Checking for duplicates and inconsistent entries

I. Introduction

One of the goals of CBMS is to provide accurate and reliable data to be used by local planners and policy makers. Inaccurate and unreliable information can give misleading results and false conclusion. Thus, to ensure accuracy of CBMS data, duplicate and inconsistent entries should be eliminated. In CBMS process, data duplicates and inconsistencies may occur in encoding, processing and digitizing stage. This session will provide procedures on how to address these entries in the existing database.

II. Checking for duplicates and data inconsistencies

Duplicates occur when there are two or more same cases or observations in the database that result to data redundancy. In some cases, there are duplicate household IDs but the information is different. Cases like these make the data inconsistent since it creates conflicting versions of the data. Redundancy and inconsistency of data may occur in the following:

A. Data Encoding

Common error in encoding is duplication of household IDs in a barangay. This could happen due to the following:

- Same questionnaire was encoded in different computers.
- Same household ID was assigned to different households.

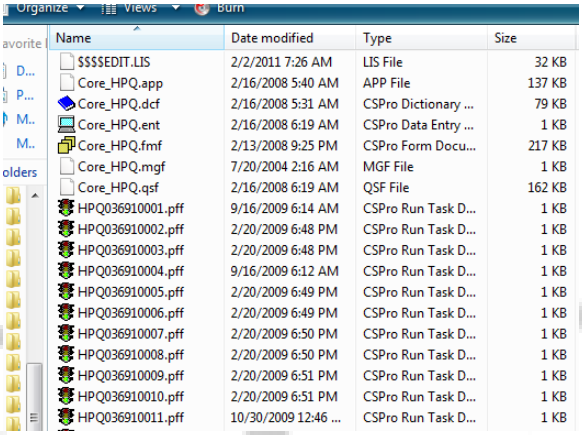
In addition, data inconsistencies can be one of the following:

- In one barangay text file, the encoded questionnaires have different barangay codes.
- No specified code for barangay, purok or household.

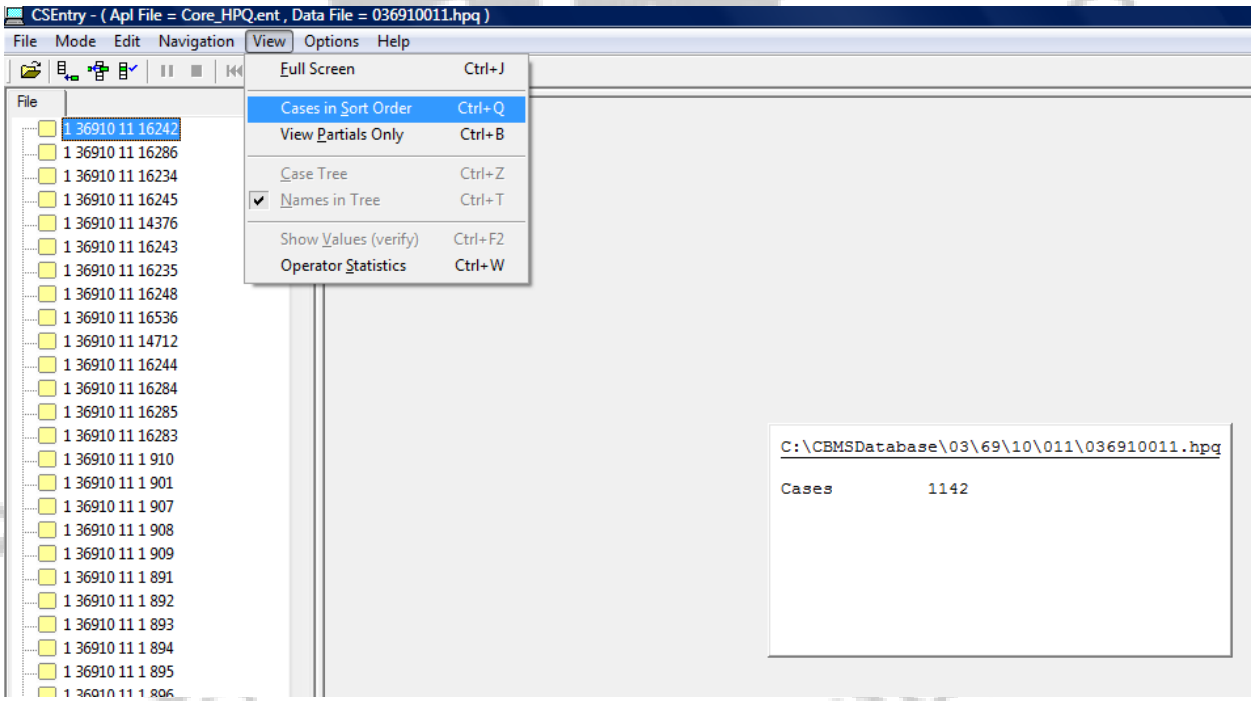
The following examples presented will address the aforementioned problems.

Example 1:

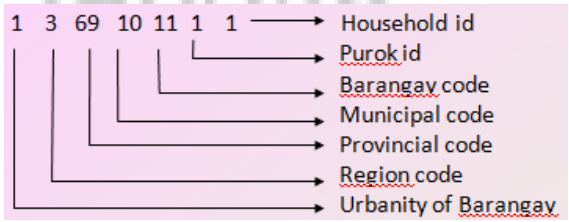
1. For this example we will check the encoded data of Brgy Carino, Paniqui, Tarlac. Open the *Encode* folder of Paniqui, Tarlac in C:\CBMSDatabase\03\69\10. Double click *HPQ36910011* to view the encoded file.



