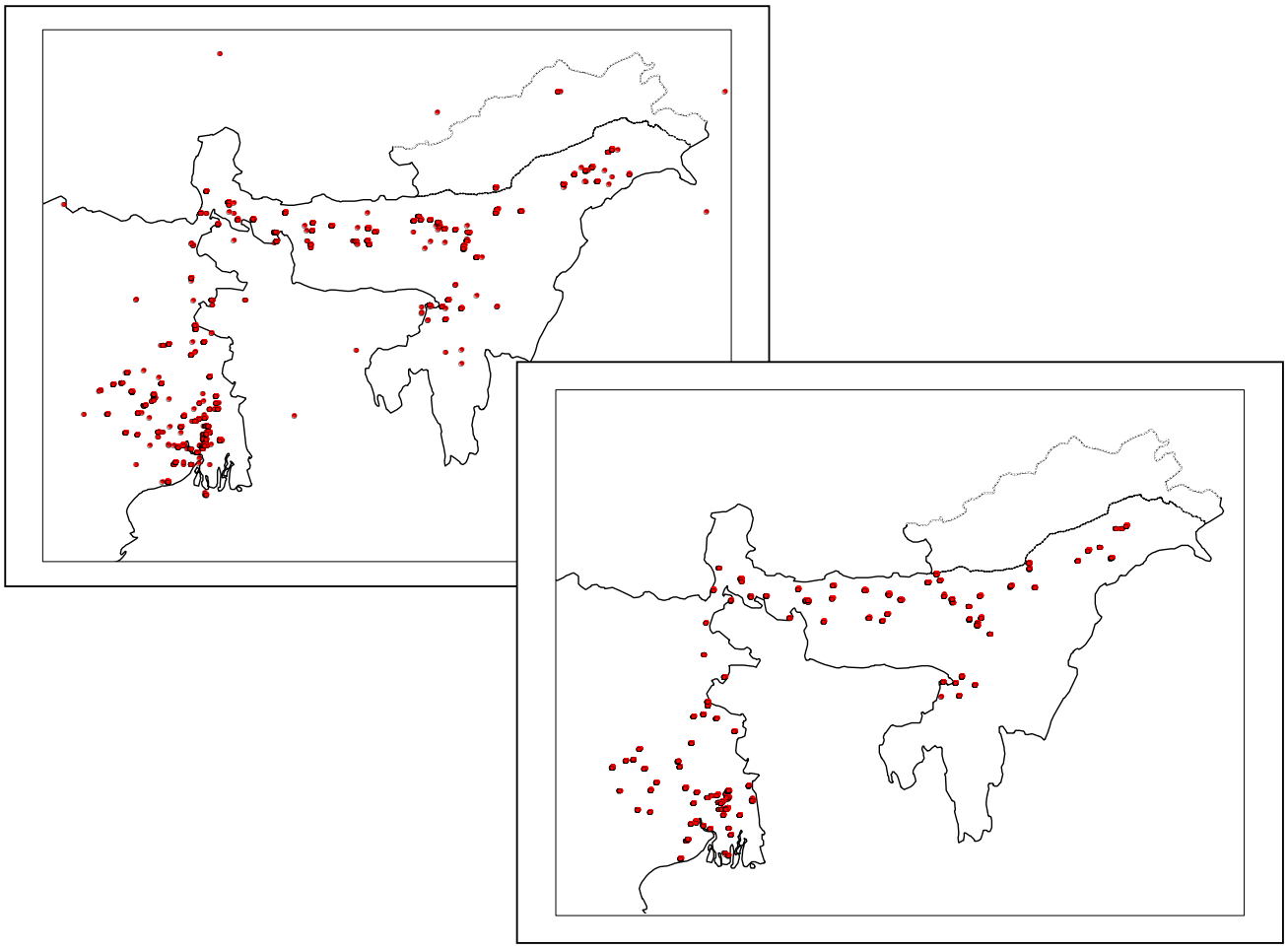
WHS
World Health Survey

Cleaning Protocol for the Geographic Component (Section 0100 and 0200) Countries using GPS devices



Cleaning Protocol for the Geographic Component (Section 0100 and 0200) Countries using GPS devices

Geneva, Switzerland

This document was prepared by F. Naville and S. Ebener

TABLE OF CONTENTS:

List of Figures:

List of Tables:

1. Introduction

This document contains the protocol that has been implemented in the context of the World Health Organization World Health Survey (WHO WHS) in order to clean section 0100 and 0200 of the questionnaire for the countries which used GPS (Global positioning System) and therefore insure the consistency and homogeneity of the geographic component of the datasets.

