

ORIGINAL: ENGLISH

SECRETARIAT OF THE PACIFIC COMMUNITY
STATISTICS FOR DEVELOPMENT PROGRAMME

UNFPA-SPC REGIONAL WORKSHOP REVIEWING THE 2010 ROUND OF
POPULATION AND HOUSING CENSUSES IN THE PACIFIC
(Noumea, New Caledonia, 21–25 May 2012)

Experience from the Pacific – Lessons learned for 2020

Session 8 : DATA PROCESSING IN THE 2010 CENSUS ROUND

(Document presented by Pierre Wong, Census and Survey Data Processing Specialist - Programmer (Northern Pacific), Secretariat of the Pacific Community)

1. INTRODUCTION

A population and/or housing census is the total process of collecting, compiling, evaluating, analyzing and releasing demographic and/or housing, economic and social data pertaining to all persons and their living quarters (United Nations, 2007). Normally the censuses have been done every ten or every five years depending on the size of the country. Because of the normally large scope of a census and the amount of money needed to complete such an exercise, it requires a lot of planning and dedication.

The primary purpose of the census is to provide information on the size, distribution, and characteristics of a country's population. And the data that is collected from the census is used for policymaking, planning and administration, as well as the management and evaluation of public programs such as education, health and workforce.

In order to produce quality and meaningful data the census must adhere to the following actions. 1) Preparatory work; 2) field enumeration; 3) data processing; 4) dissemination of results; 5) evaluation of results; and 6) the analysis of the results.

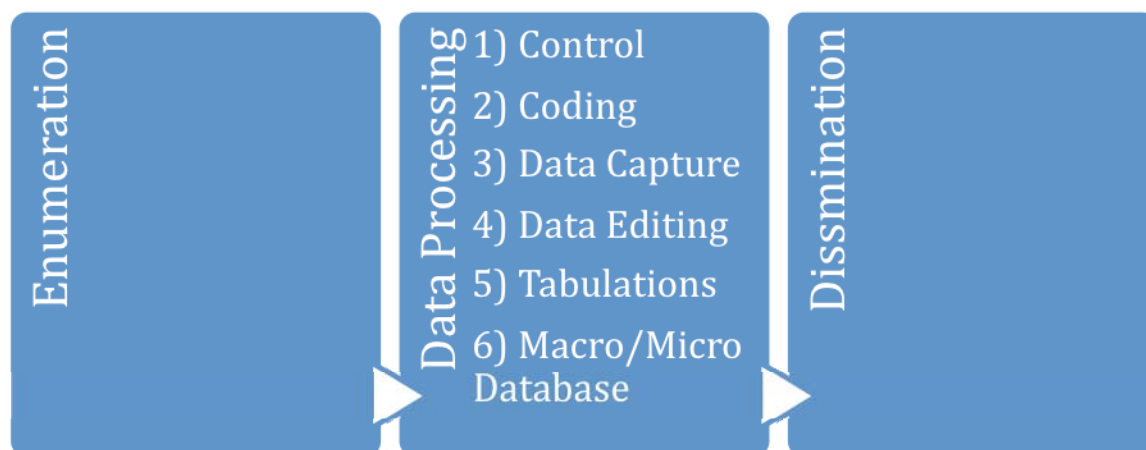
The purpose of this paper is to explore the data processing methods utilized in the Pacific island censuses.

2. PICTs CENSUS DATA PROCESSING

No matter how comprehensive and accurate the census enumeration is, the usefulness and timeliness of the dissemination of results will be affected unless the collected data is properly processed. People often mistake data processing as the activity of editing the data either manually or with the aid of a computer. The whole process is just more than editing the data and normally goes along the following manner. 1) Control; 2) coding; 3) data capture; 4) data editing; 5) tabulations; and 6) the creation of micro/macro databases.

Before we begin to explore the data processing methods utilized during the 2010 – 2011 Pacific Islands Censuses. We have to make several things clear. Data processing is not an activity that is

entirely done manually with out the aid of a computer. And although the method of computer based editing has been documented to be the approach of editing future censuses, it still requires the complex and logical thinking of a human being. So data processing centers are often left with the following question. When should the data processing team use manual editing and when should they use computer based editing? Because the census in its self is an expensive exercise the questionnaires are constantly being checked for completeness and accuracy. Every phase of the data processing cycle constantly requires human interactions.