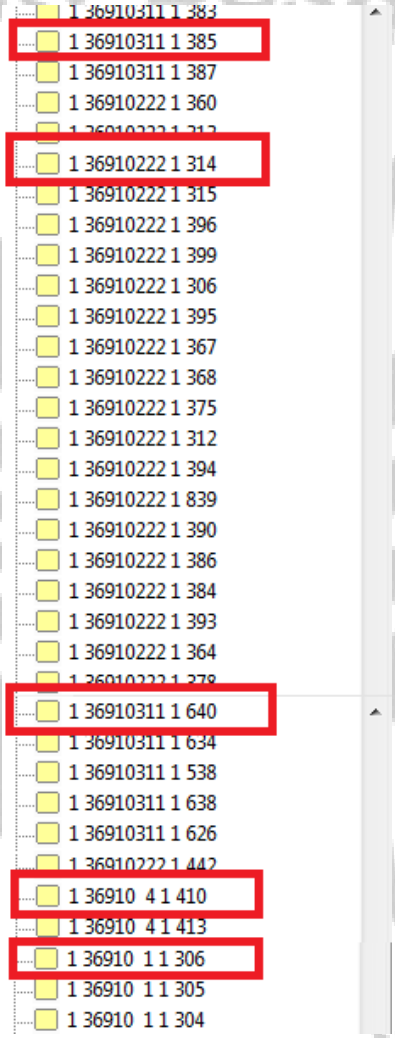
2. Brgy. Carino has 1,142 encoded households. For easy viewing, sort the file in ascending order by barangay, purok and household ID. Go to View and click *Cases in Sort Order.*



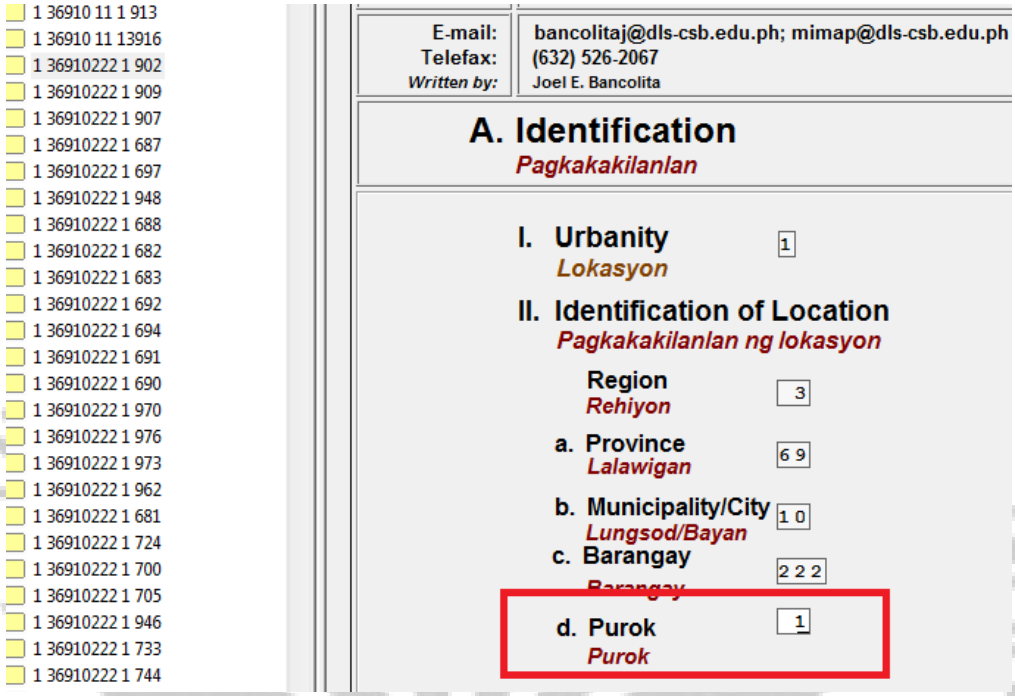
Recall that codes in the right hand side correspond to the following:



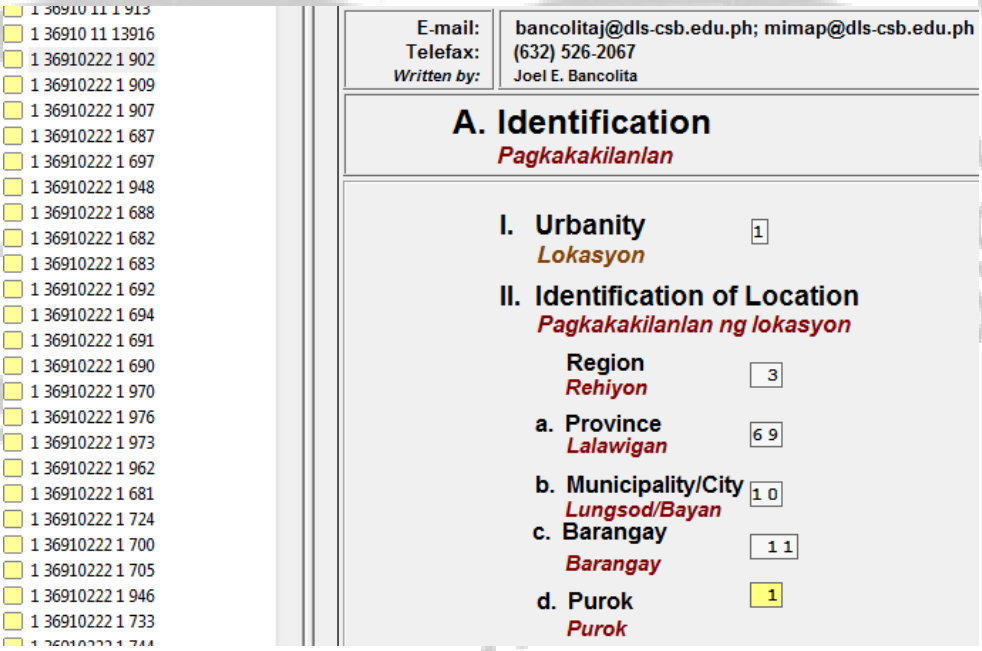
3. Browse the encoded households; notice that the barangay text file for ____, Paniqui, Tarlac has 5 barangay codes-222, 311, 4, 1, and 11(which should be the correct code). Recall that ID items such as urbanity, region code, province code, municipality code and barangay code are called *Persistent* items, where the codes are automatically assigned by the system. However, some encoders may have mistakenly modified the barangay code while encoding.



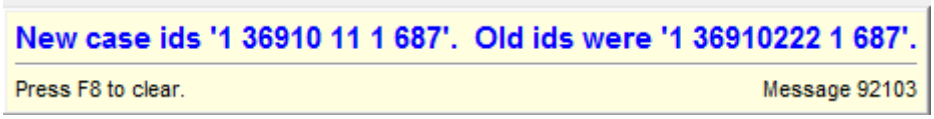
4. Before correcting all these, it should be checked first if indeed all the households belong to barangay 11 or if these households were incorrectly encoded as barangay 11. Check with the master list of the barangay to ensure correctness of the assigned household IDs. Suppose all these households belong to barangay 11, to correct the barangay code, open the first household questionnaire with incorrect barangay code. Click on the purok field.



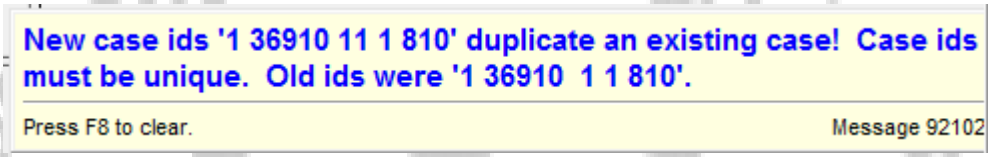
5. Press F7 so that the cursor will go to barangay field. Then, change the barangay code.



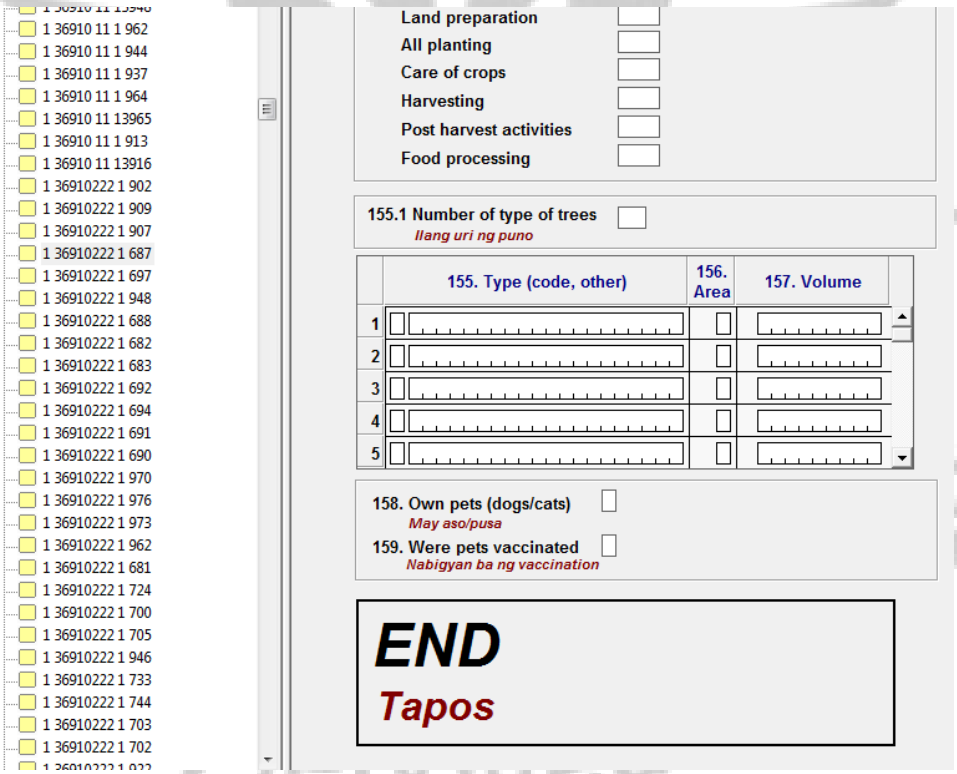
6. To save the changes, go to the end of the questionnaire. Use arrow keys to go to the next page or Page Down key. The system will then verify the change in the case ID. Press F8 to accept the new case ID.



