15th National Convention on Statistics

Development of Geo-enabled Master Sample Frame Prototype Model from Preliminary 2020 CPH Data Files in Marinduque and Batanes Using Python QGIS and R

By

Florante C. Varona
Johanna G. Abad

For additional information, please contact:

Authors' name	Florante C. Varona
Designation	OIC-Assistant National Statisticain
Affiliation	Philippine Statistics Authority
Address	17th Floor Cypebod Centris Three, cor. EDSA and Quezon Avenue, Quezon City
Tel. no.	277-3733
E-mail	f.varona@psa.gov.ph; varonaflorante@gmail.com

Co-Authors' name	Johanna G. Abad
Designation	Senior Statistical Specialist
Affiliation	Philippine Statistics Authority
Address	17th Floor Cypebod Centris Three, cor. EDSA and Quezon Avenue, Quezon City
Tel. no.	277-3733
E-mail	J.Abad@psa.gov.ph; jgabad21@gmail.com

Development of Geo-enabled Master Sample Frame Prototype Model from Preliminary 2020 CPH Data Files in Marinduque and Batanes Using Python QGIS and R

by

Florante C. Varona[1]
Johanna G. Abad[2]

ABSTRACT

This paper aims to introduce procedures in the development of geo-enabled master sample frame prototype model from the preliminary results of 2020 CPH in the provinces of Marinduque and Batanes using Python QGIS and R. This proposed approach will automate the PSU formation process to reduce substantially the time required for the task. In the previous 2013 master sample frame development, the PSU formation process would require several months to complete since the task is still being done manually using census cartographic maps. The 2023 geo-enabled master sample is expected to be used for the first time in July 2023 for the 2023 Family Income and Expenditure Survey (FIES) survey round, the only way to achieve this target is to innovate and automate processes in the master sample frame development.

The 2023 geo-enabled master sample frame will be constructed based on the results of the 2020 CPH with the corresponding 2020 CPH post-EA Reference File as the initial MS PSU frame while the lists of housing units/households containing HH head names and addresses will be extracted from the geo-validated 2020 CPH Form 2 and 2020 CPH Form 3 data files to form the Secondary Sampling units (SSUs) for each PSU.

I. Introduction

A statistical sampling frame requires accurate, complete, and up-to date information to represent the target population. The Philippine Statistics Authority (PSA) develops and maintains its own sampling frame for its household-based surveys and other statistical operations that it may serve. This sampling frame is extracted from the latest census data that the agency has. Being that the PSA has successfully conducted the 2020 Census of Population and Housing (2020 CPH) in September of the reference year, a new sampling frame or master sample can then be developed using the census results. Currently, the 2013 Master Sample (2013 MS) is being used for all household-based surveys of the PSA, including surveys of other agencies such as the Consumer Finance Survey, Consumer Expectations Survey, National Nutrition Survey, etc. The 2013 MS uses data from the 2015 Census of Population.

As part of the enhancement of the survey operations and statistical outputs of the PSA, the first use case of the new master sample is targeted to be the 2023 Family Income and Expenditure Survey. To be able to deliver according to this

[1] OIC-Assistant National Statistician, National Censuses Service (NCS)
[2] Senior Statistical Specialist of the Census Planning and Coordination Division (CPCD), National Censuses Service (NCS)

timeline, procedures in the development stages should be re-visited and innovate where possible to make the entire process efficient - not only cutting on the time that will be dedicated to the development of the new sampling frame, but also the automation and streamlining of specific steps will be cost-efficient compared to the previous master sample.

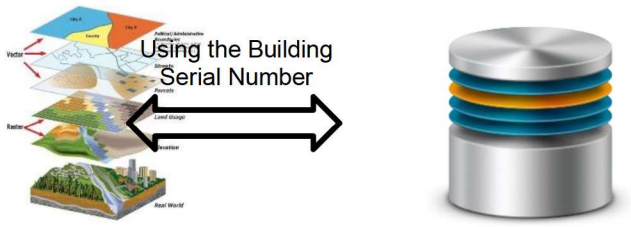
This research aims to introduce procedures in the development of geo-enabled master sample frame using Python QGIS and R through the prototype model generated from the preliminary results of 2020 CPH in the provinces of Marinduque and Batanes. This proposed approach will automate the Primary Sampling Unit formation process to substantially reduce the time required for the task. In the previous 2013 MS frame development, the PSU formation process required several months to complete since it was done manually using census cartographic maps. By utilizing the geospatial data that PSA has acquired and created over time, what was done for months prior can now be done in a shorter span of time, provided that inputs are complete and has been pre-processed accordingly.

II. Integration of Statistical and Geospatial Information

Geospatial data typically combines descriptive information about a data point and its relative location information. Geospatial data helps the data user visualize where and on what surface on the earth the data point is located which provides a more informed analysis on a dataset. As PSA, we have an archive of survey and census data, most of which are cross-sectional data wherein only common statistical analysis can be done or only conclusions on aggregated data can be inferred.

