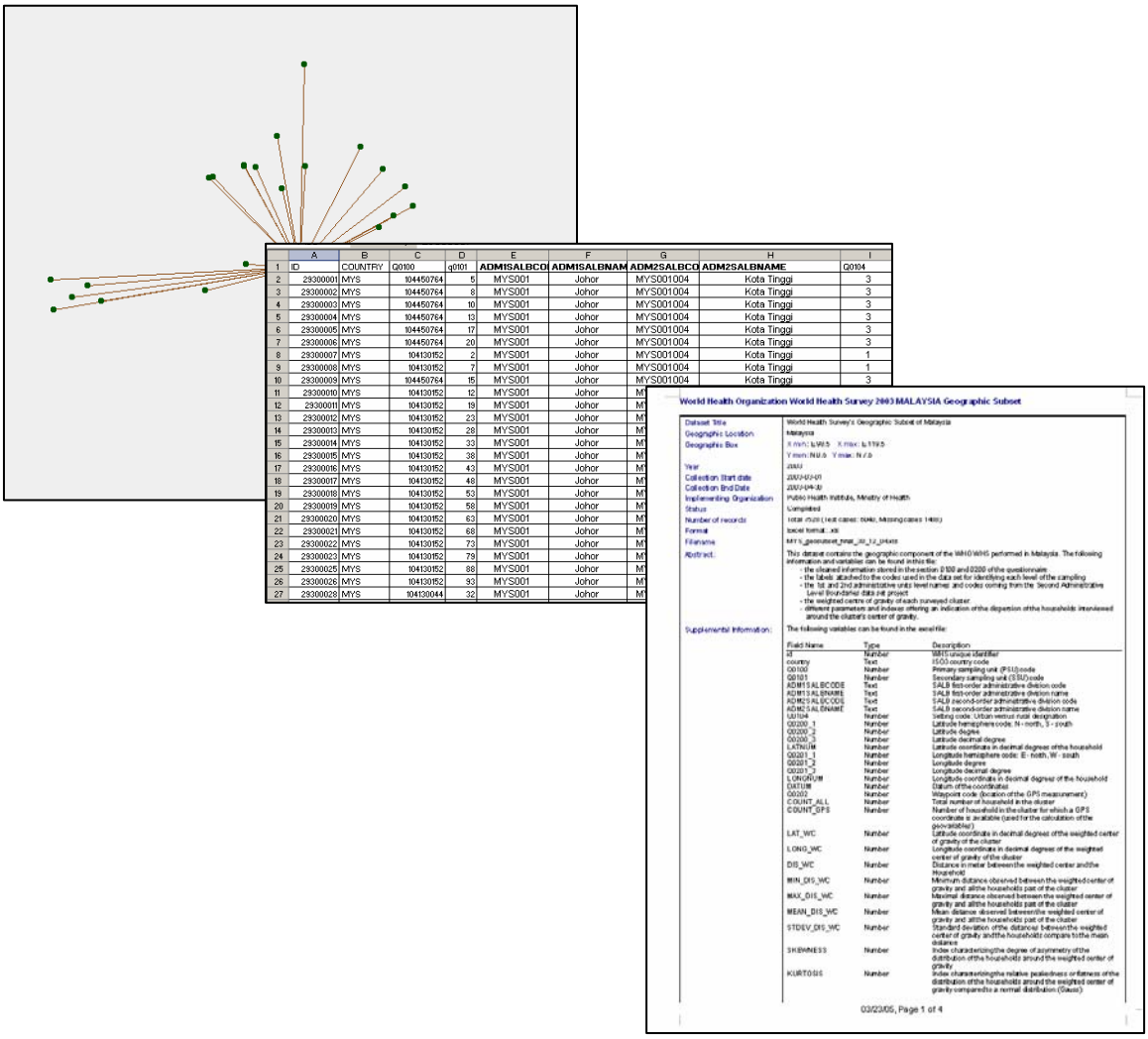
WHS

World Health Survey

Generation of the GEO Subset

Countries using GPS devices



	A	B	C	D	E	F	G	H	I
1	ID	COUNTRY	Q0100	q0101	ADM1SALBCO	ADM1SALBNAM	ADM2SALBCO	ADM2SALBNAME	Q0104
2	29300001	MYS	104450764	5	MYS001	Johor	MYS001004	Kota Tinggi	3
3	29300002	MYS	104450764	8	MYS001	Johor	MYS001004	Kota Tinggi	3
4	29300003	MYS	104450764	10	MYS001	Johor	MYS001004	Kota Tinggi	3
5	29300004	MYS	104450764	13	MYS001	Johor	MYS001004	Kota Tinggi	3
6	29300005	MYS	104450764	17	MYS001	Johor	MYS001004	Kota Tinggi	3
7	29300006	MYS	104450764	20	MYS001	Johor	MYS001004	Kota Tinggi	3
8	29300007	MYS	104130152	2	MYS001	Johor	MYS001004	Kota Tinggi	1
9	29300008	MYS	104130152	7	MYS001	Johor	MYS001004	Kota Tinggi	1
10	29300009	MYS	104450764	15	MYS001	Johor	MYS001004	Kota Tinggi	3
11	29300010	MYS	104130152	12	MYS001	Johor	M		
12	29300011	MYS	104130152	19	MYS001	Johor	M		
13	29300012	MYS	104130152	23	MYS001	Johor	M		
14	29300013	MYS	104130152	28	MYS001	Johor	M		
15	29300014	MYS	104130152	33	MYS001	Johor	M		
16	29300015	MYS	104130152	38	MYS001	Johor	M		
17	29300016	MYS	104130152	43	MYS001	Johor	M		
18	29300017	MYS	104130152	48	MYS001	Johor	M		
19	29300018	MYS	104130152	53	MYS001	Johor	M		
20	29300019	MYS	104130152	58	MYS001	Johor	M		
21	29300020	MYS	104130152	63	MYS001	Johor	M		
22	29300021	MYS	104130152	68	MYS001	Johor	M		
23	29300022	MYS	104130152	73	MYS001	Johor	M		
24	29300023	MYS	104130152	79	MYS001	Johor	M		
25	29300025	MYS	104130152	88	MYS001	Johor	M		
26	29300026	MYS	104130152	93	MYS001	Johor	M		
27	29300028	MYS	104130044	32	MYS001	Johor	M		

World Health Organization World Health Survey 2003 MALAYSIA Geographic Subset

Dataset Title	World Health Survey's Geographic Subset of Malaysia
Geographic Location	Malaysia
Geographic Box	X min: 99.6 X max: 119.5 Y min: 0.8 Y max: 7.0
Year	2003
Collection Start date	2003-01-01
Collection End Date	2003-04-30
Implementing Organization	Public Health Institute, Ministry of Health
Status	Completed
Number of records	total: 7529 (Hot cases: 6948, Missing cases: 1481)
Format	local format: .xls
Filename	MYS_geosubset_year_00_12_04.xls
Abstract:	This dataset contains the geographic component of the WHO WHS performed in Malaysia. The following information and variables can be found in this file: - the cleaned information stored in the section: B100 and B200 of the questionnaire - the labels attached to the codes used in the data set for identifying each level of the sampling - the 1st and 2nd administrative units level names and codes coming from the Second Administrative Level Boundaries data set project - the weighted centre of gravity of each surveyed cluster - different parameters and indexes offering an indication of the dispersion of the households interviewed around the cluster's center of gravity.
Supplemental Information:	The following variables can be found in the excel file:

Field Name	Type	Description
id	Number	WHS unique identifier
country	Text	ISO3 country code
Q0100	Number	Primary sampling unit (PSU) code
Q0101	Number	Secondary sampling unit (SSU) code
ADM1SALBCODE	Text	SALB first-order administrative division code
ADM1SALBNAME	Text	SALB first-order administrative division name
ADM2SALBCODE	Text	SALB second-order administrative division code
ADM2SALBNAME	Text	SALB second-order administrative division name
U0104	Number	Setting code: Urban versus rural designation
Q0200_1	Number	Latitude hemisphere code, N - north, S - south
Q0200_2	Number	Latitude degree
Q0200_3	Number	Latitude decimal degree
LATNUM	Number	Latitude coordinate in decimal degree of the household
Q0201_1	Number	Longitude hemisphere code: E - north, W - south
Q0201_2	Number	Longitude degree
Q0201_3	Number	Longitude decimal degree
LONGNUM	Number	Longitude coordinate in decimal degree of the household
DATUM	Number	Datum of the coordinates
Q0202	Number	Waypoint code (location of the GPS measurement)
COUNT_ALL	Number	Total number of household in the cluster
COUNT_GPS	Number	Number of household in the cluster for which a GPS coordinate is available (used for the calculation of the geovariables)
LAT_WC	Number	Latitude coordinate in decimal degree of the weighted center of gravity of the cluster
LONG_WC	Number	Longitude coordinate in decimal degree of the weighted center of gravity of the cluster
DIS_WC	Number	Distance in meter between the weighted center and the Household
MIN_DIS_WC	Number	Minimum distance observed between the weighted center of gravity and all the households part of the cluster
MAX_DIS_WC	Number	Maximal distance observed between the weighted center of gravity and all the households part of the cluster
MEAN_DIS_WC	Number	Mean distance observed between the weighted center of gravity and all the households part of the cluster
STDEV_DIS_WC	Number	Standard deviation of the distance between the weighted center of gravity and the households compare to the mean distance
SKEWNESS	Number	Index characterizing the degree of asymmetry of the distribution of the households around the weighted center of gravity
KURTOSIS	Number	Index characterizing the relative peakedness or flatness of the distribution of the households around the weighted center of gravity compared to a normal distribution (Gauss)

03/23/05, Page 1 of 4

Generation of the GEO subset
Countries using GPS devices

Geneva, Switzerland

This document was prepared by Imed Ben Hamadi, Steeve Ebener, Fanny Naville and Zine El Morjani

1. Introduction

This document contains the protocol that has been implemented in the context of the World Health Organization World Health Survey (WHO WHS) for generating the GEO subset for countries where the GPS (Global positioning System) devices have been used.

The steps describes in this protocol could be repeated in other surveys as long as the geographic information collected correspond to the WHO WHS variables.

The final file resulting from the application of this protocol contains the geo variables observed for the test and missing cases.

1.1. WHO WHS Geographic component

The WHO WHS has been launched in 2001 within 71 countries located in the different WHO regions. It has been designed to fill existing data gaps, to supplement national and sub-national health information systems and to provide reliable and valid data in a cost-effective manner that can be used to inform policy debates.

GPS devices have been used in 27 of the 71 countries part of the survey in order to collect the location of each of the surveyed household representing a data set of more than 175'000 records.

By integrating Geography, the WHO WHS becomes the second biggest effort, after the DHS+, which collects the geographic location of the surveyed households adding therefore value to the survey itself.

In the context of the WHO WHS specific data collection protocol and data cleaning protocols have been used to ensure the homogeneity and quality of its geographic component. These protocols can be downloaded from the WHO WHS Web site at the following address: http://www3.who.int/whs/P/instrumentandrel8293.html.

The geographic component of the WHO WHS has been collected in two sections of the WHS questionnaire: The Sampling Information section (Figure 1) and the Geocoding Information section (Figure 2).

0100. Sampling Information (To be filled in by the supervisor)

	Sampling			
0101	Primary Sampling Unit (PSU) Name/Code			
0102	Secondary Sampling Unit (SSU) Name/Code			
0103	Tertiary Sampling Unit (TSU) Name/Code			
0104	Quarternary Sampling Unit (QSU) Name/Code			
	Additional Information			
0105	Setting	Urban 1	Peri-urban /Semi-urban 2	Rural 3
		Other 4	Specify: _____	

Figure 1 - Sampling Information section of the WHO WHS questionnaire

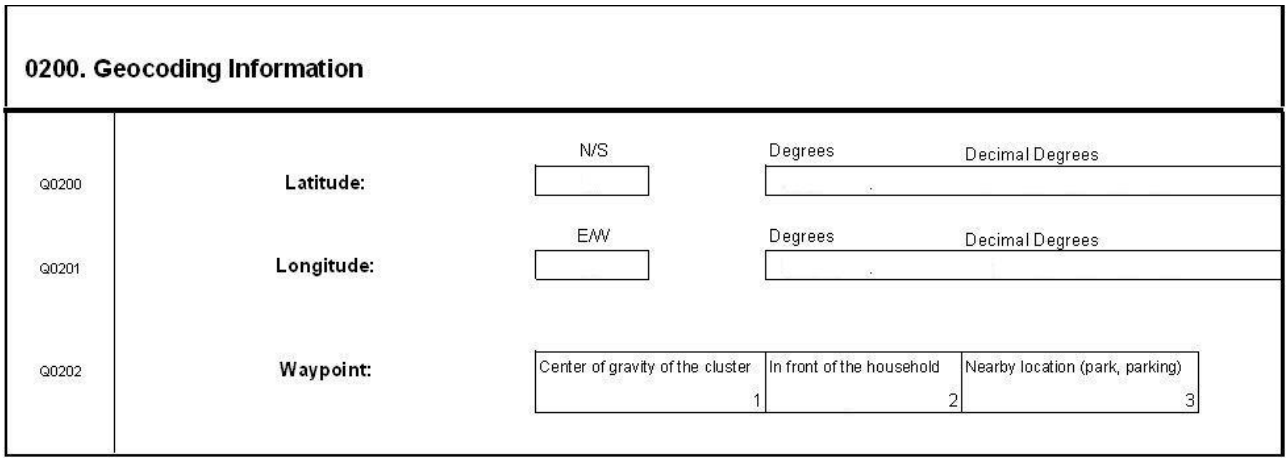


Figure 2 - Geocoding Information section of the WHO WHS questionnaire

The sampling codes, based on each country sampling plan (as drawn by the implementing organization and approved by WHO), have been collected in the sampling section, while the Geocoding Information section has been used to collect the GPS coordinates of each surveyed Households.

1.2. Generation of the WHS GEO Subset

The generation of the GEO subset is based on the ***ISO3*_geo_cleaned_*date*.xls** file resulting from the application of the processes reported in the "**Cleaning Protocol for the Geographic Component (Section 0100 and 0200), Countries using GPS devices"** document that can be downloaded from the WHO WHS web site (http://www3.who.int/whs/).

The present protocol is used to produce:

- a sub set of variables to be used in the context of spatial analysis or modelling
- a specific metadata record for its documentation.

As the GPS coordinates have been collected for each of the surveyed households, the generation of new variables also allows solving the confidentiality issues that might be raised in some countries by providing an alternative information regarding the location and extension of each cluster.

This protocol has been applied on a country by country basis.

1.2.1. Folders and Files organization

In order to homogenise and simplify the treatment of the different files a specific folder structure has been generated (Figure 3)

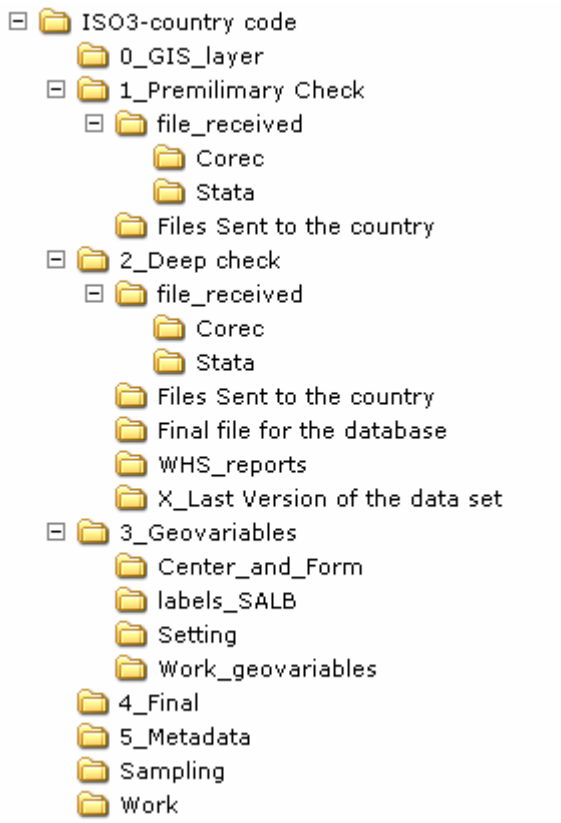


Figure 3 - Folder structure used during the application of the protocol

The folder in which each new file should be located is indicated in the protocol.