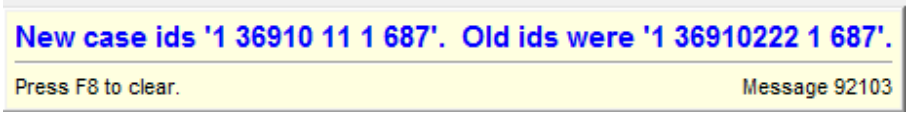
7. For duplicate IDs, the system will automatically display a message notifying the user. Modification of the barangay code will not be accepted hence, it is suggested to check the master list and the hard copy of the questionnaire. If the duplicates are two different household then assign a new ID to one of the households.



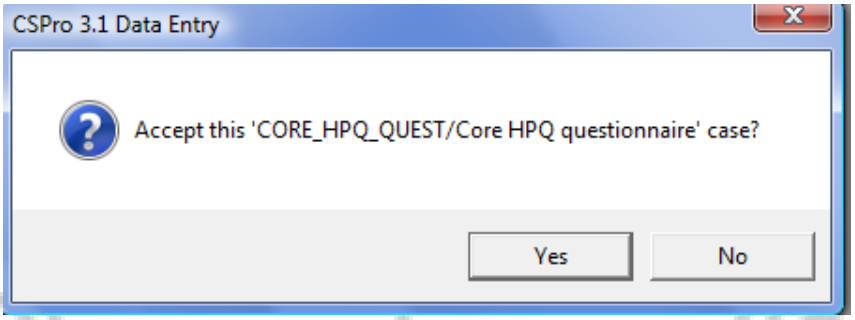
8. Suppose there are no duplicate IDs, the cursor will automatically go to the next page. If no other changes to be made to the encoded data, then go to the end of the questionnaire by pressing page down.



9. Press F12 to save the encoded questionnaire. The encoding system will again verify the case ID. Press F8 to clear.



10. A message will prompt the user to accept the questionnaire. Click Yes.

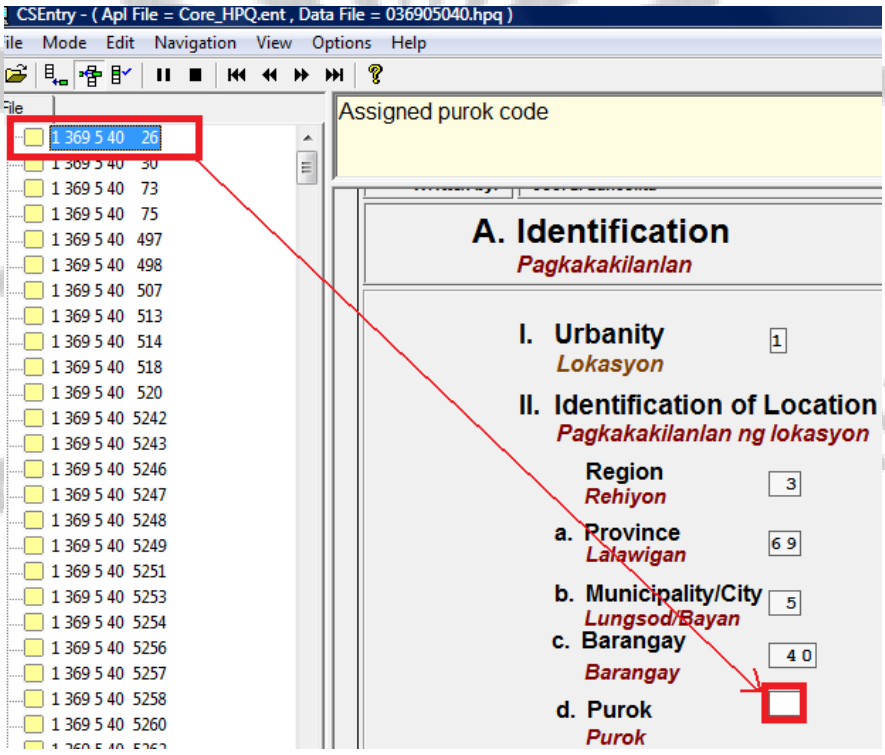


If no duplicate household IDs, all subsequent questionnaires with incorrect barangay code will be modified.

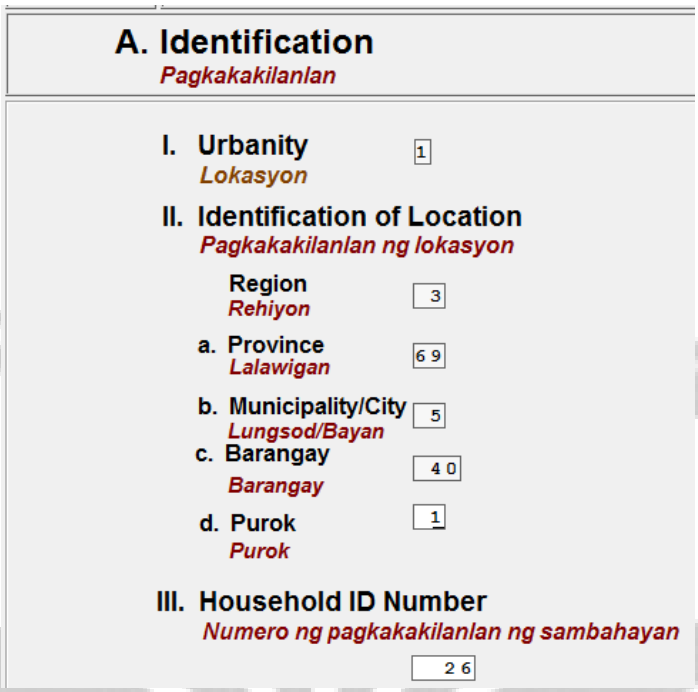
Example 2:

1. Open the folder Encode of Concepcion, Tarlac in C:\CBMSDatabase\03\69\05. Double click *HPQ36905040* to view the encoded file.