By integrating geospatial and statistical data one can pinpoint where in a barangay or enumeration area a specific data point is located. Not only will this help in actual survey operations-leading to lesser or no occurrence of sample housing units that cannot be located-but most of all in providing an additional perspective or insights in the evaluation and analysis of statistical survey results.

To be able to integrate the two datasets, we will use the Building Serial Number (BSN) assigned to a building structure during the 2020 CPH. This identifier will allow us to make our map and census data meet.



Geospatial data
(Including admin maps, EA maps, geotagged building points, road networks, rivers, landmarks, etc.)

Statistical Data
(Collected census and survey data)

III. 2023 Geo-enabled Master Sample (GeoMS)

The starting point of the development of a geo-enabled master sample frame is preparing two initial frames: the Primary Sampling unit (PSU) frame and the Secondary Sampling Unit (SSU) frame.

The PSU frame contains the list of Enumeration Areas (EAs)/barangays/merged EAs which represents the area frame of a sampling domain. In the 2013 MS and the 2023 GeoMS, the provinces and Highly Urbanized Cities (HUCs) are set as the sampling domain. The geospatial counterpart of the PSU frame are the barangay and EA maps, which are polygons digitized on the base map.

Meanwhile, the SSU frame is the list of housing units and its corresponding households enumerated during the census. The SSU frame has an identifier to which PSU a housing unit/household is included. Its geospatial counterpart are the geotagged building points that the PSA has collected since 2016. Each geotagged point refers to a building structure, it can have a living quarter or not.

The 2023 GeoMS will use the EA Reference file of the 2020 CPH as its initial PSU frame prior to evaluation and merging of qualified areas. For the SSU frame, the enumerated households with the name of household head, address and identified demographic information will be extracted from CPH Forms 2 and 3. This paper will focus on the innovation of the PSU and SSU formation phase of the MS frame development.

A PSU is defined to be an area that contains 100 to 300 households within its bounds. This can be a lone EA/barangay or formed by merging two or more smaller EAs/barangays. To create the final PSU frame of a sampling domain, one must evaluate first the size of all EAs within the domain to check which areas has to be merged to reach the size measure of 100 to 300 households. However, prior to merging, some EAs/barangays are excluded in the initial PSU formation, such as zero-household EAs and Least Accessible Areas.

For the 2023 GeoMS, the Field Offices (FOs) identified Least Accessible Areas in their provinces for exclusion in the PSU frame of their sampling domain. Least Accessible Areas can be a i) Least Accessible Barangay if the while barangay fits the criteria of a LAA; ii) Least Accessible EA if only a portion of the barangay fits the criteria of a LAA and that portion has been delineated as EA during the 2020 CPH; iii) Least Accessible Purok/Sitio/Segment if the portion of the barangay that fits the criteria of a LAA is not delineated as separate EA during the 2020 CPH. All EAs or barangays in the final list of LAAs of a sampling domain is to be excluded in its PSU frame.

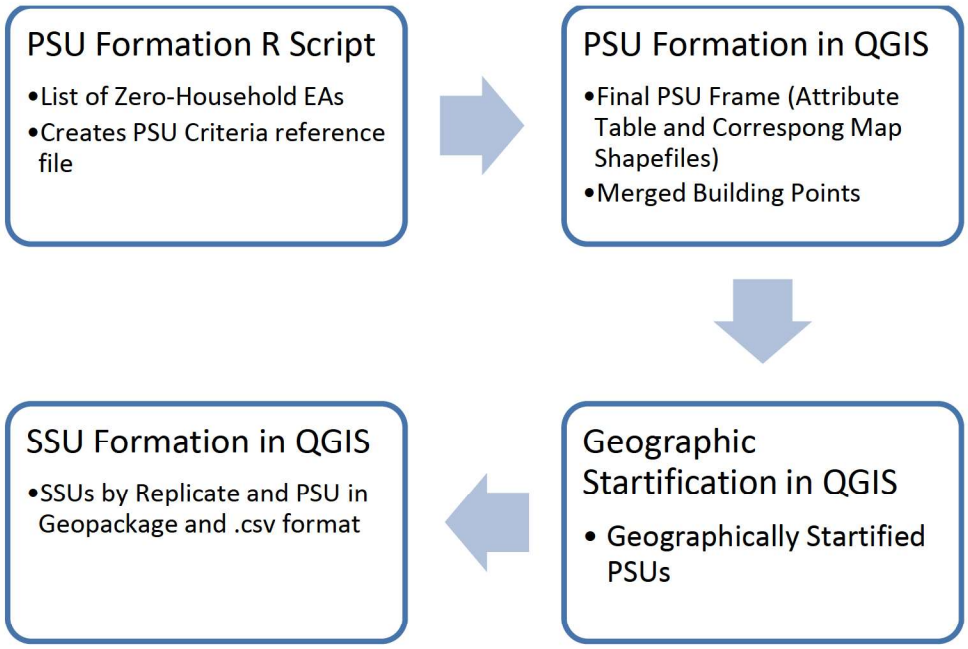
The criteria for merging areas that have less than 100 households to its adjacent EAs/barangays to form a PSU are:

1. The sum of the number of households of the resulting PSU must be within 100 to 300 households
2. The merged area units must come from within the same barangay (if both are EAs) or must come from within the same municipality (if both are barangays)
3. The areas for merging must have the same urbanity
4. The resulting PSU must be accessible within itself

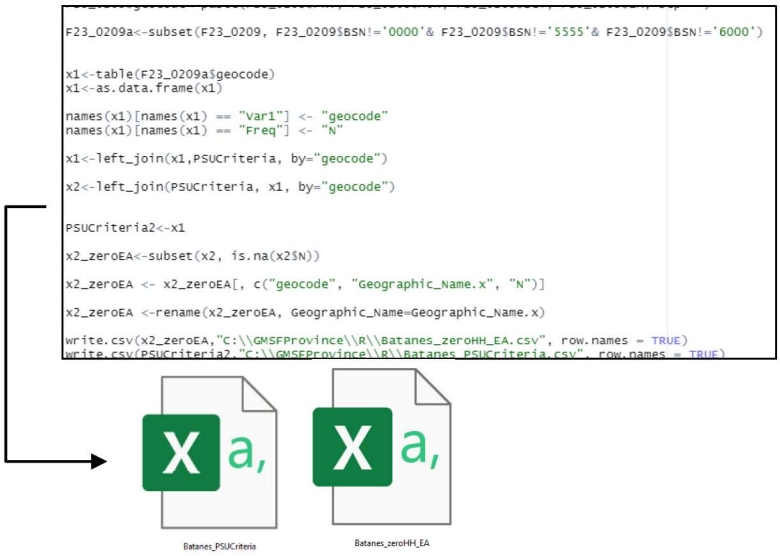
All the criteria stated above, except for the last one, is included in the considerations of the automated PSU formation program.

The complete procedure for the PSU and SSU formation using R and Python QGIS involves running four scripts in two separate software: (*See Figure 1: Process Flow for PSU and SSU Formation*)

Figure 1: Process Flow for PSU and SSU Formation

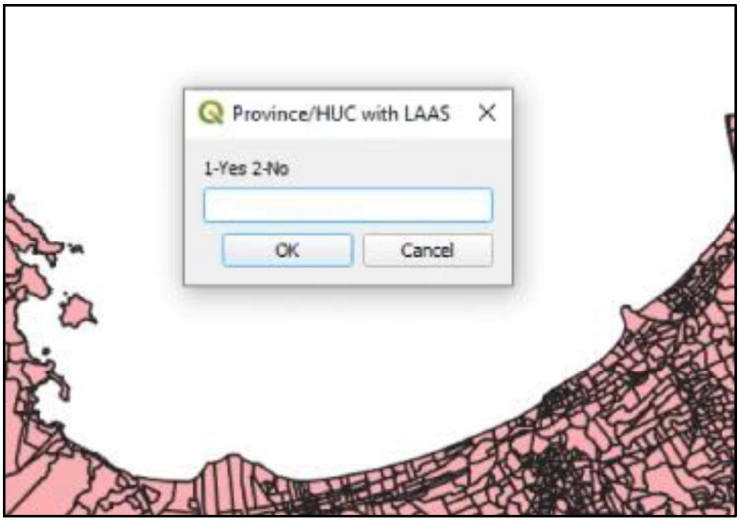


1. **PSU Formation** R script – This will generate the list of EAs/barangays with no listed households during the 2020 CPH based on its Form 2 and 3. Also, the PSU Criteria reference file which has the Urbanity Indicator and Number of households per EA. The output files will be used as reference to run the automated PSU formation.

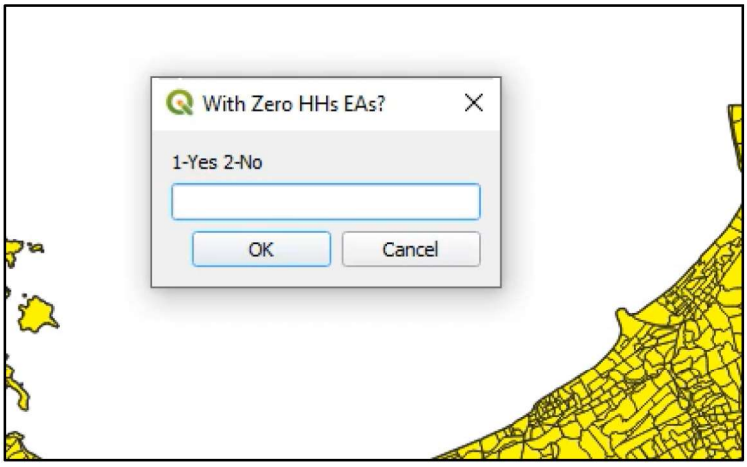


2. **PSU Formation** Python code in QGIS – This code will initiate the PSU formation process in the sampling domain. By asking for inputs for each phase of the procedure, the Python code automatically excludes, merges, and extracts the attribute information together with the map shapefiles.

   a. <u>Exclusion of Least Accessible Areas</u> (if applicable) – The user will input either 1 (The sampling domain has approved LAAS) or 2 (There are no LAAs in the sampling domain). To assure correctness of input, the user is advised to use the final list of LAAs for the domain as reference.