During the whole process the corresponding ISO3 country code is integrated in the file name in order to identify to which country each file correspond to.

In order to identify the different version of file resulting from a same operation each file name is ended by a date expressed using the following format: dd_mm_yy.

The whole process followed in the context of this protocol is illustrated in the Flow Chart reported in Figure 4.

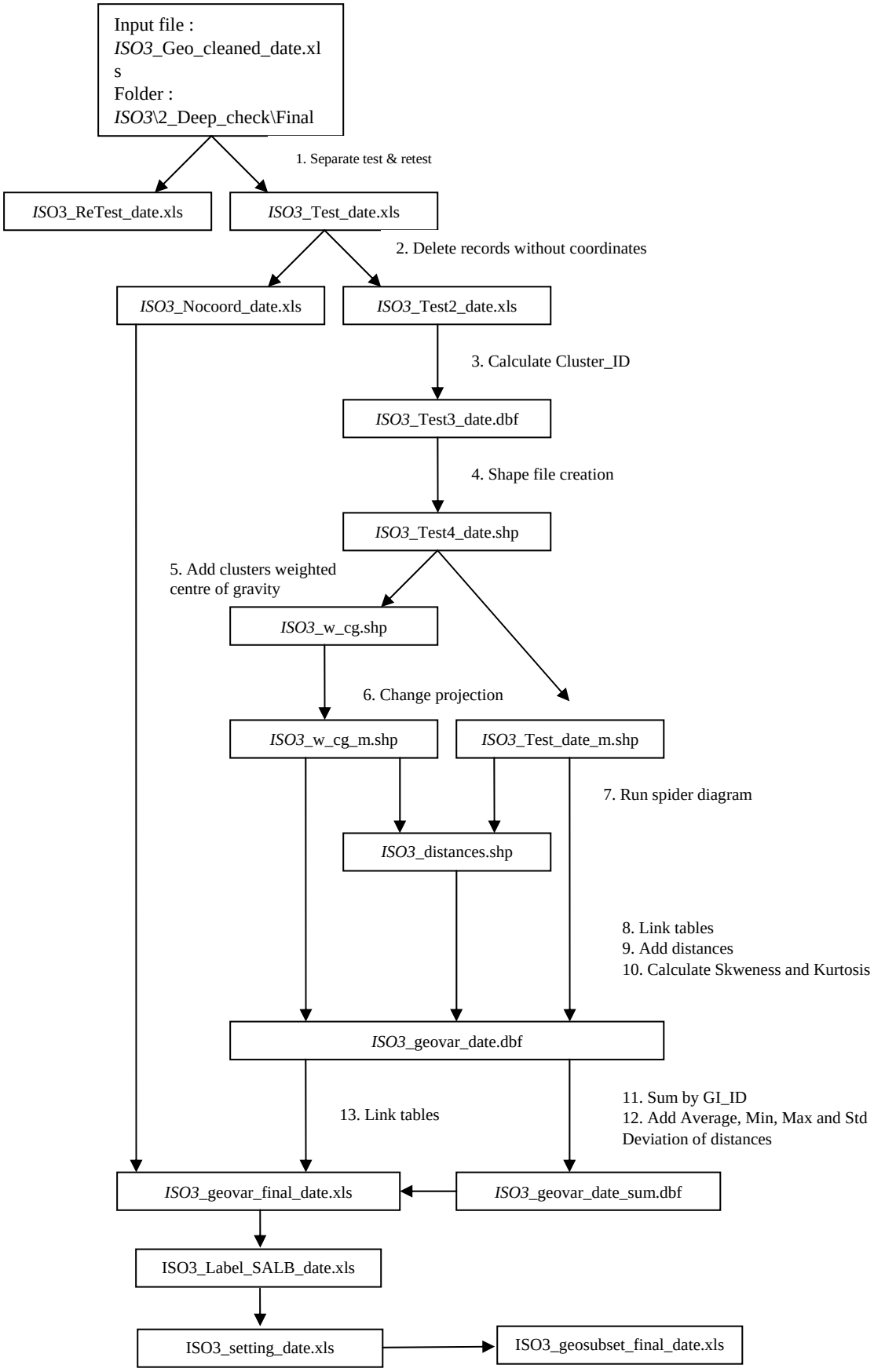


Figure 4 - Process and naming conventions used in the context of this protocol

1.2.2. Additional Materials

The following software are necessary in order to perform the steps described in this protocol:

- Excel
- ArcView 3.x or higher

In this protocol reference is done to a set of files and documents that can be found in the Annex_GEO_subset.zip file downloadable from the WHO WHS web site (http://www3.who.int/whs/P/instrumentandrel8293.html). These are:

- The Arcview project of reference (Ref_WHS.apr)
- ArcView extensions (prjctr.avx, XTOOLSMH.avx)
- ArcView script (spider.ave)

Country specific files, provided by the survey institutions, have also been used, these are:

- the *ISO3*_Samp_key_table_date.xls which contains the label for the codes enter in the data set for the different sampling level (PSU, SSU, TSU,…)

2. Preparation of the File

2.1. Separation between Tests and Retests

Firstly, as we let the retest untouched the separation between the two types of records has to be operated following this procedure:

1) open the *ISO3*_geo_cleaned_date.xls file in Excel
2) insert a new column on the right of the id column and called it "d"
3) extract the last digit of the Household ID (id column) by using the formula:
   =RIGHT(x,1) (where x is the corresponding cell in the id column)
   When d= 1 we have a test case, when it is equal to 2 we have a retest
4) Copy the "d" column and paste it on itself using the paste special/values option
4) Using the sort function of the Data menu, separate the test cases (to be saved as: *ISO3*_Test_date.xls) from the retest cases (to be saved as : *ISO3*_Retest_date.xls) in the *ISO3*\3_Geovariable\Center_and_Form folder

It might happen that values in some of the cells are stored as text instead of numbers presenting the use of the sort function. If this would be the case convert the cell content to numbers before processing this operation

2.2. Separation of the records without coordinates

The records without coordinates are not taken in the process. They therefore needs to be removed from the text case file generated in the previous section using the following steps (they will be integrated again in the final sub set, see section 3.4):

1) open the *ISO3*_Test_date.xls file in Excel
2) sort the table using any of the columns that contains part of the lat/long coordinates for the households (e.g. LAT or LONG)
3) Select the records without coordinates, cut and paste them in a new file called *ISO3*_nocoord_date.xls in the *ISO3*\3_Geovariable folder
4) Save the file with the records presenting coordinates in a file called *ISO3*_Test2_date.xls in the *ISO3*\3_Geovariable\Center_and_Form folder

2.3. Individualisation of the clusters (unique ID for each cluster)

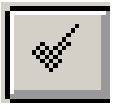
The following steps allows to give a unique ID to each cluster part of the sample numbering them from 1 to X (where X is the total number of surveyed cluster).

1) Open the sampling key correspondence table file (*ISO3*_Samp_key_table_date.xls) in Excel
2) Create a new column (Cluster_ID) which will contain a unique cluster ID based on the coding scheme used by the survey institution. If the coding scheme used already generate unique IDs place them in the new column. If this is not the case generate one by, for example, merging the codes of the different sampling level together
3) Insert a new column next to the "Cluster_ID" one and call if "GI_ID"
4) Sort the table in ascending order for the "Cluster_ID" and fill the "GI-ID" with unique number ID starting from 1 until the last cluster. This code will be used by ArcView for generating the weighted center of gravity and other related geo variables
3) Save the resulting file as "*ISO3*_Samp_key_table2_date.dbf" in the "Sampling" folder
4) Open the *ISO3*_Test2_date.xls file in excel and add a new column called "Cluster_ID".
5) in the new "Cluster ID" column, generate the unique ID according to the model applied in step 2 for the "*ISO3*_Samp_key_table_date.xls" file
6) Save the resulting table as "*ISO3*_Test3_date.dbf" in the *ISO3*\3_Geovariable\Center_and_Form folder

2.4. Generation of the working ArcView project

Some of the geo variables being calculated using scripts and extension running with the ArcView software it is firstly need to generate an ArcView project which does contain them and which will be sued for the coming processes. This is done using the following process:

1) Copy the ArcView project of reference called Ref_WHS.apr from the Annex_GEO_subset.zip file to the *ISO3*\3_Geovariable folder

2) Copy the ArcView extension prjctr.avx, XTOOLSMH.avx and the spider.ave script from the Annex_GEO_subset.zip file to the …:\ESRI\AV_GIS30\ARCVIEW\EXT32 folder located on your hard drive
3) Start the ArcView software
4) Open the Ref_WHS.apr ArcView project and save it as *ISO3*_WHS.apr in the *ISO3*\3_Geovariable folder
5) Change the working directory to correspond to the *ISO3*\3_Geovariable\work folder
6) Activate the prjctr.avx extension in the *ISO3*_working.apr project by checking the "Projector!" box that now appears in the File/Extensions menu
7) Upload the spider.ave scrip in the *ISO3*_working.apr project by:
   - creating a new script window
   - uploading the content of the spider.ave file into this script window clicking on the load text file icon  and specifying the path to this file
   - compiling the script clicking on the "compile" icon  in the menu bar

8) Save the changes done in the *ISO3*_WHS.apr project file

2.5. Generation of the working Shape file

Some of the geo variables being calculated based on ArcView shape files it is necessary to convert the working table into this format using the following steps:

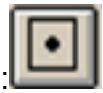
1) Open the *ISO3*_WHS.apr project generated in the previous section
2) Import the "*ISO3*_Samp_key_table2_date.dbf" & "*ISO3*_Test3_date.dbf" tables in the ArcView project using the "Add table" function of the project menu
5) Join the "*ISO3*_Samp_key_table2_date.dbf" table to the "*ISO3*_Test3_date.dbf" one using the "Cluster_ID" column as the common field
6) Make sure that the joint as link value for each of the records part of the *ISO3*_Test3_date.dbf table. If this would not be the case check if this is not linked to an error in the coding structure of the common ID or if this ID is in fact not missing in the "*ISO3*_Samp_key_table2_date.dbf" table
7) If the joint succeed export the resulting table as "*ISO3*_Test4_date.dbf" in the *ISO3*\3_Geovariable\Center_and_Form folder
8) Import the "*ISO3*_Test4_date.dbf" table in the project
9) Open a new view and choose the "Add Event Theme" option of the "View" menu. Select "*ISO3*_Test4_date.dbf" as the table field, "Long" for the X coordinate and "Lat" for the Y coordinate
10) Convert the resulting "*ISO3*_Test4_date.dbf" Event Theme into a shape file saving it as "*ISO3*_Test4_date.shp" in the *ISO3*\3_Geovariable\Center_and_Form folder
11) save the changes done in the *ISO3*_working.apr project file

3. Generation of the GEO Variables

From that point we will be able to calculate and integrate the different GEO variables in the original cleaned version of the geographic component of the WHO WHS

3.1. Calculation of the weighted center of gravity

The following operations allows the calculation of the localization of the surveyed clusters's weighted center of gravity (WCG) and their storage in a shape file:

1) In the view select the *ISO3*_test4_date.shp shape file
2) Display the attribute table, and sort the GI_ID column in the descending order and identify the highest value.
3) From the view click the icon "Points to weighted center of gravity": 
   - In the first window which appears, enter the *ISO3* code
   - of the concerned country
   - In the second window, enter the highest GI_ID value observed in step 2) and start the process

   - This script creates a shape file ISO3_w_cg.shp representing the WCG of each cluster including the following fields in its attribute table:
   - Shape: type of feature
   - ID: source field of the GI_ID
   - GI_ID: Cluster ID
   - Num_pts = Number of points taken into account in the Cluster
   - Y_wc = Latitude of the WCG
   - X_wc = Longitude of the WCG

4) Save the changes done in the project after making sure that the *ISO3*_w_cg.shp is saved in the *ISO3*\3_Geovariable\Center_and_Form folder

3.2. Change of the projection from decimal degrees to metric

Before being able to apply the other methods aiming at generating the other GEO variables it is necessary to change the projection system of the latitude and longitude information we have for the location of the households and the weighted center of gravity of each cluster. This is done using the following steps

3) In the view menu of the *ISO3*_working.apr project, select view properties and indicate "decimal degrees" as the Map Unit
4) Select the *ISO3*_Test4_date.shp shape file and from the view click the "change projection" icon. 
   - In the 1st window precise meters as the output units

- In the 2nd window (Projection properties) let the standard button being check, select UTM 1983 as the Category and choose the Zone for the country in question in the Type field. The corresponding UTM Zone can be found on the shape file that can be downloaded from https://zulu.ssc.nasa.gov/mrsid/
- in the 3rd window (Recalculate area…..etc) select "No"
- in the 4th window (Add as theme to a view) select "Yes"
- in the 5th indicate "New view" and name the file "*ISO3*_test_date_m.shp"

5) Select the "*ISO3*_w_cg.shp" shape file and repeat the process reported in step 4 naming the final file "*ISO3*_w_cg_m.shp" adding it to the new view generated under step 4
6) Join the "*ISO3*_w_cg_m.shp" table to the "*ISO3*_Test_date_m.shp" attribute tables using the "GI_ID" as the common field
7) Make sure that each record in the "*ISO3*_Test_date_m.shp" attribute tables has been attributed a latitude and longitude value for the weighted center of gravity and export the resulting table under the name "*ISO3*_geovar_date_dbf" in the *ISO3*\3_Geovariable\Center_and_Form folder

3.3. Calculation of the distances between each household and the weighted center of gravity for each the cluster

This part of the process allows the calculation the distance between each HH and its corresponding cluster weighted center of gravity using the spider.ave script.

This script creates spider diagrams based on liner distances between the points contained in two points themes (1 for centers & the other for points). The output is a new theme with corresponding distances from each point of the cluster to the center. This output is obtained as follow:

1) Make sure that the *ISO3*_w_cg_m.shp and the *ISO3*_Test_date_m.shp layer are in the view but not activated
2) Open the spider script window
3) Run the script by clicking on the run icon  and by précising:
   - in the 1st window: *ISO3*_w_cg_m.shp as the theme to use for centers
   - in the 2nd window: *ISO3*_Test_date_m.shp as the theme to use for points
4) Save the resulting file as *ISO3*_distances.shp in the *ISO3*\3_Geovariable\Center_and_Form folder
5) Save the changes done to the project and close Arc view

3.4. Calculate Skewness and Kurtosis

These two indexes will allow to get an idea of the cluster extend/shape based on the distance between the HH and its corresponding cluster's center of gravity as it is not possible to redistribute the exact location of each household for confidentiliaty reasons. The calculation of these two indexes is done using the following steps:

1) In Excel, open the *ISO3*_geovar_date.dbf (see section 3.2) and the *ISO3*_distances.dbf files
2) Copy the DISTANCE column from *ISO3*_distances.dbf and paste it in the *ISO3*_geovar_date.dbf file (if no sorting has been done since the generation of these two files the order of the records is the same in both of them)
3) Copy the GI_ID column from *ISO3*_geovar_date.dbf and paste it in *ISO3*_distances.dbf
4) Make sure that all the records have been attributed new values with the operation reported under point 2 and 3 and save the changes in *ISO3*_distances.dbf and close the file
5) In *ISO3*_geovar_date.dbf add a column called "SKEW" and calculate the skewness for the first cluster applying the excel SKEW function to all the distance reported for it in the DISTANCE column. Manually do the same for all the other clusters.
6) Add a column called "KURTOSIS" and calculate the kurtosis for the first cluster applying the excel KURT function to all the distance reported for it in the DISTANCE column. Manually do the same for all the other clusters.
7) Save the changes in the *ISO3*_geovar_date.dbf file and close it

3.5. Creation of the GEO variable file

Before going to the next step of the process driving to the generation of the geo subset it is important to generate an intermediate file which will contain all the variables calculated until now using this protocol and to add the records without coordinates to the file in question (see section 2.2) using the following process:

1) Add the *ISO3*_geovar_date.dbf table to the *ISO3*_WHS.apr ArcView project
2) Sum the DISTANCE, SKEW and KURTOSIS field by Cluster_ID using the sum icon  (after having selected the Cluster_ID or the GI_ID column)
3) Save the summary definition table as *ISO3*_geovar_date_sum.dbf
4) join the *ISO3*_geovar_date_sum.dbf table to the *ISO3*_geovar_date.dbf one using Cluster_ID or GI_ID as the common field
5) Export the resulting table as *ISO3*_geovar_final_date.dbf
6) Save the project and close Arc view
7) Open *ISO3*_geovar_final_date.dbf and *ISO3*_nocoord_date.xls (see section 2.2) in Excel
8) In the *ISO3*_nocoord_date.xls file, add one column called Cluster_ID and generate the unique ID according to the model applied in step 2 of Section 2.3
9) Add the records from *ISO3*_nocoord_date.xls at the bottom of the table in the *ISO3*_geovar_final_date.dbf file making sure that the correspondence in terms of columns is respected between both sets
10) Make sure that the file contains the following columns with the corresponding format and content:

- ID	Number	WHS unique identifier
- COUNTRY	Text	ISO3 country code
- Q0100	Number	Primary sampling unit (PSU) code
- Q0101	Number	Secondary sampling unit (SSU) code
- Q0102	Number	Tertiary sampling unit (TSU) code

- Q0103	Number	Quaternary sampling unit (QSU) code
- **CLUSTER_ID**	Number	Unique ID generate for each cluster
- Q0104	Number	Setting code: Urban versus rural designation
- Q0105s	Text	Setting specification
- Q0200_1	Number	Latitude hemisphere code: N - north, S - south
- Q0200_2	Number	Latitude degree
- Q0200_3	Number	Latitude decimal degree
- **LATNUM**	Number	Latitude coordinate in decimal degrees of the household
- Q0201_1	Number	Longitude hemisphere code: E - north, W - south
- Q0201_2	Number	Longitude degree
- Q0201_3	Number	Longitude decimal degree
- **LONGNUM**	Number	Longitude coordinate in decimal degrees of the household
- **DATUM**	Text	Datum of raw coordinates
- Q0202	Number	Waypoint code (location of the GPS measurement: 2 - In front of the household, 3 - nearby location
- **COUNT_GPS**	Number	Number of household in the cluster which have been taken into account for the calculation of the geovariables (Num_pts field)
- **LAT_WC**	Number	Latitude coordinate in decimal degrees of the weighted center of gravity of the cluster
- **LONG_WC**	Number	Longitude coordinate in decimal degrees of the weighted center of gravity of the cluster
- **DIS_WC**	Number	Distance in meter between the weighted center and the Household (DISTANCE field)
- **MIN_DIS_WC**	Number	Minimum distance observed between the weighted center of gravity and all the households part of the cluster
- **MAX_DIS_WC**	Number	Maximal distance observed between the weighted center of gravity and all the households part of the cluster
- **MEAN_DIS_WC**	Number	Mean distance observed between the weighted center of gravity and all the households part of the cluster
- **STDEV_DIS_WC**	Number	Standard deviation of the distances between the weighted center of gravity and the households compare to the mean distance
- **SKEWNESS**	Number	Index characterizing the degree of asymmetry of the distribution of the households around the weighted center of gravity
- **KURTOSIS**	Number	Index characterizing the relative peakedness or flatness of the distribution of the households around the weighted center of gravity compared to a normal distribution (Gauss)

The variables reported in bold in this list correspond to the ones that have been added to the original file used at the beginning of this process (see section 2.1)

11) If needed, delete the unnecessary columns, correct the spelling of the headers and arrange the order of the columns to correspond to the one in the list reported under point 8)
12) Sort the whole table by cluster (CLUSTER_ID column) and manually complete the following fields for the records with no GPS coordinates:
    - CLUSTER ID
    - COUNT_GPS
    - LAT_WC
    - LONG_WC
    - MIN_DIS_WC
    - MAX_DIS_WC
    - MEAN_DIS_WC
    - STDEV_DIS_WC
    - SKEWNESS
    - KURTOSIS

This is not applied to the DIST_WC column as it has not been possible to calculate a distance for these records

13) Add a new column called COUNT_ALL in which will be stored the total number of household observed in each cluster. To fill this column:
    - insert a new worksheet in the excel file
    - in this new sheet create a pivot table for which the database extend will be the column containing the unique cluster ID
    - once the pivot table generated drag the cluster ID field from the pivot table field list to the center of the table (place indicated by "Drop data Item Here"). This will have for result to calculate the number of records for each cluster.
    - populate the COUNT_ALL column on the first sheet by using the VLOOKUP function of excel using the CLUSTER_ID as the link between the first and the second worksheet
    - copy the content of the COUNT_ALL column and paste it in the same place using the paste special/value function.
14) Save the final file as *ISO3*_geovar_final_date.xls in the R:\WHO_WHS\Survey_2002\DATA_CLEANING\Countries\ISO3\3_Geovariables\Center_and_Form

4. Integration of the sampling level labels, SALB data set and final setting information in the GEO variables

In order for the user to more easily link the sampling level codes with a geographic object (Administrative unit, populated place,..) the labels provided by the survey institutions for each of the sampling level as well as the Second Administrative Level Boundaries data set (SALB) names and codes have been added to the GEO variables calculated in the previous sections. This work is done in two steps:

- integration of the SALB names and codes in the key correspondence table provided by the survey institution
- link between the new version of the key correspondence table and the *ISO3*_geovar_final_date.xls file

4.1. Integration of the SALB data in the key correspondence table

The first step of this process consist in homogenising the content of the key correspondence table received from the survey institution. This is done as follow:

1) In excel, open the *ISO3*_Samp_key_table2_date.dbf file (see section 2.3) and organize the structure of the file in order to only have the following fields:
    - Cluster_ID or GI_ID (number): WHS Cluster unique ID
    - Q0100_code (number): PSU code used by the survey institution
    - Q0100_label (text): PSU label provided by the survey institution (if available)
    - Q0101_code (number): SSU code used by the survey institution (if applicable)

- Q0101_label (text): SSU label provided by the survey institution (if applicable and available)
- Q0102_code (number): TSU code used by the survey institution (if applicable)
- Q0102_label (text): TSU label provided by the survey institution (if applicable and available)

- Q0103_code (number): QSU code used by the survey institution (if applicable)
- Q0103_label (text): QSU label provided by the survey institution (if applicable and available)
- Admin1_survey: first-order administrative division name provided by the survey institution
- Admin2_survey: second-order administrative division name provided by the survey institution

2) Call this first worksheet "Labels" and save the resulting file as *ISO3*_Samp_key_work_date.xls in the *ISO3*\3_Geovariables\Labels_SALB folder

From there, depending on the type of information at disposal two type of process can be applied in order to integrate the SALB administrative units names and codes into the key correspondence table provided by the survey institution:

A) It has been possible for the survey institution to provide the complete list of 1st and 2nd level administrative units names (or to recreate this list based on information coming from the sampling plan) used for defining the sampling frame used for the survey as well as the link between the 2nd level admin unit names and any of the sampling level (PSU, SSU, TSU). If this is the case follow the steps going from A1) to A13) bellow.

B) One of the two information reported in point A) is missing. If this is the case follow the steps going from B1) to B11) bellow.

Process when all the information is available

A1) In Excel, open the *ISO3*_Samp_key_work_date.xls file generated under point 1) above and name the worksheet in which key sampling information is located "Source". Insert 2 new worksheet called "Temporary" and "Comparison". Save the file as *ISO3*_Samp_key_work2_date.xls

A2) Having the complete list of 1st and 2nd level administrative units used by the survey institution in hands look at the historic changes data set posted on the SALB web site (http://www3.who.int/whosis/gis/salb/salb_coding.htm) and try to identify the match between the different period of representativity reported there and the list of admin units reported in the complete list provided by the survey institution. If this link is not possible, for example because the historic changes are not complete, take contact with the SALB project coordination team in order to obtain this information (contact information on the web site itself).

A3) In *ISO3*_Samp_key_work2_date.xls copy the list of 1st and 2nd level administrative units names from the Source worksheet into the Temporary one

A4) Sort the whole table in the ascending order firstly by 1st administrative level units name and then by ascending 2nd administrative level unit names.