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

The final file resulting from the application of this protocol contains the corrected and completed variables of the Sampling of Geocoding sections for the test and missing cases.

1.1. The 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

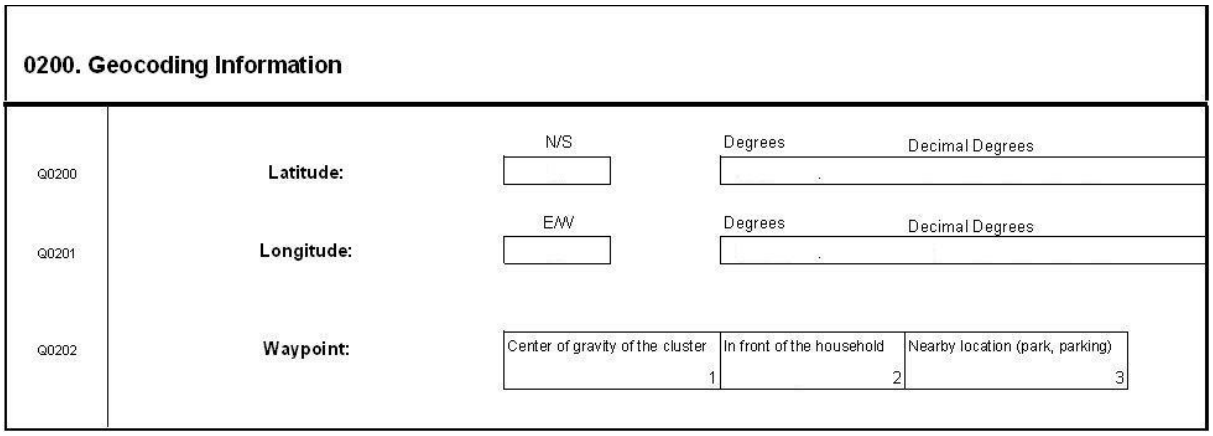


Figure 2 - Geocoding Information section of the WHO WHS questionnaire

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. Cleaning of the WHS Geographic Component

The cleaning of the WHS geographic component of the WHO WHS is based on the data received from the field and collected in the sections 0100 and 0200 of the WHS questionnaire for the countries using GPS devices.

The whole process is applied on a country by country basis and structured in the following way:

In a first phase, the steps mentioned in the chapter 2 "Preliminary Check" are applied on the first set of data received from the country. The aim of this check is to early detect any acquisition problem like GPS settings error (coordinate system or units), inadequate use of the GPS or systematically missing data, and explain to the survey institution how to correct it.

In a second phase, the steps mentioned in the chapter 3 "Deep Check" are applied once the whole dataset is in house. Corrections are proposed and submitted to the site for validation before integration in the main data set.

During the whole process it is important to follow up closely with the implementing institution in order to make sure that:

- necessary actions are taken regarding the data collection or data entry errors observed during the preliminary check
- answers are obtained regarding the errors observed in the context of the deep check

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).

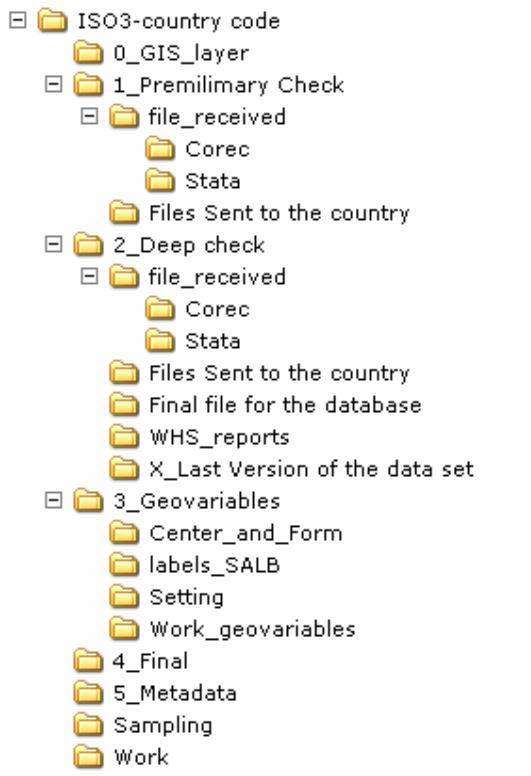


Figure 3 - Folder structure used during the application of the protocol

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.