2. To sort the encoded questionnaires, go to *View* and click *Cases in Sort Order.* Notice that the PSGC code of the first questionnaire lacks purok code. Hence, the box for purok code is blank.



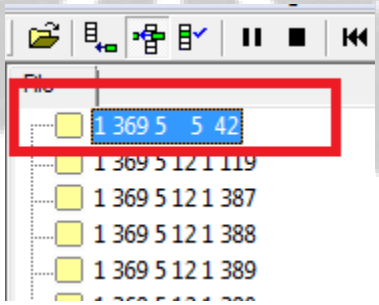
3. Look for the household number 26 in the master list and obtain the assigned purok code for the said household. For example, the assigned purok code is 1, then click on the purok field and type 1.



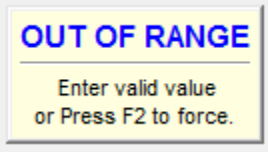
4. To save the changes made to the questionnaire, refer to Example 1 and do steps C to G.

Example 3:

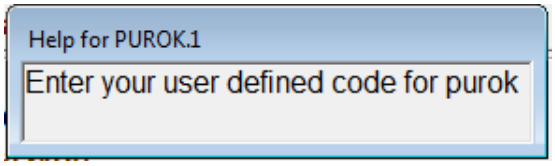
1. For this example, we will use the barangay data of Datung-a-matas. Open the *Encode* of Concepcion, Tarlac in C:\CBMSDatabase\03\69\05. Double click *HPQ36905012* to view the encoded file.
2. Sort the encoded questionnaires. Notice that the PSGC code of the first questionnaire lacks barangay code.



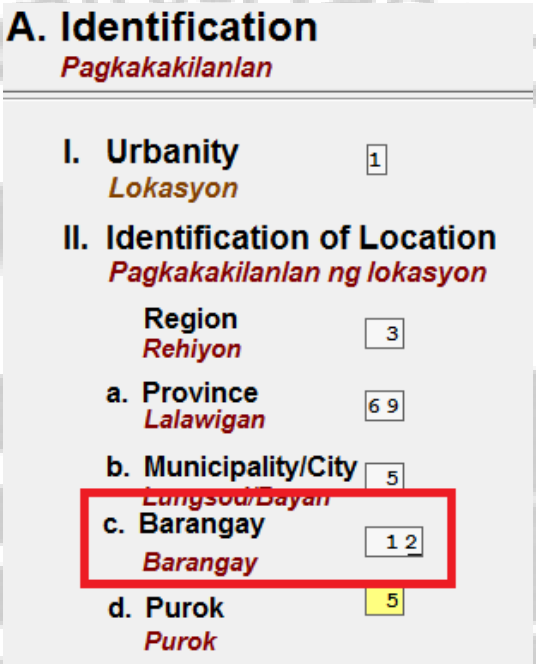
3. Double click the encoded questionnaire. Press F2 to force open the encoded questionnaire.



4. A window will appear requesting for a purok code. Press escape key.



5. Next, press F7 and edit the barangay code.



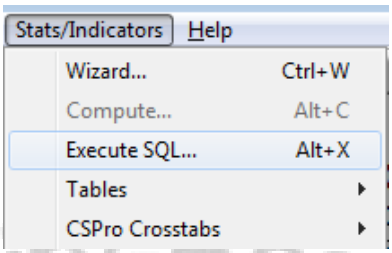
6. Again, refer to example 1 and do steps C to G to save the modified questionnaire.

B. Data Processing

There are cases that some duplicate household IDs are not detected in data encoding, and thus affects the processed data. Moreover, in processing the municipal data using StatSim, there are instances that a barangay text file is imported twice. This causes duplication of data and doubles the result. Checking for duplicates in the processed data will be presented using the examples below.

Example 1

1. To check the duplicates, first open the CBMS StatSim database (For example, *tarlac_concepcion_40)* Go to *Stats/Indicators* in the menu bar. Select *Execute SQL* or Alt+X.