A5) As this table will still contain several time the same information, use a pivot table in order to create one which will contain each unit only once. This is done using the following process:

- select the two columns containing the 1st and 2nd level admin unit name provided by the institution and start the "PivotTable and PivotChart" option of the Data menu
- place the pivot table in the same worksheet
- from the "Pivot Table Filed list" drag the Admin2_survey field into the "Drop Row Field here" section of the pivot table and do the same for the Admin1_survey one. This will create a table which contain the link between each 2nd level admin unit with the corresponding one
- drag the Admin2_survey field into the "Drop Data Item" section of the pivot table which will provide the number of cluster observed in each 2nd level admin unit
- copy the content of the pivot table and paste it in the Comparison one using the paste special/value option
- complete the table in the Comparison worksheet in order to have the name of the corresponding 1st level admin unit in front of each 2nd level units and delete the "total" lines generated by the pivot table process

A6) Open the *ISO3*_SALB_table.xls downloaded from the SALB web site or obtained from the SALB coordination team. Copy the 1st and 2nd administrative level names and codes corresponding to the representativity identified under step A2) and paste it into the Comparison worksheet of the *ISO3*_Samp_key_table3_date.xls file. This table is already sorted by ascending order of the 1st and 2nd administrative level unit names

A7) In the Comparison worksheet, without touching the order or spelling of the column provided by the survey institution try to match the administrative divisions names for the 1st and 2nd admin level coming from SALB with the ones provided by the survey institution keeping the corresponding SALB codes attached to them. The following step could help in this regards:

- Insert 2 news columns on the right in the Comparison worksheet and name them "Adm1_comp" and "Adm2_comp"
- In the first record of the "Adm1_comp" column: insert the following formula:

  IF(cell (AX)=cell (BX),1,0)

  With A being the SALB 1st administrative unit, B the 1st administrative unit reported by the survey institution and X the row of the cell in question. Paste this first cell down the all list of admin units.
- Apply the same process for the "Adm2_comp" column with this time A being the SALB 2nd administrative unit, B the 2nd administrative unit reported by the Survey and X the row of the cell in question. Values in the Adm1_comp or Adm2_comp column is equal to for which the result of the application of the formula is equal to 0. These cases correspond to situation where there is a difference between the information reported in SALB and the one provided by the survey institution. This might be because of some spelling errors, incomplete name or a real difference between the two lists.
- Identify the first difference in the list starting from the top. If this is the case, correct the information from the survey institution as the one coming from SALB has been validated by the country. In some cases

If the number of 2nd level administrative units reported by the survey institution is higher than the one reported in the SALB data set (e.g more detailed subdivision of the urban

areas…), it is possible to use the corresponding SALB map in combination with a map showing the delimitation of the units in question in order to identify to which 2nd administrative unit reported in SALB they should be linked to.

A8) Still in the Comparison worksheet, once the link between the SALB names/codes and the survey information has been established, create a new column called "Adm1_adm2_WHS". Fill this column with the result of the merging between the 1st and 2nd level administrative units names provided by the survey institution using the Excel "concatenate" function.

A9) Sort the table by ascending order of the "Adm1_Adm2_WHS" column"

A10) In the Source worksheet, repeat the operations reported in points A8) and. add new 4 columns as follow:

- ADM1SALBCODE Text SALB first-order administrative division code
- ADM1SALBNAME Text SALB first-order administrative division name
- ADM2SALBCODE Text SALB second-order administrative division code
- ADM2SALBNAME Text SALB second-order administrative division code

A11). Use the VLOOKUP function of excel to populate the columns generated under step A10) with the information from the "Comparison" worksheet using the Adm1_Adm2_WHS column as the link between the two worksheets. Make sure that new figures are attributed to each of the record.

A12) Copy the content of the newly populated columns and paste them on themselves using the paste special/value function.

A13) Delete the other worksheet to keep only the source one and save the resulting file as *ISO3*_samp_key_SALB_date.xls in the *ISO3*\3_Geovariables\Labels_SALB folder.

Process when some information are missing

B1) Download the January 2000 shape file for the country in question from the SALB web site (http://www3.who.int/whosis/gis/salb/salb_MDATA.htm). If this map is not available liaise with the SALB coordination team to know if an other map which would correspond as one of the representativity part of the SALB historic changes could be used. If this is the case, make sure that the SALB codes are integrated in the attribute table of this map before goinf further in the process.

B2) Open the *ISO3*_WHS.apr project in ArcView and add the SALB (or other source) administrative boundary map, in a new view

B3) If necessary, change the projection of the map from decimal degrees to metric using the steps reported in section 3.2 saving the file as *ISO3*_SALB_m.shp in the in the *ISO3*\3_Geovariable\Labels_SALB folder

B4) Copy the *ISO3*_SALB_m.shp file in the same view than the *ISO3*_test_date_m.shp, the *ISO3*_w_cg_m.shp (see section 3.2) and the *ISO3*_distance.shp layers (see section 3.3)

B5) Overlaying the points layers on top of the identify in which second administrative level unit each cluster is located, this can be done:

- automatically if you do have access to the Spatial analyst ArcView extension by converting the 2nd level administrative boundaries shape file into a grid which would contain the SALB codes as the attribute and by then making a summaries by zone on it using the *ISO3*_w_cg_m.shp as the reference layer and the

cluster_ID or the GI_ID as the field which defines the zones. Export the resulting table as Link_cluster_SALB.dbf

- manually if the Spatial Analyst extension is not available, selecting all the cluster located in a same administrative units and by entering manually the SALB code in a new column , called "2admin_code" in the attribute table of the *ISO3*_w_cg_m.shp file. Saving the resulting table as Link_cluster_SALB.dbf.

B6) In both of the case mentioned in step B5 a special care will have to be given to cluster located at the border between two administrative units. This check has to be visually done and additional information maybe found in order to obtain a confirmation regarding the administrative units in which these particular clusters falls (for example using the name of the village or town in which they are located as collected in section 0300 of the WHS questionnaire)

B7) In Excel, open the Link_cluster_SALB.dbf file and the *ISO3*_Samp_key_work_date.xls files

B8) Create a new worksheet in the *ISO3*_Samp_key_work_date.xls file, call it "SALB" and copy the content of the Link_cluster_SALB.dbf file into it

B9) In the "Label" worksheet add 4 new columns as follow:

- ADM1SALBCODE (text): SALB first-order administrative division code
- ADM1SALBNAME (text): SALB first-order administrative division name
- ADM2SALBCODE (text): SALB second-order administrative division code
- ADM2SALBNAME (text):SALB second-order administrative division name

B10) Use the VLOOKUP function of excel to populate the columns generated under step B9) with the information in the "SLAB" worksheet" using the Cluster_ID or the GI_ID column as the link between the two worksheet. Make sure that new figures are attributed to each of the record.

B11) Make sure that SALB names and codes have been attributed to each cluster. If this is the case, copy the content of the 4 columns and paste them on themselves using the Paste special/values option. Delete the other worksheet to keep only the source one and save the resulting file as *ISO3*_Label_SALB_date.xls in the *ISO3*\3_Geovariables\Labels_SALB folder

4.2. link between the new version of the key correspondence table and the GEO variable file

The integration of the information reported in the new key correspondence table generated in the previous section is done using the following steps:

1) In excel, open the *ISO3*_geovar_final_date.xls file (see section 3.4) and add the following columns:

- Q0100_label	Text	Primary sampling unit (PSU) label (indication of the type of unit)
- Q0101_label	Text	Secondary sampling unit (SSU) label (indication of the type of
- Q0102_label	Text	Tertiary sampling unit (TSU) label (indication of the type of unit)
- Q0103_label	Text	Quaternary sampling unit (QSU) label (indication of the type of unit)

- ADM1SALBCODE Text SALB first-order administrative division code
- ADM1SALBNAME Text SALB first-order administrative division name
- ADM2SALBCODE Text SALB second-order administrative division code
- ADM2SALBNAME Text SALB second-order administrative division name

2) Open the *ISO3*_samp_key_SALB_date.xls, copy its content and paste it in a new worksheet in the *ISO3*_geovar_final_date.xls file
3) Use the VLOOKUP function of excel to populate the columns generated under step 1 using the unique cluster ID as the link between the two worksheets. Make sure that new figures are attributed to each of the record.
4) Copy the content of the newly populated columns and paste it in the same place using the paste special/value function.
5) Delete the other worksheet to keep only the one which contain the complete table and save the resulting file as *ISO3*_Label_SALB_date.xls in the *ISO3*\3_Geovariables\Labels_SALB folder