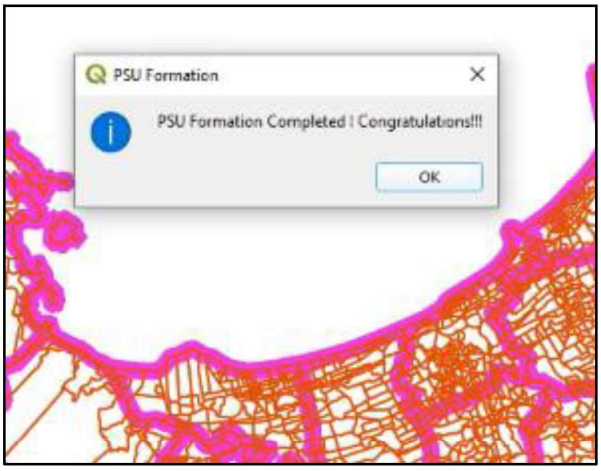
b. Exclusion of Zero Household EAs (if applicable) – EAs or barangays with no listed households is to be excluded in the PSU frame of a sampling domain. This is done during the initial stages to ensure that the EA or barangay will not be included in the "for merging" areas. Similar to the previous step, the user shall encode "1" referring to existence of zero household EA in the domain or "2" for none.



c. Merging of EAs/barangays with less than 100 households – After excluding LAAs and Zero household EAs, the program automatically proceeds to:

i. identification of qualified areas for merging
ii. evaluating its adjacent areas and identify to which area it can be merged
iii. merge the attribute information and map layers to correspond to the initially formed PSUs

Afterwards, a pop-up will show on the QGIS of the user such as:



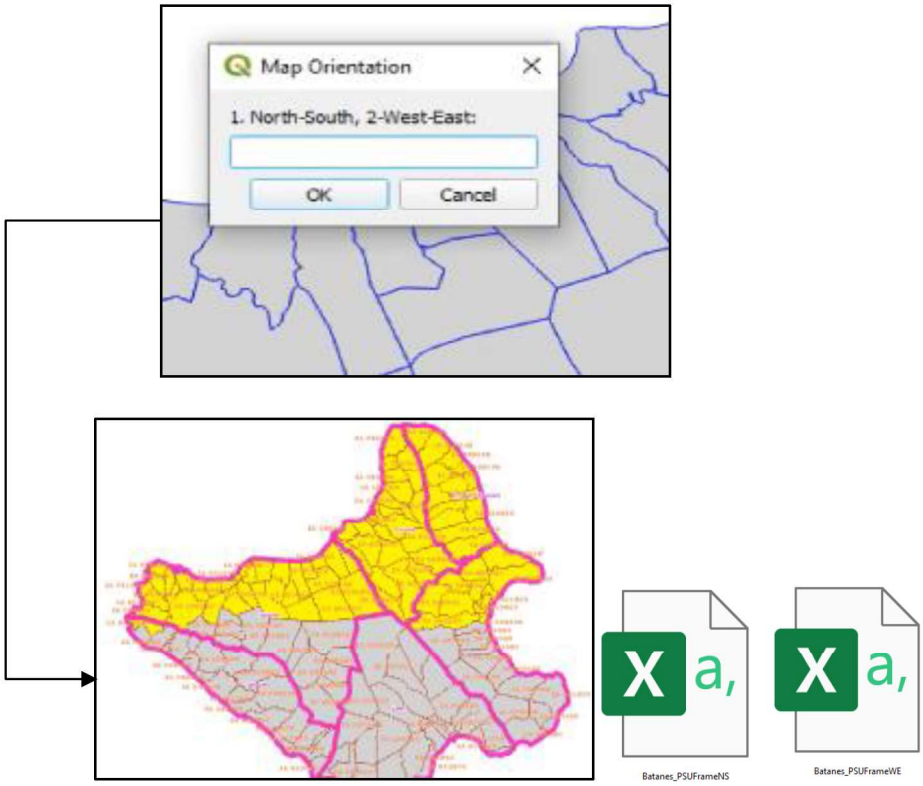
Note: An important criterion in the finalization of PSUs is also included in the program. The total number of PSUs to be formed must be divisible by the allotted replicate size (e.g. 3 PSUs (small domains), 6 PSUs (regular province), 8 PSUs (regular HUCs)).

```
--------------------------------------------------------------------------------
Replicate Divisibility Requirement in PSU Formation
--------------------------------------------------------------------------------

The total number of unaffected barangays/EAs for the initial PSU Formation= 3731
The total number of merged barangays/EAs for the initial PSU Formation= 22
The replicate size for this province/huc domain = 6
The total number of replicates that can be created =625.5
The total initial number of PSUs (3753) is not yet divisible by the replicate size (6)
The required total number of PSUs to make it divisible by the replicate size (6) is 3756
There is a need to unmerge from the dissolve_case2 layer a total of 3 set/s of merged barangays/EAs
--------------------------------------------------------------------------------
```