3. Copy the SQL statement provided in the box below, and paste it in *Execute SQL window.*

```
SELECT mun, brgy, purok,_freq, count(hcn) as count
FROM (SELECT prov, mun, brgy, purok, hcn, count(hcn) as _freq
FROM hpq_hh
GROUP BY prov, mun, brgy, purok, hcn) as hpq_hh_count
GROUP BY mun,brgy,purok,_freq;
```

4. A table will be displayed showing the number of households (*Count* column) by purok, barangay, and municipality. The *_freq* column shows the duplicate household IDs. The figure under the column *_freq* should be "1" to indicate that the household IDs are unique.

mun	brgy	purok	_freq	count
05	040	NULL	1	26
05	040	01	1	154
05	040	02	1	84
05	040	03	1	47
05	040	04	1	150
05	040	05	1	139
05	040	05	2	1
05	040	06	1	38
05	040	07	1	55
05	040	08	1	153
05	040	52	1	1
05	040	65	1	1

In the table above, the *_freq* column indicates that in the imported data, one household ID in purok 5 was used twice. But the rest of the household IDs in barangay Santa Cruz are unique. Other inconsistencies are also shown such as Purok code "null", 52 and 65. To address the missing and incorrect purok codes, refer to examples 1 and 2 in the data encoding. On the other hand, to identify the duplicate household ID, copy the syntax provided below, and paste it in *Execute SQL window.*

```
SELECT mun, brgy, purok,_freq, hcn
FROM (SELECT prov, mun, brgy, purok, hcn, count(hcn) as _freq FROM hpq_hh GROUP BY prov,
mun, brgy, purok, hcn) as hpq_hh_count WHERE brgy=40 AND purok=5 AND _freq=2
GROUP BY mun,brgy,purok,hcn;
```

mun	brgy	purok	_freq	hcn
05	040	05	2	8745

5. The result table shows that household ID 8745 is not unique. To correct this, check first if household information for both cases is the same. Also, verify the household ID using the master list. If indeed this is a case of duplication then delete one of the encoded data, and re-process the text file using the corrected text file.

Example 2:

1. Open the database *tarlac_concepcion_8* and go to *Stats/Indicators* in the menu bar. Select *Execute SQL* or Alt+X.

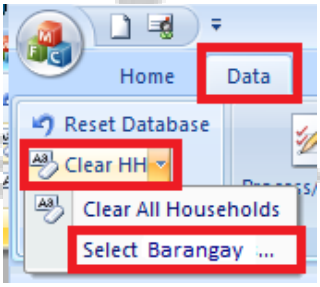
3. Copy the SQL statement provided in the box below, and paste it in the *Execute SQL window.*

```
SELECT mun, brgy, purok,_freq, count(hcn) as count
FROM (SELECT prov, mun, brgy, purok, hcn, count(hcn) as _freq
FROM hpq_hh
GROUP BY prov, mun, brgy, purok, hcn) as hpq_hh_count
GROUP BY mun,brgy,purok,_freq;
```

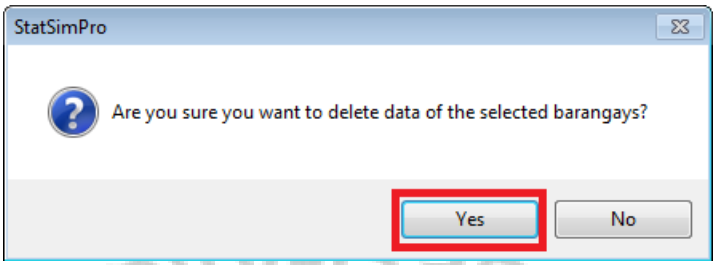
4. For this example, encoded questionnaires for Brgy. Castillo were imported twice as indicated in the frequency column of the table.

mun	brgy	purok	_freq	count
05	008	01	2	123
05	008	02	2	93
05	008	03	2	221
05	008	04	2	137

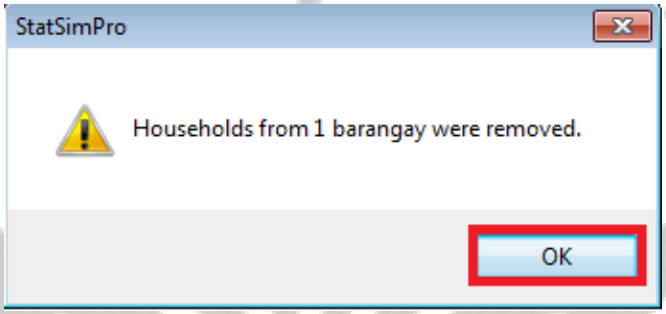
To delete the imported text file, go to the Data tab then select Clear HH and click Select barangay.



A new window with a list of imported barangays will be displayed. Select Brgy. Castillo. The system will ask if the user is sure to delete the selected barangay. Click OK.



The processed data of Brgy. Castillo has been deleted from the database. A confirmation message such as below will appear. Just click OK.



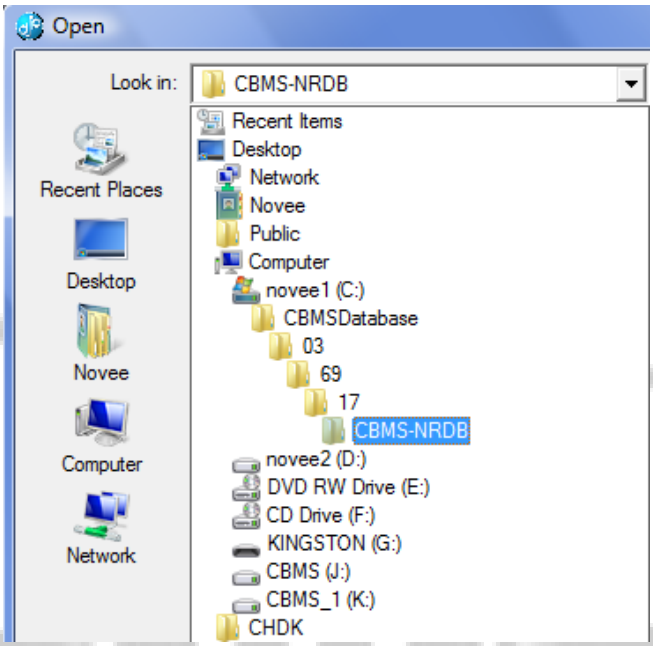
5. Go to Home tab and import again the barangay text file of Castillo, and reprocess the data.