Control

The initial phase of data processing cycle starts with control. It is imperative that control measures are taken into consideration to make sure that every household listed in the household listing is accounted for. The control process begins with control clerks sorting and recording the disposition of each questionnaire and placing those questionnaires into enumeration area (EA) batches. The control process is also where the first manual edits are made. Several technical missions were provided to the Republic of the Marshall Islands (RMI) and Kiribati census offices to instruct control clerks of the proper management of census forms. In both countries, as soon as a batch enters the processing office several steps are taken to ensure that the questionnaires have complete and valid data. The first step was to balance the number of questionnaires entering the processing office with the household listing by verifying the identification section of each questionnaire. The next step was to verify the number of people listed on the cover sheet with the amount of person records recorded in the questionnaire. The control clerk would also check for completeness of age and sex items as well as any missing information that should have been collected by the enumerator. Because this step requires constant human interaction this is normally the process that takes the most time. As in RMI, when a EA batch enters the processing office control clerks were trained to follow skip patters and identify missing information that are used in computer aided editing. When questionnaires with missing information are identified, the batch goes into a holding section until questionnaires with missing information have had several attempts to recover the missing information. After sufficient attempts have been made to complete questionnaires with missing information the questionnaire's cover balance sheet is entered into a database or spreadsheet. In Kiribati, the balance sheet contained several items that were used for control and preliminary counts. These items included Island, EA, Village, HH number, Sex, and Age. As soon as the EA batch has been entered into the preliminary database the batch is placed into the coding section.

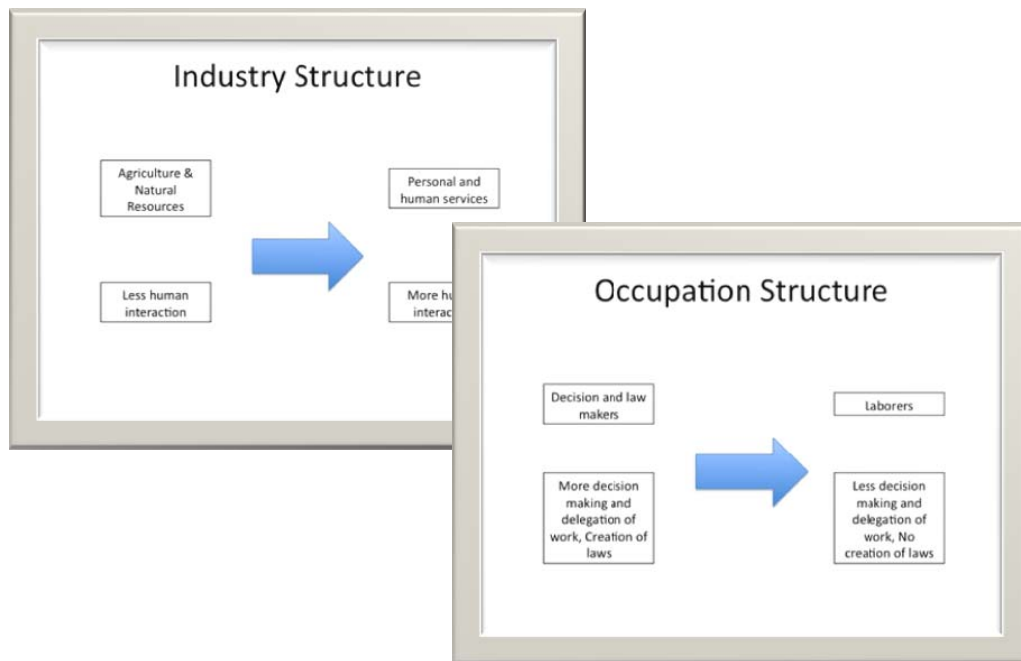


RMI – Control and Quality Assurance

Coding

Coding is the process in which responses in the questionnaire are assigned a numeric representation. This process allows questionnaires to be keyed into the computer with greater speed and accuracy. Because data processing computers are not designed to logically interpret words and sentence structures, human interaction is used once again to logically decide which numeric codes represent responses in the questionnaire. For example males are numerically coded a value of 1 and females are numerically coded a value of 2. There are several benefits of coding. One, coding makes the process of analyzing the data easier because responses are numerically quantified. Two, large amounts of information can be saved in the computer utilizing only a few bytes of hard drive space. And three, provides a minimum level of data security.