The program will also automatically adjust the number of merged PSUs according to the allotted replicate size for the sampling domain. Once done the replicate divisibility requirement should print a congratulatory message.

```
Replicate Divisibility Requirement in PSU Formation (With Adjustment)
--------------------------------------------------------------------------------

The total adjusted number of unaffected barangays/EAs for the initial PSU Formation= 3737
The total adjusted number of merged barangays/EAs for the initial PSU Formation= 19
The replicate size for this province/huc domain = 6
The adjusted total number of replicates that can be created =626.0
The total adjusted number of PSUs(3756) is already divisible by the replicate size (6)
--------------------------------------------------------------------------------
 Congratulations !!!
```

3. **Geographic Stratification** Python code in QGIS – After the PSUs of a domain has been finalized, the PSUs shall undergo implicit geographic stratification. This will endure that the sampling units will represent the entire physical area of the sampling domain. The Python code will sort and group the formed PSUs into its respective geographic location based on the geographic orientation (North-South or East-West) inputted in the Python code.



4. **SSU Formation** Python code in QGIS – When all details and corresponding geospatial data for the PSUs has been completed, the merged geotagged points for the domain can then be overlayed on the PSU maps and from which the PSU id and other attributes will be joined to the geotagged points. The census data will also be integrated to the geospatial data during this step for provinces that used Paper and Pen Personal Interview (PAPI) during the census.

IV. Prototype Applied to Map-based Census CAPI Area (Batanes and Marinduque)

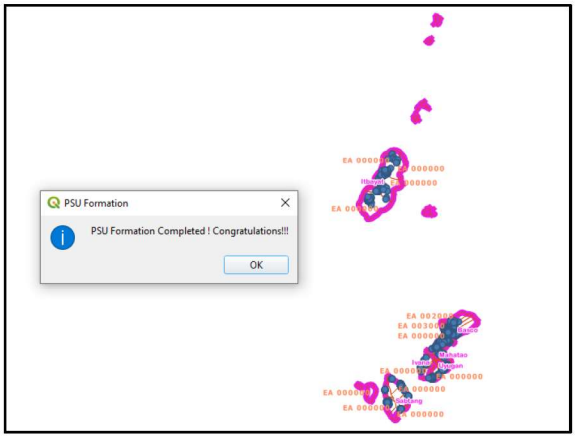
Using the preliminary data from the 2020 CPH, we can apply the prototype model to the provinces of Batanes and Marinduque. Being that the two provinces used Computer-Assisted Personal Interview (CAPI), the integration of census data to the geotagged points will be skipped as it is already integrated.

Batanes

Number of EAs during the 2020 CPH: 38 EAs
Number of LAAs: None
Number of Zero Household EAs: None

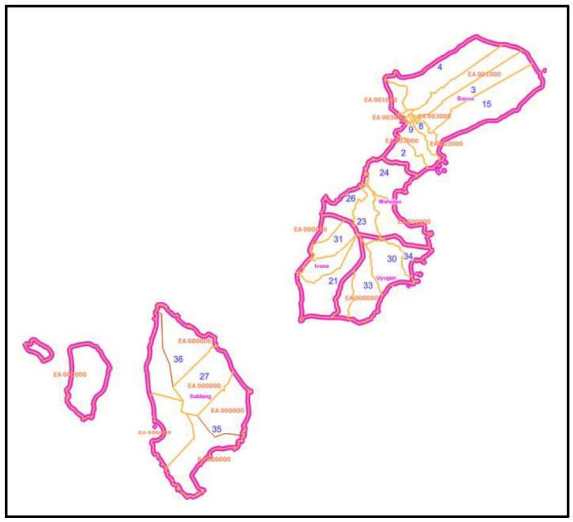
After providing the inputs for criteria for exclusion, the program will proceed in the evaluation and merging of EAs/barangays as necessary. Prior to the creation of final PSU layer, the program will ask the user to input the replicate size of the sampling domain. For Batanes, a small province, the replicate size set during the 2013 MS and retained for the 2023 GeoMS is 3 PSUs per replicate.

After ensuring that the number of final PSUs is divisible by the set replicate size, a pop-up will show that the PSU formation process has been completed.

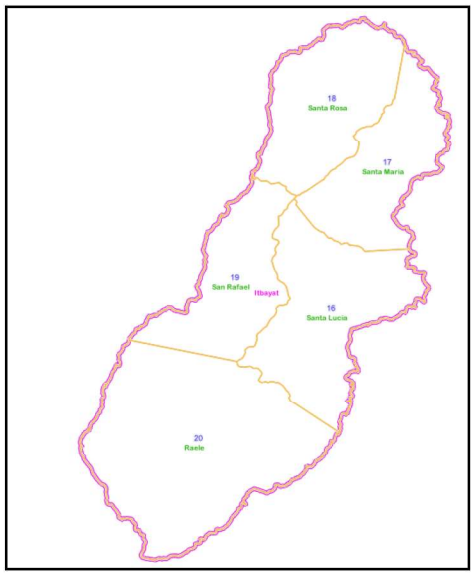


The program automatically creates a separate layer for each step completed so that the user can trace the merging process backwards to ensure that only qualified EAs has been merged to its adjacent and criteria-fit neighbors.