C. Digitizing

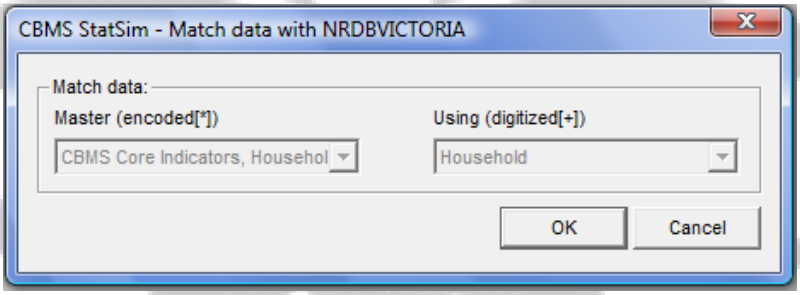
In digitizing, duplicates occur when one household has 2 or more digitized dots in the NRDB file. To detect the duplicates, matching using StatSim is performed. Recall that matching process ensures that each encoded household has a corresponding digitized location in the NRDB file. The following example will explain the matching process.

Example 1

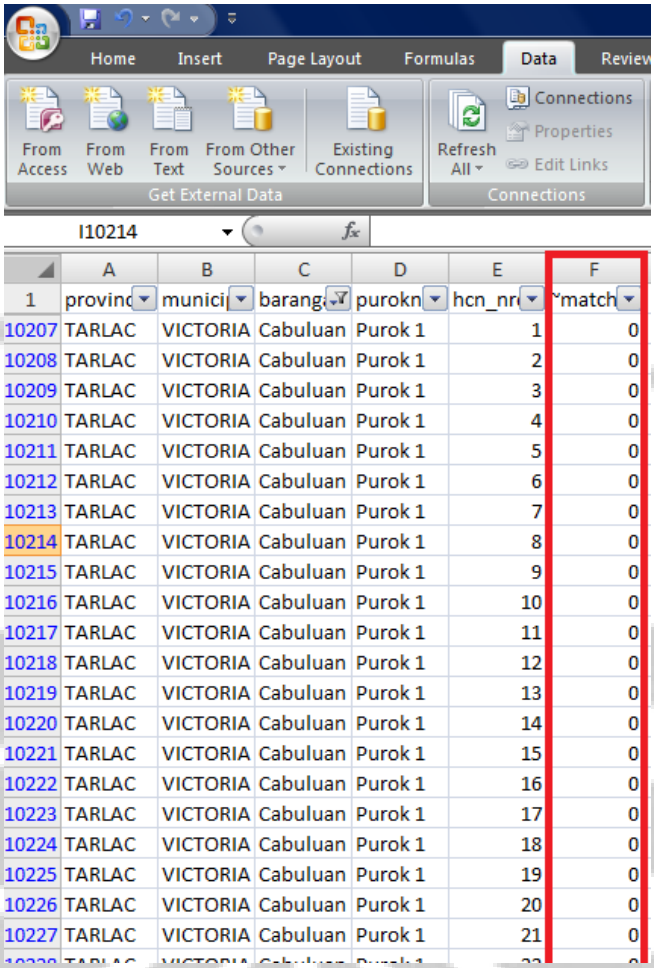
1. Open the CBMS StatSim database- *tarlac_victoria.* This database contains processed data for 2 barangays-Cabuluan and San Francisco in Victoria, Tarlac. Go to *Data* and select *Matching.* To open the NRDB, in the *Look in* field, go to C:\CBMSDatabase\03\69\17\CBMS-NRDB and select *NRDBVICTORIA.mdb.*



Another window will appear verifying that the processed data will be matched with digitized households in NRDB. Click OK.



3. A report in excel format showing the matched and unmatched households is generated. In doing the matching, our goal is to match all the encoded data with the digitized household. In the report, this is indicated by zero (0) under the column *~match*. Appearance of 1 and -1 denotes mismatches. To check if households in the Barangay Cabuluan and San Francisco are matched, filter only the data for said barangay in the excel file. Report shows that all households in the *~match* column are zero. However, it is still necessary to check if all encoded and digitized households have no duplicates.



