As previously stated, the proper design of the census forms is of paramount importance. The census uses four major forms to collect the data from the field. These forms are:

- Household and housing conditions.
- Establishments.
- Buildings.
- Residents in public houses.

Legal size (A3) regular (100 gm) paper was used. The household and housing conditions form includes sufficient space for seven household members. Households with more than seven members may use additional sheets.

The forms were designed to be processed using scanners as part of the ICR technology. It is important to follow the appropriate design process in order to develop an efficient form that reduces the processing time and maximizes the reliability of the ICR process. Some of the guidelines are:

- The instructions should be written in clear simple language.
- The data fields must be clearly defined.
- The fields should be listed by name and number of characters required for each field identified.
- Fields that require ICR should be identified.

- Size of form and weight of paper will play a role in the determination of the type of scanner.
- Registration marks should be used. These are special markings needed to aid the registration system in de-skewing the scanned image. The marks should be at least $\frac{1}{8}$ inch away from the edge of the paper.
- A margin of at least $\frac{1}{4}$ inch (6.4 mm) around the entire frame should be provided.
- A drop out color should be used to improve the recognition process.
- For best results registration marks should be placed as far apart as possible on the form.
- Size of each character box should be a minimum 5x 6 mm to be suitable for filling it with data.
- The form should include white space between each field character box and between each field to prevent the intersection of data and to get good result during recognition.
- For maximum recognition rates ensure that responses are coded with numeric characters.

5. System Requirements: Hardware and Software

5.1 Hardware

Three main hardware components must be carefully selected. These are:

- Pc's (for scanning , recognition , verification , validation, and tabulation).
- Servers (for data storage, validation and tabulation).
- Scanners. Scanner selection is critical for the success of the process.

The following factors must be considered:

- Paper size.
- Paper handling (Automatic document feeder).
- Resolution.
- Scanning speed.
- Drop out color.

We used two types of scanners Fujitsu M4099 D and Kodak 3520 which processed 90 page a minute.

5.2 Forms processing capabilities.

Forms processing features include Form ID, registration, de-skewing and form template removal. Form ID allows sorting of forms in a batch by allowing unique identification of graphical object or character strings. Recognition and de-skewing features automatically align and re- size images to their original dimension and provide more precise from template removal leading to much higher recognition accuracy.

Fujitsu M 4099 D specification

Duplex Image scanners

Features

Speed	: 90 ppm simplex scanning 180 imp duplex scanning
Document size	: from A3 to A7
Resolution	: 100 to 400
Capacity of ADF	: Maximum 1000 Sheet

5.3 Software requirements.

The software that will be used in the census consists of modules for recognition, verification, validation and tabulation. The scanning module is bundled with the scanner. The recognition module must recognize and interpret different data forms including hand written, machine – printed barcodes, check boxes, marks, numeric field, alphabetical field and mixed field. It should be able to convert the scanned image file into a text / ASCII file. In the case of CAPMAS, the software development team is responsible for the development and deployment of the remaining modules. All the modules will be integrated into one application.

6. Recognition Approach

Character recognition is an important process. Different approaches are often used to increase the recognition rate. These include the use of the same recognition engine with different settings and the parallel use of different engines with different capabilities, since some engines are better recognizing numeric while others are better recognizing alpha characters. CAPMAS used two different engines during the third pilot project (discussed later) census to determine the best approach that delivers a high degree of confidence.

There are many companies globally but only a handful of companies provide this type of product like Intelligent Data Capture Solution (Indicius) from Neura Script company, and Eyes & Hands From Read and Soft company. Also COSEKE in Tanzania is using OCR for Anydoc software.

7. Automatic Coding

CAPMAS replaced manual coding with automatic coding in order to reduce time and achieve more accurate results. Automatic coding is the process of selecting a code that matches the response given to a question. Possible answers to questions and their appropriate codes are included in a coding library, and during the data entry a list of codes appear on the operator's screen. The operator may select a code from the list that matches the answer, or enter a different one.

8. Decentralization

CAPMAS, in order to improve the collection and analysis of data, divided Egypt into four provinces. Each province covers a number of governorates and includes a province head office, a governorate office and local offices. CAPMAS established clear guidelines of the responsibility of each level, these guidelines are:

Role of local offices in governorate:

- Committees receive the documents for each governorate.
- Manual verification of the documents.

Role of provinces center:

- Committees deliver the documents to the data processing department.
- Manual verification.
- Data Entry process (Scanning – recognition – correction).
- Automatic coding.
- Load data to the head quarters system.

Role of the head quarter:

- Validation programs for each governorate.
- Edit programs.
- Elementary tabulation.
- Statistical revision for the tabulation.
- Final tabulation for each governorate.
- Final tabulation for all the country.

9. Implementation Procedures

CAPMAS has carried out three pilot projects to test the appropriateness of the use of ICR in the census data entry process. The first included 7500 households, and data entry was limited to Latin characters only. The main problems identified during this pilot project were the inappropriate design of the questionnaire that led to inaccurate filling of the form; and the second was the heavy weight of the register that included the forms and carried by the field personnel. After the redesign of the form and the register, a second pilot project was conducted using 20,000 households. During this phase, the register contained 30 forms and dropout technology was introduced. The results showed improvements over the first pilot project.

A third pilot project was conducted covering 100,000 households. The pilot project assessed two main aspects; the use of Arabic characters and the decentralization system. The recognition process used a CAPMAS-developed module to recognize Arabic numbers (Hindu numbers) and Arabic alphabetic for the first time in the Arabic countries. The pilot project was conducted in two provinces (Cairo and Ismailia). The result of the third pilot project (97% recognition rate) is extremely encouraging. Recognition rate is the proportion of data that is automatically accepted.

10. Challenges

Several problems arose during the preparation phases of the ICR/OCR implementation. The following restates the challenges.

A – Designing of documents

Some points should be taken into consideration:

The paper quality (if it is too thin or dirty).

Using trial and run method is the best way to choose the suitable dropout color.

B – Selecting the scanner

The type of scanner depends on :

Speed.

Ability to handle a large number of pages.

Duplex (read both sides in the same time).

Resolution.

Driver and interface.

C – Choosing the image software

Many types of imaging software are available in the market and each with different degrees of accuracy, some modifications may be necessary with census data processing. CAPMAS used two different engines in the third (3rd) pretest to choose the most accurate one for the 2006 population census.

D – Variation in hand writing

Due to variations in hand writing a special training program for enumerators is required to ensure they fill in the forms using ICR friendly characters. The recognition rate must be as high as possible to minimize manual data verification, and corrections.

11. Conclusion

Using ICR technology is a progressive step, which shortens the data processing time and therefore the period from field level data collection to the production of reports and other dissemination applications. The three years leading up to data collection have been extremely useful as they allowed the team to develop and comprehensively test the data processing applications before data collection started. It is expected that data entry and tabulation will be complete within 9 months of the end of data collection which will enable the team to produce the census documents and disseminate the census results in very good time.