Number of PSUs formed: 36 PSUs



The 2023 GeoMS PSUs formed for the sampling domain of Batanes with its initial PSU number and corresponding polygon layer. The picture on the left shows the PSUs for the main islands of the province containing the municipalities of Sabtang, Basco, Mahatao, Uyugan, and Ivana.



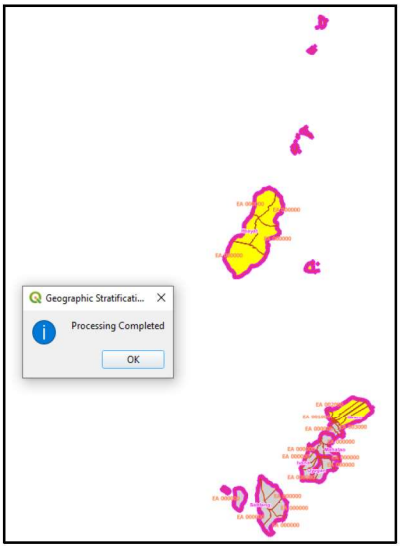
The picture on the left shows the PSUs for the main island of the municipality of Itbayat.

After forming the initial PSUs of the domain, the province will then conduct field validation, specifically on the PSUs formed by merging EAs/barangays. The FOs will validate the accessibility within the merged areas to ensure that the formed PSUs can be easily traversed by the enumerator during survey operations. An access road will be drawn by the field verifier to serve as evidence that the EAs/barangays can be accessed within each other. This process will be done on October during the PSU and SSU validation field operations of the Census Planning and Coordination Division (CPCD).

Assuming that the FO has completed the validation of the PSUs and that all merged areas are accessible within each other, the user can then proceed to the first implicit stratification for the 2023 GeoMS: by Geographic Orientation.

Since the topographic orientation of the province of Batanes is vertical, it is logical to stratify its PSUs from North to South.

Topographical Orientation of the Domain: North-South



Similar to the previous steps, a dialog box will pop-up to inform the user if the processing has been completed.

As seen in the illustration, one can easily visualize the geographic stratification done by the automated program.

The attribute table of the PSU layer will contain all variables used for the PSU formation such as the number of households listed for the EA and its strata for the geographic stratification.

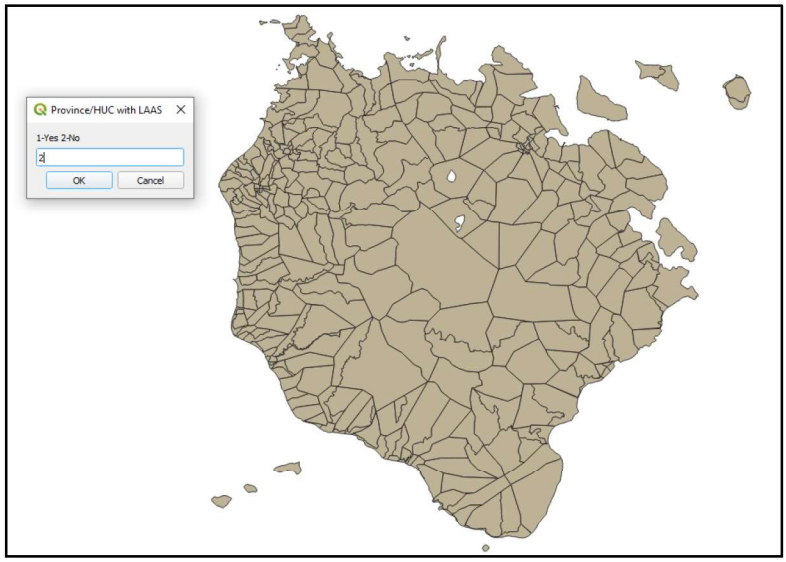
	fid	geocode	EA_Name	PSU_Name	reg	prv	mun	N	PSU_number	xcoord	ycoord	geographic_stratum
1	1	0901006001000	EA 001000	Kayhuvokan - E...	2	9	1	175	1	121.970195	20.449287	1
2	2	0901005000000	EA 000000	Chanarian	2	9	1	155	2	121.967261	20.42982	2
3	3	0901001004000	EA 004000	Ihubok II (Kayv...	2	9	1	201	3	122.002075	20.463192	1
4	4	0901003002000	EA 002000	San Antonio - E...	2	9	1	213	4	121.985459	20.471802	1
5	5	0901001003000	EA 003000	Ihubok II (Kayv...	2	9	1	218	5	121.973294	20.44809	2
6	6	0901001001000	EA 001000	Ihubok II (Kayv...	2	9	1	258	6	121.998139	20.468871	1
7	7	0901001002000	EA 002000	Ihubok II (Kayv...	2	9	1	197	7	121.972291	20.449441	1
8	8	0901006002000	EA 002000	Kayhuvokan - E...	2	9	1	246	8	121.973772	20.447104	2
9	9	0901002002000	EA 002000	Ihubok I (Kaych...	2	9	1	193	9	121.969852	20.446134	2
10	10	0901002001000	EA 001000	Ihubok I (Kaych...	2	9	1	264	10	121.968334	20.448342	1
11	11	0901006003000	EA 003000	Kayhuvokan - E...	2	9	1	209	11	121.979229	20.438191	2
12	12	0901002003000	EA 003000	Ihubok I (Kaych...	2	9	1	182	12	121.973571	20.435756	2
13	13	0901003003000	EA 003000	San Antonio - E...	2	9	1	262	13	121.967116	20.450959	1