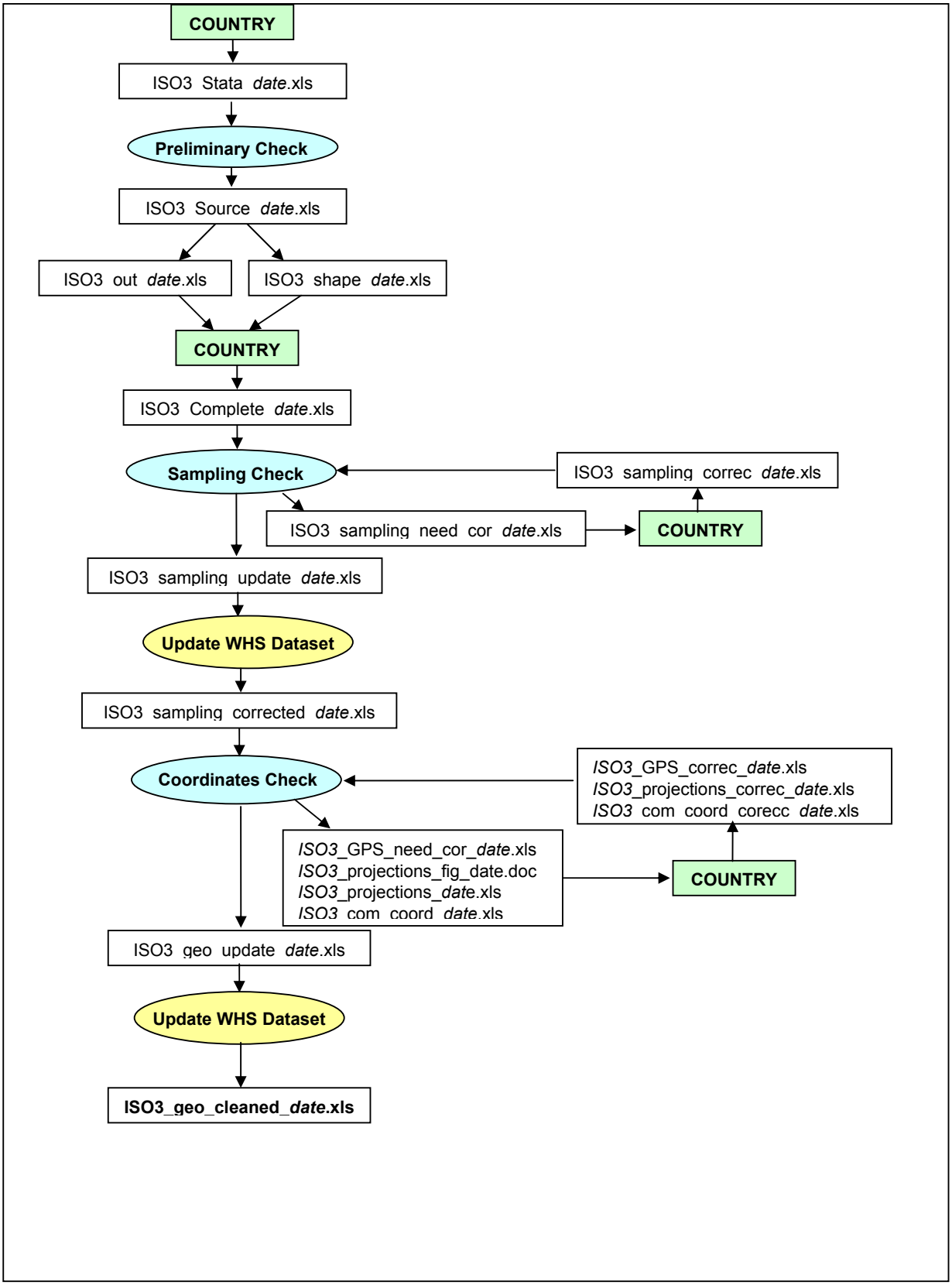


Figure 4 - Process and naming convention used in the context of this protocol

1.2.2. Additional Materials

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

- StatTransfer
- Excel
- ArcView 3.2

This protocol refers to a set of files and documents that can be found in the "Annexes_GI_cleaning.zip" file that can be downloaded from the WHO WHS Website (http://www3.who.int/whs/P/instrumentandrel8293.html). These are:

- Arcview project of reference ("ISO3_WHS.apr")
- ArcView extensions ("centroid.avx" and "XTOOLSMH.avx")
- Reference files for data conversion ("reference.xls" and "reference_deep_check.xls"), file preparation ("reference_header.xls" and "reference_com_coord.xls") and data set update ("reference_update").
- Stata code for the check of the sampling informnation consistency ("general.ado")

In the context of the WHO WHS and in order to face the high number of countries involved, other materials have been used to support the cleaning process:

- An Access database containing:
    - each country's specific information (such as implementing organization contacts, Sampling details)
    - information related to the GPS devices (number of devices sent, name of the person representing the focal point)
    - information allowing to follow up the correspondence with the sites (date and name of the files sent to and received from the survey institution)
- An Excel follow up table created to manage the whole process and flow of information. This table is continuously updated allowing to get a clear idea of the progress status for each country.
- A lineage document where all the information that would explain a particularity of the data cleaning process or the information stored in the particular subset created is reported. This document is called Lineage_Cleaning_ISO3.doc and located directly under the main folder "*ISO3"*.

2. Preliminary check

This chapter concerns the steps to be applied when the first set of data is received from the country. These steps are listed in order of priority.

2.1. Sampling Information Availability

Before doing any technical work it is important to verify that the following pieces of information are available:

- the confirmation from the Country Officer that the sampling frame has been approved
- the full sampling information which includes:
    - the sampling size as contracted
    - the sampling size + the estimated number of non respondents
    - the sampling size finally agreed
    - the sampling size finally agreed + estimated number of non respondents

- the confirmation regarding the type and number of units used at each level of the sampling frame (PSU, SSU,…) as entered in the database.
- a key correspondence table allowing to make the link between the codes reported in the dataset and the labels of the different sampling units. The key correspondence table is saved under the form of an Excel table named *ISO3*_Samp_key_table_*date*.xls located in the folder "Sampling" (Figure 4).
- the distribution of each cluster within the 1st and 2nd level administrative units of the country in question at the moment of the survey.

If some of these pieces of information are not available it is necessary to start a process, in collaboration with the Country Officer in order to obtain them from the survey institution.

2.2 Data Preparation

In order to perform the preliminary check of the data, specific variables are extracted from the main data set. These variables are:

- Section 0000. Coversheet:
  - -Q0002 : **id**: Household ID

- Section 0100. Sampling Information:
  - **Q0100**: Primary Sampling Unit (PSU) Name/Code
  - **Q0101**: Secondary Sampling Unit (SSU) Name/Code
  - **Q0102**: Tertiary Sampling Unit (TSU) Name/Code
  - **Q0103**: Quartenary Sampling Unit (QSU) Name/Code
  - **Q0104**: Setting
  - **Q0105s**: Specified Setting

- Section 0200. Geocoding Information :
  - **Q0200_1, Q0200_2, Q0200_3** : Latitude
  - **Q0201_1, Q0201_2, Q0201_3** : Longitude
  - **Q0202**: Waypoint

In addition to these data, the following variables not collected in the field are also generated:

- **casemiss:** variable created to identify the missing cases:
  - 1 = record with missing individual questionnaire
  - 0 = records with completed individual questionnaire
- **country:** country's ISO3 code

2.2.1 Transfer from Stata to Excel

In order to perform the rest of the check, the data have to be transferred from Stata to Excel.

To do so, use the StatTransfer Program, filling the Transfer window as follows:
1. **Input File Type**: Stata
2. **File Specification**: Enter the corresponding country stata file
3. **Output file type**: Excel
4. **File Specification**: *ISO3*\1_Premilimary Check\file_received\Stata and save the file as *ISO3*_stata_*date*.xls.

2.2.2 Data Standardization in Excel

First, it is important to make sure that the list of variables in the Excel table correspond to the one reported in the section 2.2. The person in charge of the database would need to be contacted if some variables are missing. In case these data are not available the reasons why should be indicated in the Lineage_Cleaning_*ISO3*.doc.