	A	B	C	D	E	F
1	provinc	municip	barang	purokn	hcn_nr	match
10207	TARLAC	VICTORIA	Cabuluan	Purok 1	1	0
10208	TARLAC	VICTORIA	Cabuluan	Purok 1	2	0
10209	TARLAC	VICTORIA	Cabuluan	Purok 1	3	0
10210	TARLAC	VICTORIA	Cabuluan	Purok 1	4	0
10211	TARLAC	VICTORIA	Cabuluan	Purok 1	5	0
10212	TARLAC	VICTORIA	Cabuluan	Purok 1	6	0
10213	TARLAC	VICTORIA	Cabuluan	Purok 1	7	0
10214	TARLAC	VICTORIA	Cabuluan	Purok 1	8	0
10215	TARLAC	VICTORIA	Cabuluan	Purok 1	9	0
10216	TARLAC	VICTORIA	Cabuluan	Purok 1	10	0
10217	TARLAC	VICTORIA	Cabuluan	Purok 1	11	0
10218	TARLAC	VICTORIA	Cabuluan	Purok 1	12	0
10219	TARLAC	VICTORIA	Cabuluan	Purok 1	13	0
10220	TARLAC	VICTORIA	Cabuluan	Purok 1	14	0
10221	TARLAC	VICTORIA	Cabuluan	Purok 1	15	0
10222	TARLAC	VICTORIA	Cabuluan	Purok 1	16	0
10223	TARLAC	VICTORIA	Cabuluan	Purok 1	17	0
10224	TARLAC	VICTORIA	Cabuluan	Purok 1	18	0
10225	TARLAC	VICTORIA	Cabuluan	Purok 1	19	0
10226	TARLAC	VICTORIA	Cabuluan	Purok 1	20	0
10227	TARLAC	VICTORIA	Cabuluan	Purok 1	21	0

4. A detailed report on the matching process is also generated in StatSim. The StatSim report shows 4 tables. The first table is for checking the duplicates in encoded households. This is the same report generated by Example 1 in Data Processing. The table shows the number of households and the household instance per purok, barangay, and municipality.

CBMS Statistics Simulator

CBMS Statistics Simulator Merge/Match Report

Checking Duplicates in Encoded Households

prov	mun	brgy	purok	household instance	frequency
69	17	008	01	1	88
69	17	008	02	1	102
69	17	008	03	1	96
69	17	020	01	1	43
69	17	020	02	1	38
69	17	020	03	1	127
69	17	020	04	1	36
69	17	020	05	1	41

In our example, the since all household instance is one then no duplicates in the encoded questionnaires.

5. The next table in the report checks the duplicate households in CBMS-NRDB. Since we are only checking for 2 barangays, copy the table and paste it in excel. Again, filter the data for barangay Cabuluan and San Francisco only. Notice that in Purok 1 and Purok 2 of Brgy. Cabuluan the household instance is 2. This indicates that in Purok 1 there is one household which was digitized twice in NRDB. In addition, in Purok 2 of the same barangay, there are 2 households which were digitized twice.

CBMS Statistics Simulator

CBMS Statistics Simulator Merge/Match Report

Checking Duplicates in CBMS-NRDB Households

	purok	barangay	municipality	province	household instance	frequency
31	Purok 1	Cabuluan	VICTORIA	TARLAC	1	87
32	Purok 1	Cabuluan	VICTORIA	TARLAC	2	1
50	Purok 1	San Franci	VICTORIA	TARLAC	1	43
67	Purok 2	Cabuluan	VICTORIA	TARLAC	1	100
68	Purok 2	Cabuluan	VICTORIA	TARLAC	2	2
85	Purok 2	San Franci	VICTORIA	TARLAC	1	38
101	Purok 3	Cabuluan	VICTORIA	TARLAC	1	96
115	Purok 3	San Franci	VICTORIA	TARLAC	1	127
142	Purok 4	San Franci	VICTORIA	TARLAC	1	36
167	Purok 5	San Franci	VICTORIA	TARLAC	1	41

6. The next table in the report gives information on what households were digitized twice. Copy this table and paste it in excel.

CBMS Statistics Simulator Merge/Match Report
Households with more than one instance in digitized

household	purok	barangay	municipality	province	household instance
174	Purok 1	Balayang	VICTORIA	TARLAC	2
571	Purok 1	Balayang	VICTORIA	TARLAC	2
651	Purok 1	Balayang	VICTORIA	TARLAC	2
51	Purok 1	Balbaloto	VICTORIA	TARLAC	2
19	Purok 1	Bantog	VICTORIA	TARLAC	2
31	Purok 1	Bantog	VICTORIA	TARLAC	2
32	Purok 1	Bantog	VICTORIA	TARLAC	2
42	Purok 1	Bantog	VICTORIA	TARLAC	2
16	Purok 1	Cabuluan	VICTORIA	TARLAC	2
265	Purok 1	Calibungan	VICTORIA	TARLAC	2
392	Purok 1	Calibungan	VICTORIA	TARLAC	2

7. Filter the table showing only the information for Brgy. Cabuluan. In the example, household IDs 16, 254 and 259 should be checked in the NRDB file. If indeed the households were digitized twice then delete one dot for each household. Again, this example shows that even if the excel report indicates that all households are matched it still necessary to check if there are duplicates in the digitized household.

	A	B	C	D	E	F
218	**CBMS Statistics Simulator**					
219	**CBMS Statistics Simulator Merge/Match Report**					
220	*Households with more than one instance in digitized*					
221						
222	household	purok	barangay	municipality	province	household instance
231	16	Purok 1	Cabuluan	VICTORIA	TARLAC	2
266	254	Purok 2	Cabuluan	VICTORIA	TARLAC	2
267	259	Purok 2	Cabuluan	VICTORIA	TARLAC	2

8. A summary table on the result of the matching process is also included in the StatSim report. The number of zero cases in the excel report should be the same with the data in the summary table.

CBMS Statistics Simulator
~match: -1=Found in digitized but not in encoded; 1=Found in encoded but not in digitized; 0=Matched
Household Instance: Number of occurence of household, Frequency: Instances of occurence

~match	household instance	frequency
-1	1	10205
0	1	571