4.3. Integration of the cleaned setting information

In order to homogenise the content of the q0104 question (setting) a parallel process has produced new figures for all the countries based on a two categories:

- urban (codified as 1)
- rural (codified as 3)

It is therefore necessary to replace the figures initially reported in this field in the *ISO3*_Label_SALB_date.xls using the following steps:

1) from the people in charge of the complete WHS database obtain an excel table which contains two fields:
   - the record id (id)
   - the new setting information (q0104)
2) In excel, open the *ISO3*_Label_SALB_date.xls file and add a new worksheet
3) Copy the content of the file provided under point 1) in the newly created worksheet
4) Use the VLOOKUP function of excel to replace the figures in the q0104 column of the original *ISO3*_Label_SALB_date.xls with the one located in the new worksheet using the record ID as the link between the two worksheets. Make sure that a new figure is attributed to each of the record.
4) Copy the content of the q0104column and paste it in the same place using the paste special/value function.
5) Delete the other worksheet to keep only the one which contain the complete table and save the resulting file as *ISO3*_setting_date.xls in the *ISO3*\3_Geovariables\Setting folder

5. Finalization of the GEO subset file

In order to homogenize the content and make a final check of the file generated using the present protocol the last following steps are applied:

1) In excel, open the *ISO3*_setting_date.xls
2) Make sure that all the column are named and organized in the order reported in Annex 1. Make the necessary modification if this would not be the case
3) Make sure that the format of the figures reported in each column also correspond to the ones reported in Annex 1
4) Make a final visual check of the content of the table making sure that:
    - the total number of records which should correspond to the number of test cases that were observed in section 2.1
    - all the records contains figures for the fields that should necessary contain information (e.g. COUNT_ALL, KURTOSIS,…)
5) Save the final file as *ISO3*_geosubset_final_date.xls in the *ISO3*\4_Final folder

6. Generation of the metadata file

In order to document the content of the GEO subset file a specific Metadata profile has been generated for the context of the WHO WHS.

This profile is filled for each country using the information provided by the survey institution as well as the ones generated during the data cleaning process and the application of the present protocol. The result is a Metadata record such as the reported in Annex 2 for Malaysia.

Annex 1 - List of potential fields in the final GEO subset

Field Name	Type	Description
ID	Number	WHS unique identifier
COUNTRY	Text	ISO3 country code
Q0100	Number	Primary sampling unit (PSU) code
Q0100_label	Text	Primary sampling unit (PSU) label (*indication of the type of unit*)
Q0101	Number	Secondary sampling unit (SSU) code
Q0101_label	Text	Secondary sampling unit (SSU) label (*indication of the type of unit*)
Q0102	Number	Tertiary sampling unit (TSU) code
Q0102_label	Text	Tertiary sampling unit (TSU) label (*indication of the type of unit*)
Q0103	Number	Quaternary sampling unit (QSU) code
Q0103_label	Text	Quaternary sampling unit (QSU) label (*indication of the type of unit*)
CLUSTER_ID	Number	Unique ID generate for each cluster
ADM1SALBCODE	Text	SALB first-order administrative division code
ADM1SALBNAME	Text	SALB first-order administrative division name
ADM2SALBCODE	Text	SALB second-order administrative division code
ADM2SALBNAME	Text	SALB second-order administrative division name
Q0104	Number	Setting code: Urban versus rural designation
Q0105s	Text	Setting specification
Q0200_1	Number	Latitude hemisphere code: N - north, S - south
Q0200_2	Number	Latitude degree
Q0200_3	Number	Latitude decimal degree
LATNUM	Number	Latitude coordinate in decimal degrees of the household
Q0201_1	Number	Longitude hemisphere code: E - north, W - south
Q0201_2	Number	Longitude degree
Q0201_3	Number	Longitude decimal degree
LONGNUM	Number	Longitude coordinate in decimal degrees of the household
DATUM	Text	Datum of raw coordinates
Q0202	Number	Waypoint code (location of the GPS measurement: 2 - In front of the household, 3 - nearby location
COUNT_ALL	Number	Total number of household in the cluster
COUNT_GPS	Number	Number of household in the cluster which have been taken into account for the calculation of the geovariables (Num_pts field)
LAT_WC	Number	Latitude coordinate in decimal degrees of the weighted center of gravity of the cluster
LONG_WC	Number	Longitude coordinate in decimal degrees of the weighted center of gravity of the cluster
DIS_WC	Number	Distance in meter between the weighted center and the Household
MIN_DIS_WC	Number	Minimum distance observed between the weighted center of gravity and all the households part of the cluster
MAX_DIS_WC	Number	Maximal distance observed between the weighted center of gravity and all the households part of the cluster
MEAN_DIS_WC	Number	Mean distance observed between the weighted center of gravity and all the households part of the cluster
STDEV_DIS_WC	Number	Standard deviation of the distances between the weighted center of gravity and the households compare to the mean distance
SKEWNESS	Number	Index characterizing the degree of asymmetry of the distribution of the households around the weighted center of gravity
KURTOSIS	Number	Index characterizing the relative peakedness or flatness of the distribution of the households around the weighted center of gravity compared to a normal distribution (Gauss)

Annex 2 - Example of Metadata record for Malaysia

Dataset Title	World Health Survey's Geographic Subset of Malaysia
Geographic Location	Malaysia
Geographic Box	**X min:** E 98.5 **X max:** E 119.5 **Y min:** N 0.5 **Y max:** N 7.5
Year	2003
Collection Start date	2003-03-01
Collection End Date	2003-04-30
Implementing Organization	Public Health Institute, Ministry of Health
Status	Completed
Number of records	Total 7528 (Test cases: 6040, Missing cases 1488)
Format	Excel format: .xls
Filename	MYS_geosubset_final_30_12_04.xls
Abstract:	This dataset contains the geographic component of the WHO WHS performed in Malaysia. The following information and variables can be found in this file: - the cleaned information stored in the section 0100 and 0200 of the questionnaire - the labels attached to the codes used in the data set for identifying each level of the sampling - the 1st and 2nd administrative units level names and codes coming from the Second Administrative Level Boundaries data set project - the weighted centre of gravity of each surveyed cluster. - different parameters and indexes offering an indication of the dispersion of the households interviewed around the cluster's center of gravity.
Supplemental Information:	The following variables can be found in the excel file:

Field Name	Type	Description
id	Number	WHS unique identifier
country	Text	ISO3 country code
Q0100	Number	Primary sampling unit (PSU) code
Q0101	Number	Secondary sampling unit (SSU) code
ADM1SALBCODE	Text	SALB first-order administrative division code
ADM1SALBNAME	Text	SALB first-order administrative division name
ADM2SALBCODE	Text	SALB second-order administrative division code
ADM2SALBNAME	Text	SALB second-order administrative division name
Q0104	Number	Setting code: Urban versus rural designation
Q0200_1	Number	Latitude hemisphere code: N - north, S - south
Q0200_2	Number	Latitude degree
Q0200_3	Number	Latitude decimal degree
LATNUM	Number	Latitude coordinate in decimal degrees of the household
Q0201_1	Number	Longitude hemisphere code: E - north, W - south
Q0201_2	Number	Longitude degree
Q0201_3	Number	Longitude decimal degree
LONGNUM	Number	Longitude coordinate in decimal degrees of the household
DATUM	Number	Datum of the coordinates
Q0202	Number	Waypoint code (location of the GPS measurement)
COUNT_ALL	Number	Total number of household in the cluster
COUNT_GPS	Number	Number of household in the cluster for which a GPS coordinate is available (used for the calculation of the geovariables)
LAT_WC	Number	Latitude coordinate in decimal degrees of the weighted center of gravity of the cluster
LONG_WC	Number	Longitude coordinate in decimal degrees of the weighted center of gravity of the cluster
DIS_WC	Number	Distance in meter between the weighted center and the Household
MIN_DIS_WC	Number	Minimum distance observed between the

		weighted center of gravity and all the households part of the cluster
MAX_DIS_WC	Number	Maximal distance observed between the weighted center of gravity and all the households part of the cluster
MEAN_DIS_WC	Number	Mean distance observed between the weighted center of gravity and all the households part of the cluster
STDEV_DIS_WC	Number	Standard deviation of the distances between the weighted center of gravity and the households compare to the mean distance
SKEWNESS	Number	Index characterizing the degree of asymmetry of the distribution of the households around the weighted center of gravity
KURTOSIS	Number	Index characterizing the relative peakedness or flatness of the distribution of the households around the weighted center of gravity compared to a normal distribution (Gauss)

The variables mentioned in bold correspond to information not collected during the survey itself but collected or calculated separately afterwards.