The steps to follow from that point depend on the units' setup of the GPS devices used in the fields. The Decimal Degrees should normally have been used but it may appear that the devices have been set up to another format. The **Table 1** describes the steps to follow depending of the format used for the collection of the Geographic coordinates. Under the step 2 in the table the file called "reference.xls" is needed, this file is located in the "Annexes_GI_cleaned.zip" file that can be downloaded from the WHO WHS web site.

Format of the coordinates collected:	**<u>Decimal degrees (hddd.ddddd°):</u> (i.e.: 46.25098°)**	**<u>Degrees and minutes (hddd°mm)</u> (i.e.: 46°15)**	**<u>Degrees, minutes and minutes decimal (collected without separator : hddd°mmmmm)</u> (i.e.: 46°15059') Etrex default setting as received in Stata**	**<u>Degrees, minutes and seconds (hddd°mm'ss)</u> (ie:46°15'03.5")**
Step 1	In the country Excel file exported from Stata (*ISO3*_stata_*date*.xls), insert a column on the right side of the following columns: id, q0105s, q0200_1, q0200_3, q0201_1, q0201_3			
Step 2	From the reference.xls file select the 2nd row copy and paste it in the header of the *ISO3*_stata_*date*.xls.	From the reference.xls select the first row below the note "**Deg Conversion (hddd°mm to hddd.ddddd°)**" copy and paste it in the header of *ISO3*_stata_*date*.xls.	From the reference.xls file select the first row below the second note "**Deg Conversion (hddd°mm.mmm (collected as hddd°mmmmm) to hddd.ddddd°)**" copy and paste it in the header of *ISO3*_stata_*date*.xls.	From the reference.xls file select the first row below the note "**Deg Conversion (hddd°mm.ss' (collectyed as hddd°mmss) to hddd.ddddd°)**" copy and paste it in the header of *ISO3*_stata_*date*.xls.
Step 3	Copy the first header red cell and paste it on the whole corresponding column → the formula is copied to the whole column. Do the same for each red header cell.			
Step 4	Select the whole file, copy it and **paste only the values** into a new Excel file. Enter "LAT" and "LONG" in the new columns resulting from the lat/long variables calculation. Save this new file as *ISO3*_source_*date*.xls under the folder 1_Premilimary Check\file_received. Close the *ISO3*_stata_*date*.xls without saving changes.			

Table 1- Steps to follow to standardize the data in Excel

2.3. GPS coordinates Check

This chapter lists a set of tests and controls that have to be applied when the data collection teams are still in the field in order to early detect and stop eventual misuse of the GPS devices (wrong set up of the units for example) or systematically missing data.

The problems that need to be reported to the implementing institution are:

- GPS coordinates not reported in decimal degrees => ask the institution to check which setup has been used and to make the necessary modification on the GPS devices if needed.
- important number of records without any GPS coordinates => ask the institution to make sure that the coordinates are collected and reported for each households.
- important number of coordinates appearing outside the country => ask the institution to make sure that the GPS users are verifying that the number of Satellites' signals received are enough for the reading (check of the location accuracy) or that the initial set up has been done properly (refer to the document "GPS Test and Use in the field" that can be downloaded from the WHS Web site).

These potential errors are identified using the processes reported in the coming sections. **The Lineage_cleaning_*ISO3*.doc file has to be open in order to report all the changes done during this process.**

The email containing these observations has to be **immediately sent** to the implementing institution with a follow up to make sure that the problems have been solved.

2.3.1 Lat/Long Coordinates Display in ArcView

This process aims to display the GPS coordinates in order to identify eventual setting errors or coordinates appearing outside the country border.

Once the dataset has been standardized in Excel (section 2.2.2) it is possible to visualize the GPS locations in ArcView proceeding as follows:

2.3.1.1 File Preparation

1) Open the Excel *ISO3*_source_*date*.xls.
2) In Excel select the whole worksheet and apply the Column> Autofit function from the Format menu (this allows to display the full values in each cells).
3) Make sure that the headers of the columns are as follow:

- id
- d
- country
- casemiss
- q0100
- q0101
- q0102
- q0103
- q0104
- q0105s
- q0200_1
- q0200_2
- q0200_3
- LAT
- q0201_1
- q0201_2
- q0201_3
- LONG
- q0202

The column "d" corresponds to the last digit of the id code where "1" stands for the Test cases and "2" for the Retest cases.