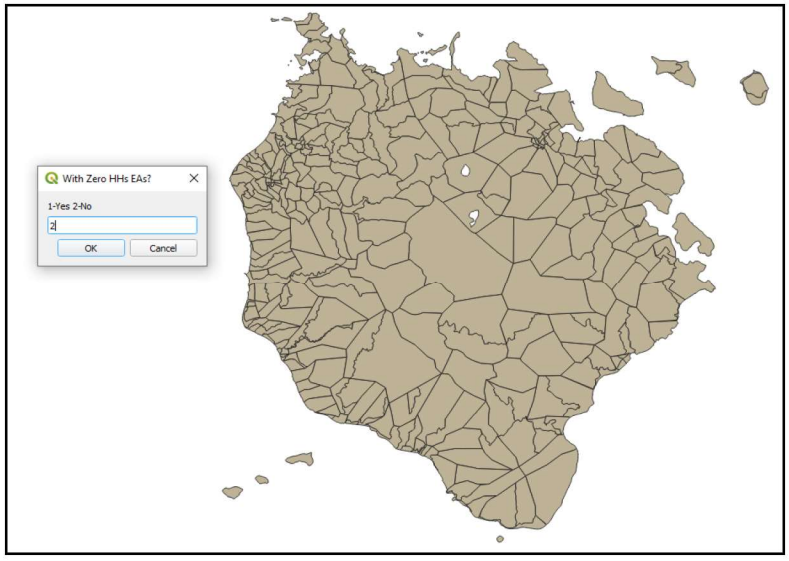
Since the province of Batanes used CAPI, visualizing the data points on the PSU for which it belongs to is as simple as overlaying the building points on the existing uploaded map layers.

Similarly, the procedure can then be applied to the province of Marinduque. Being an island province, there are also no Least Accessible Areas identified for this sampling domain.

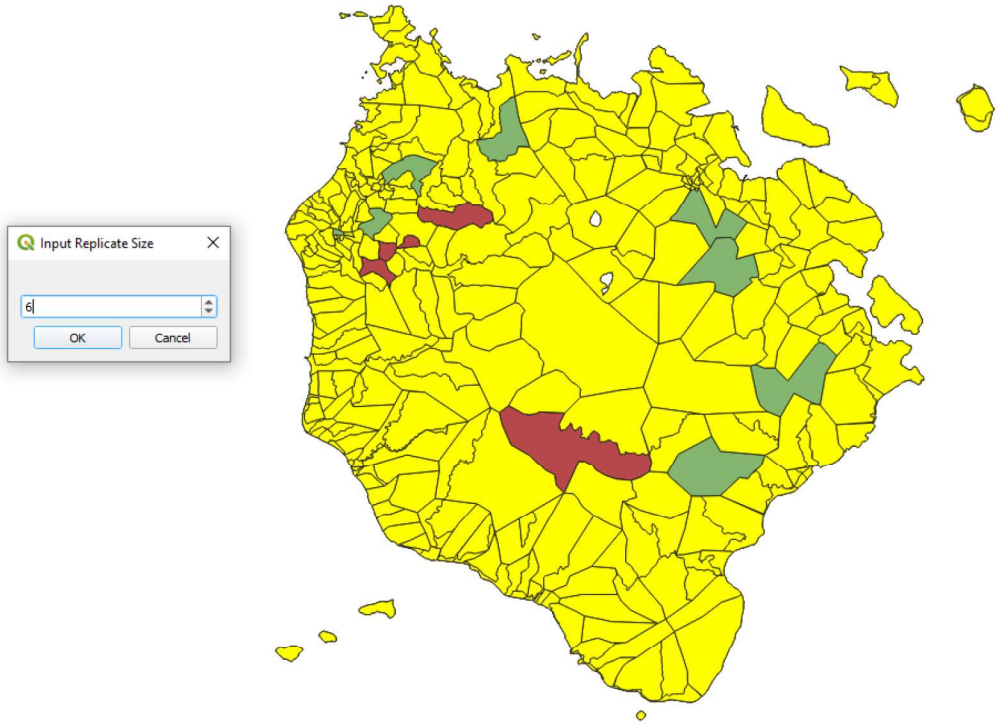
<u>Marinduque</u>

Number of EAs during the 2020 CPH: 326 EAs
Number of LAAs: None
Number of Zero Household EAs: None