There are normally two types of coding activities utilized during a census. The first being the computer assisted coding process and the other relied on personnel to manually code items in the questionnaire. During the 2010 rounds of PITCs census, we decided that manually coding questionnaires was sufficient because the number of items in the questionnaire that need a numeric code was small. Code books were developed to properly assign codes in the census questionnaires. Some codes assigned to items in the questionnaire were country specific and other codes use international standards such as International Standard Industry Classification (ISIC) and International Standard Classification of Occupation (ISCO). To save time during the coding process most of the common responses are printed onto the questionnaire. This allows the enumerator to write the associated response code into the response sections of the questionnaire. Because the time between censuses are can span five or ten years there is a chance that the ranking of common response will change. And in the case with Kiribati we took advantage of the pilot census to develop new codes for country specific items that would have otherwise been grouped into a “Other” category. Because proper coding is important to any census one whole day was dedicated to training coders in both RMI and in Kiribati to properly code ISIC and ISCO related items.



RMI Sample ISIC and ISCO training slides

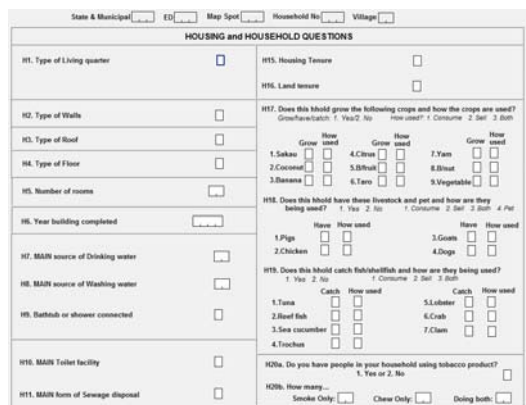
Often times there are responses that have not been coded in the questionnaire or the code books and new codes have to be created. To make sure that every coder was using the most updated codes a whiteboard was placed at coding sections of the data processing centers so that every time an item needed a new code it was placed on the whiteboard for other coders to reference.

	Occ.	Indus.	Indus. for Sale	Indus. for only consumption	
Fishing	61521	0311	9810		(11) Diploma
Toddy Cutter	61122	0161	9810		Pb. Religion (10) Bahai
Baker	74120	1071	9820		(11) Te Rauri-kamau (A) Te koana
Market donor food	91110	4781			(13) Muslim
" Non-food	91120	4782			PB. Education
Nimoko (tobacco)	74160	1200			(11) Diploma (12) Certificate
Gardener	61130	0161	9810		P10. area/field/subject
House work	99990		9820		(37) Theology (Religion) (38) Construction
Boat builder	71210	3012			
Mat weaving	34711	1312			

Kiribati – New Code Reference Board

Data Capture

There were two forms of data capturing employed during the 2010 rounds of PICTs censuses. The first was scanning which used high output document scanners to convert census questionnaires into electronic data. The other was manual keying. The entry screens developed in the Federated States of Micronesia (FSM), RMI and Kiribati were developed using CSpro and incorporated several consistency checks. Because data entry operators are not subject matter specialist few edit correction activities were implemented during the data entry phase. These checks included completeness of the questionnaire's identification; flow of skip patterns; missing sex and age items and questionnaire disposition as well as structural information that are restricted by its geography. For example a household in an atoll that states that its source of drinking water is from a river. In all three PICTs data processing centers, extra effort was made not to have data entry operators correct missing information so that there was little bias made to the dataset. Every datafile was processed by batches and backed up though a single server. For RMI and Kiribati, updates and backups to the system were made utilizing a script that was only executed by the data entry supervisor. The script was designed to run on the server which updated data dictionaries in each data entry terminal as well as retrieve data stored locally and placed into the file server.