If the information has not been collected in the field or it has not been possible to calculate it the corresponding geo variable appears as an empty cell in the file.

The sampling codes are linked to the sampling frame designed by the Implementing Organization and approved by WHO, they allow identifying at which cluster the geocoding information belongs to.

In the case of Malaysia the primary sampling unit (PSU) code was made of an aggregation of units as follow (a total of 10 digits):

1. State (first 2 digits) => refer to the ADMIN1SALBNAME for the corresponding label
2. Administrative district (next 2 digits) => refer to the ADMIN2SALBNAME for the corresponding label
3. Census district (next 2 digits) => no label available
4. Enumeration block number (next 3 digits) => no label available
5. Strata (urban, rural and so forth) (next 1 digit) => 1: Metropolitan Areas (75,000 and above) ; 2: Urban Large (10,000 to 9.999); 3: Urban Small (1.000 to 9.999); 4 to 9: Rural (4: areas with population of less than 1,000, 5: Journey less than 30 minutes, 6: Journey between 30 - < 2 hours, 7: Journey between 2 - < 3 hours, 8: Journey between 3 - < 4 hours, 9: Journey between 4 - 8 hours, 0: Journey more than 8 hours).

The Primary Sampling units correspond to the 399 clusters that have been surveyed

The Secondary Sampling Unit (SSU) code corresponds to the 4-digits code for living quarters in each enumeration block.

For both sampling level no label are reported in the data set because of their respective nature.

The SALB (second Administrative Level Boundaries) codes and names are coming from the project database (see web site at: http://www3.who.int/whosis/gis/salb/salb_home.htm).

The setting reported in the Q0104 field correspond to the aggregated form of the information reported in the Strata. From 9 categories the final data set contains only two: 1: urban and 3: rural

The location of the households (latitude and longitude) taken with a GPS device is reported in decimal degrees, Datum WGS 84.

Waypoint: The indication on the precise place where the interviewer was located while taking the GPS measure is reported in the waypoint variable. The possibilities offered to the interviewers were to take the measurement either in front of the household

	(corresponding to a location very close to the household as the front door or on the roof if possible) or at a nearby location (corresponding to a close location offering an open view of the sky in case it was not possible to obtain an accurate reading "In the front of the household"). The corresponding codes are reported in the file in the following way: 2 - In front of the household, 3 - nearby location. Geovariables: The household in the cluster which have been taken into account for the calculation of the geovariables correspond to the households for which a GPS coordinate has been collected during the survey. Skewness Index: This index characterizes the degree of asymmetry of the distribution of the households around the weighted center of gravity. A positive Skew indicates a distribution with an asymmetric tail extending towards more positive values. A negative Skew indicates a distribution with an asymmetric tail extending towards more negative values. Kurtosis Index: This index characterizes the relative peakedness or flatness of the distribution of the households around the weighted center of gravity compared to a normal distribution (Gauss). A positive Kurt indicates a relatively peaked distribution. A negative Kurt indicates a relatively flat distribution. These two indexes can be used to explain possible heterogeneity in the answers obtained for a same cluster to questions presenting a geography dimension (accessibility for example).
Lineage:	The data collection and cleaning process for the section 0100 and 0200 of the questionnaire as well as the protocols allowing the generation of the geo sub set file are described in documents that can be directly downloaded from the WHO WHS web site at: http://www3.who.int/whs/ (instruments and related documents section). Here are the details of the documents that can be found on the WHO WHS web site: 1. Training material which has been sent to the survey institution for the data collection process. This material groups the following documents: - GPS field Guide which contains an introduction on the GPS system as well as a basic users manual for the Garmin eTrex GPS device. - A PowerPoint presentation to be used during the training - "How to use the GPS training material" document which describes how to use the training material when training the field interviewers 2. Data collection material as follow: - "GPS data collection protocol" which describes the steps to follow in order to fill the section 0100 (Sampling Information) and 0200 (Geocoding Information) of the questionnaire - "Test and use of the GPS in the field" documents which describe how to setup and use the Garmin eTrex device in the context of the WHS. This documents also describes the processes reported in the GPS data collection protocol under the form of a small cartoon 3. Data cleaning protocol which is described in the documents entitled: " Cleaning protocol for the geographic component (section 0100 and 0200 of the questionnaire) for countries using GPS devices". This protocol has been applied in close collaboration with the survey institution in order to fill the gaps or make the necessary corrections. 4. Protocol used for the generation of the geo subset file which is described in the document entitled: "Generation of the geographic subset for countries using the GPS devices" The latitude and longitude coordinates of each households have been collected by the field supervisor using Garmin eTrex GPS devices which have been sent by WHO to the survey institution. Confidentiality: In order to insure the confidentiality of the respondents, the values stored in the section 0200 have always been treated separately from the rest of the data set once received from the field. This section forming what is called the "geographic sub set" of the WHO WHS.

Data Quality Comments	This section describes the degree of confidence that can be applied to each field in the section 0100 and 0200 as well as to the different geovariables that have been integrated or calculated after the survey itself. Sampling: For the sampling Information (section 0100) the key correspondence table provided by the survey institution, the small number of missing information in the initial data set, the application of the automatic protocol generated for the context of the survey as well as the help provided by the survey institution provides us with a good degree of confidence regarding the quality of this information. Setting: The setting information entered in the final data set has been generated using a and automatic protocol. The results obtained have been approved by the survey Institution providing therefore a good degree of confidence for this variable. SALB administrative units: The integration of the SALB administrative units names and codes, information validated by the National Mapping Agency of Malaysia, has been based on a complete list of the units considered for the design of the sampling frame and this information is of good quality. Coordinates: The fact that the GPS devices have been used for locating each of the households part of the survey and the application of a specific data cleaning protocol are insuring the quality of the latitude and longitude reading reported in the data set. It is nevertheless important to mention that the information regarding the latitude and longitude is not available for 1383 Households considered as missing cases (18 % of all the Surveyed households). This high number of missing information is due to the fact that the person using the GPS devices in the field did not understand that the location of the households should also be collected for missing cases. Geovariables: For the geovariables that are derived from the GPS coordinates the quality of the figure reported depends on the number of household in the cluster for which a coordinate was available (see COUNT_GPS field in the data set). This number varies from 4 to 24 with a mean value of 15.4 observation by cluster.
Theme Keywords	Malaysia, World Health Organization World Health Survey (WHO WHS), sampling, Household GPS coordinates, Georeferenced data, administrative units, SALB
Dataset Topic Category	Household Survey Geographic component
Restrictions	For confidentiality reasons the fields containing the latitude and longitude of each households (Q0200_1, Q0200_2, Q0200_3, LATNUM, Q0201_1, Q0201_2, Q0201_3 and LONGNUM) can not be realized to the public. This information remains the property of the country and the Ministry of Health should therefore be contacted if there would be a need to have access to this particular section of the sub set. The other fields mentioned in the "Supplemental Information" section of this document are part of the data set that can be access to the public. Please mention the following copyright and acknowledgement mention in case of use of any of this information: Copyright: ? Acknowledgement: Public Health Institute, Ministry of Health, Malaisa and World Health Organization (WHO). *WHO World Health Survey 2003*, Kuala Lumpur
Linkage	http://www3.who.int/whs/
Dataset Language	En
Dataset Character Set	usAscii
Metadata Provider	World Health Organization
Metadata Contact	EIP/KMS/EHL/STK World Health Organization 20, AV. Appia 1211 Geneva 27 Switzerland Phone: +41.22.791.47.44 Fax: +41.22.791.43.28
Metadata Date	20041117
Metadata Language	En
Metadata Character Set	usAscii
Metadata Standard	ISO 19115