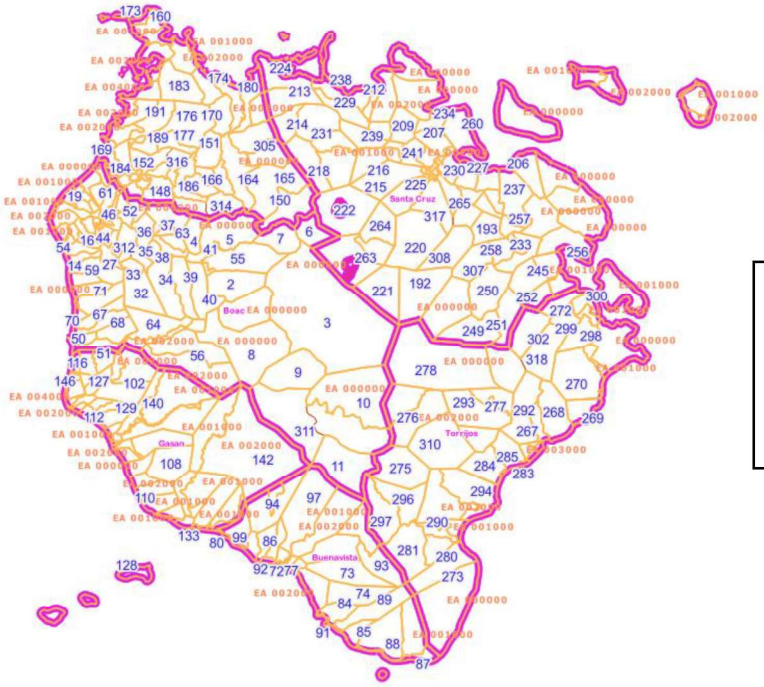




Merging of PSUs to Finalization of PSUs formed following the Replicate Divisibility Rule:

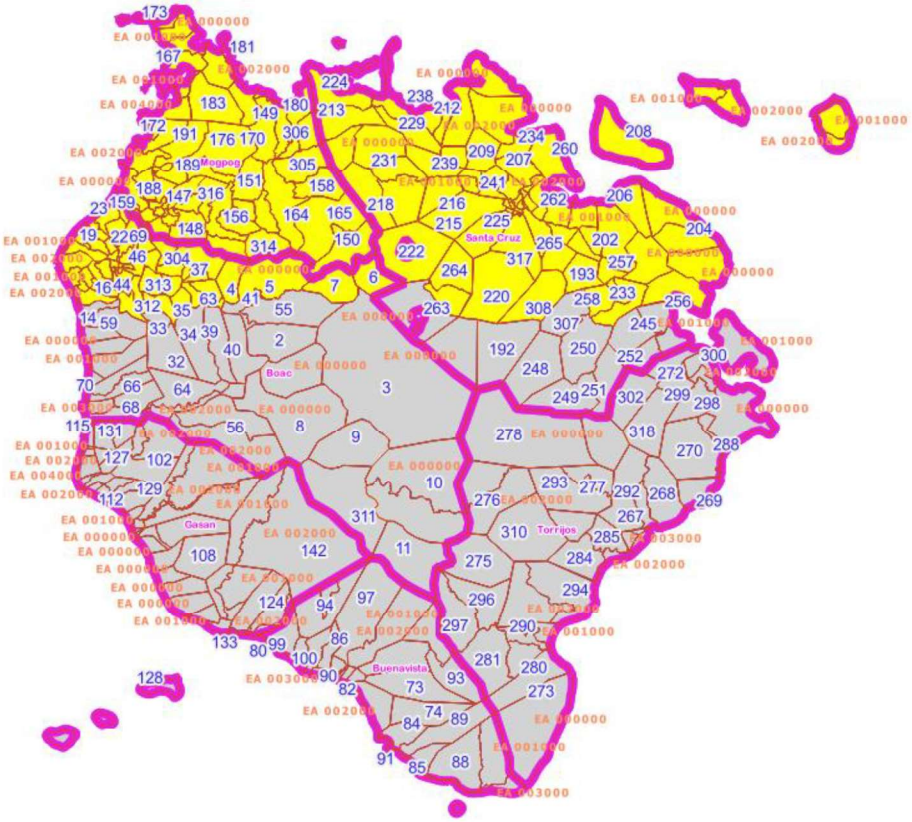


Being a regular-sized province, the allotted number of PSUs per replicate for this domain is six PSUs.



The total number of PSUs formed for the sampling domain of Marinduque using the automated program is 318 PSUs.

Once the FOs has validated that all PSUs resulting from merged EAs/barangays are accessible within each other, then the user can run the Geographic Stratification code for the province of Marinduque. Given its topographical orientation, one can use North to South as the orientation for the process.



Note that the geographic stratification does not only consider the exact geographical center of the domain in dividing it into two strata, but that the number of households is almost equally distributed into the strata.

V. Way Forward

With the need to deliver the 2023 Geo-enabled Master Sample in time for the 2023 FIES, having these automated programs using R and Python QGIS will not only significantly shorten the time needed for the evaluation and processing of PSU formation**,** but will also streamline and lighten the manual work of the Central Office personnel who will create the initial PSU frames for all sampling domains. Compared to the process employed for the 2013 MS development, one can surely see the advantage of having the automated programs to create not only reference files for two-dimensional data but turning the notch one level higher – visualizing statistical information through integration to geospatial data.