The screenshot shows a data entry application window titled "HOUSING and HOUSEHOLD QUESTIONS". It contains various questions with checkboxes and dropdown menus. For example, H1. Type of Living quarter, H2. Type of Walls, H3. Type of roof, H4. Type of Floor, H5. Number of rooms, H6. Year building completed, H7. MAIN source of Drinking water, H8. MAIN source of Washing water, H9. Bath tub or shower connected, H10. MAIN Toilet facility, H11. MAIN form of Sewage disposal, H15. Housing Tenure, H16. Land tenure, H17. Does this hold grow the following crops and how the crops are used?, H18. Does this hold have these livestock and pet and how are they being used?, H19. Does this hold catch fish/shellfish and how are they being used?, H20a. Do you have people in your household using tobacco product?, H20b. How many... Smoke Only, Chew Only, Doing both.

At different parts of the data entry application, the computer will display the following window. This option window will allow the user to add more people if they accidentally enter less people than actually reported in the main totals. This screen will only show at the last person in the roster and will show up on different characteristics depending on the Age, gender and responses to items in the questionnaire. The default response in this Option Window is no. Meaning that no other people will be added to the roster.

Would you like to enter more people?
NO
YES


Housing Form (Single Occurrence)
H1 thru H3 are self explanatory.
H4. Does this household grow? – This item will be coded as [1] if checked and [2] if blank

Questionnaire -

H4. Does this household grow? (Check box)	near by	else-where	near by	else-where
a. Breadfruit	<input checked="" type="checkbox"/>	<input type="checkbox"/>	b. Te bero	<input type="checkbox"/>
c. Babai	<input type="checkbox"/>	<input type="checkbox"/>	d. Te kaina	<input type="checkbox"/>
e. Banana	<input type="checkbox"/>	<input type="checkbox"/>	f. Coconut dwarf	<input checked="" type="checkbox"/>
g. Pawpaw	<input checked="" type="checkbox"/>	<input type="checkbox"/>	h. Coconut local	<input checked="" type="checkbox"/>
i. Sweet potatoes	<input type="checkbox"/>	<input type="checkbox"/>	j. Cabbage	<input type="checkbox"/>
k. Does this household cut toddy?	<input checked="" type="checkbox"/>	<input type="checkbox"/>	l. Other	<input type="checkbox"/>

Data Entry Application -

H4. Does this household grow?	near by	else-where	near by	else-where
Breadfruit	1	2	Te bero	2
Babai	2	2	Te kaina	2
Banana	2	2	Coconut Dwarf	1
Pawpaw	1	2	Coconut local	1
Sweet potatoes	2	2	Cabbage	2
Does this household cut toddy?	1	2	Other	2



Top left: FSM Data entry screen, Bottom left: RMI Data entry, Above: Kiribati Data Entry Training Manual.

Data Editing

Data editing for the 2010 round of PICTs census used four phases of editing. The first phase of the data editing was the control phase which control clerks checked for completeness of the questionnaire. During this phase, items were verified by contacting the respondents either by phone or by home visit. The countries took advantage of enumerators still on the field to complete any missing information especially those pertaining to the head of the household, education and fertility questions. The second phase of data editing was completed during data entry on items that had responses in places where no responses was expected and vice versa. Any information that was missing or incomplete in the questionnaire was substituted with a special code and keyed into the computer. Other than corrections

to age, sex to name association and skip patterns no other information was edited during this phase. The third phase utilized a standardized editing method called dynamic imputation. The method imputes missing or invalid items in the questionnaire with a person in the same geographical region that displays similar characteristics. The method used an approach called top-down to prevent circular and over editing of data. This method was used in FSM, Kiribati and RMI. For the Vanuatu and Solomon Islands where scanning was employed editing was done using interactive systems. But in the cases where complete EAs were missing, an imputation method was devised to link the household roster with EAs in the general vicinity. The fourth phase was more of a quality control issue and refinements to the data edits. This was normally done with the production of tables and the interaction of subject matter specialist.

Tabulation

As soon as the edit specification was completed we then tested the data with a series of quality control tabulations. The quality control tabulations showed outliers that normally would go unnoticed until the data had been broken into the EA or village level. Some of the quality control tabulations used in the PICTs 2010 census rounds checked single year age and education level, labor force participation, literacy, fertility and items that had certain skip patterns. For FSM, we used automated tabulation to produce custom tables that utilized CSPro table logic features. By using this method the table logic can be easily adapted to produce XML code that formatted the tables so that it can be imported in PopGIS. This was successfully implemented in FSM to update the FSM PopGIS. For RMI and Kiribati we used the traditional drag and drop method of tabulations. This proved to be a better alternative because it did not use table logic and the PICTs NSO's were able to produce tables on demand without the need to learn table programming.

Micro/Macro Databases

After the data has been edited and table produced the next phase is the creation of micro/macro database. In FSM, Kiribati and RMI we went a step further and started to incorporate Nation Minimum Development Indicators (NMDI) into the data file. By doing this we made sure that those referencing the National dataset followed the same definitions defined by the NMDIs and MDG.

3. WHAT WORKED WELL?

The data processing activity has many steps to ensure that the data collected are accurate and presented on time. We found out that training, testing and applying every phase of data processing produced good results. But this was not enough. We also had an obligation to monitor that individual processes were done correctly and that problems were solved immediately. During data editing of the RMI census we found out that removing an NSO staff away from the distractions of the office and having him work as an attachment with us to produce their edit specification improved the quality and timeliness of the data.

4. WHAT DID NOT WORK SO WELL?

1) Unrealistic timelines –

It is often hard to estimate how long a data processing activity will take and it is dependent on several factors such as the complexity of the questionnaire, the number of data entry operators, the number data entry terminals and the number of questionnaires. In addition, the lack subject matter experts to test edited data, hampered the quality and timeliness of the data.

2) Little training in quality control clerks – Not enough time was dedicated to the training of data control clerks.

- 3) Lack of documentation –
Most of the times we are caught up in producing outputs that we forget to document what has been done. Because of timelines and stretched human resources, we don't have the time to document data processing activities properly. The documentation process is vital in improving future data processing activities.
- 4) Level of NSO involvement and shared responsibilities –
This did not work so well in that there was no clearly defined line of responsibilities. For example, when data processing activities were undertaken in FSM there was no clear role on who will be producing data outputs.
- 5) Some census projects did not have an agreed editing plan –
No matter which mode of data capture is used in the census an editing plan must be incorporated in to the data processing plan. For example in the Solomon Islands it was agreed to utilize scanning in the capture of census questionnaires. But there was no editing plan in place to correct scanned data.

5. RECOMMENDATIONS ON WHAT CAN BE IMPROVED AND HOW?

- 1) SPC attachments –
We have seen that when staff from a NSO is dedicated to the completion of a project it give the NSO a sense of ownership.
- 2) Clearly define roles in term of references –
Define activities and roles of every person involved with data processing activities. Who will be responsible for the completion editing specifications? Who will be responsible for the completion of data tables?
- 3) International classification codes –
There is a need to utilize standardized classification codes and proper training on the use and rationale of standardized codes.
- 4) Single data dictionary and data file
In future PITSc census we will be utilizing a single dictionary and data file that will contain three sections. The first section will be the raw unedited data items. The second section will be the edited data items and the third section will be calculated variables that will use established international definitions.
 - a. Raw data
 - i. Household level – appended raw household level calculated variables
 - ii. Person level – appended raw person level calculated variables
 - b. Edited data
 - i. Household level – appended edited household level variables
 - ii. Person level – appended person level calculated variables
- 5) Documentation –
Revisit our commitment to utilize and train NSS to document all census and survey activities using metadata documentation standards. As well as develop a revision numbering systems for edits specifications and datasets.

6. CONCLUSION

There is a need to standardize data editing methodologies. We have seen that dynamic imputation can work with both manual and scanned data capturing systems. Dynamically imputed data helps demographers and subject matter specialist interpret census data. But there is also a need to still employ manual editing practices especially when the field enumeration window is still open. There is also a need properly archive final dataset with their associated data dictionaries and documentation.