4) Save the file as *ISO3*_working_*date*.xls under the "Work" folder. This file would be the working file for the application of the following steps. The folder "Work" would represent the working directory.

2.3.1.2 Data Importation in ArcView

1) Save the file *ISO3_working_date*.xls in DBF4 format (dBASE IV, *.dbf) and name it *ISO3*_working_*date*.dbf, still under the "Work" folder (Figure 3).
2) Copy the ISO3_WHS.apr project from the "Annexes_GI_Cleaning.zip" file into the *ISO3* folder.
3) Open the ISO3_WHS.apr project and save it directly under the *ISO3* folder renaming it with the corresponding country ISO3 code. Set the working directory to *ISO3*.
4) In ArcView import the *ISO3*_working_*date*.dbf table using the “Add Table” module in the Project menu (Figure 5).

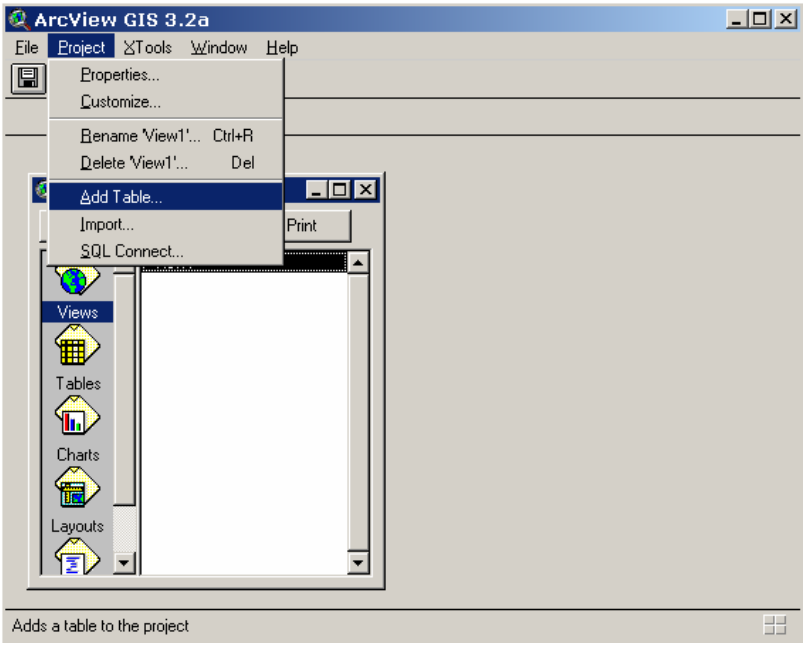


Figure 5 - The ArcView Module "Add Table"

5) Open a View and add the *ISO3*_working_*date*.dbf in that view following these steps:
   - From the View menu, choose “Add Event Theme”.
   - Click the “XY” button.
   - In the Table field, select the *ISO3*_working_*date*.dbf
   - In the “X field” select “LONG” and in the “Y field” select “LAT”.
   - Press OK
6) Make sure that the records' values contain the right number of digits (5). If this is not the case come back to the *ISO3*_working_*date*.xls file and check the format of the cells containing the LAT and LONG information (they should be in numeric format with Decimal places equal to 5).
7) Convert the theme *ISO3*_working_*date*.dbf as a shape file, name it *ISO3*_working_shape_*date*.shp and save it under the "Work" folder (Figure 3).
8) Now that the GPS coordinates are displayed in ArcView it is possible to visually check the coordinates' coherency.

2.3.2 Identification of the GPS coordinates located outside the country

For this verification it is necessary to have a shape file containing the delimitation of the international border of the country in question. In the context of the WHO WHS the UN cartographic section international border standard has been used (http://www.un.org/Depts/Cartographic/english/htmain.htm).

The first steps consist to individualize the GPS coordinates outside the country border in a new shape file.

1) Overlay the international border with the GPS coordinates points.
2) Identify the records outside the country border by selecting them on the screen using the "Select Feature" tool. The records without coordinate information should also appear outside the country as 0°Lat / 0°Long.
3) Verify that the theme *ISO3_working_*shape*_date*.shp is selected, go to the Theme Menu and select "Convert to shape file". Name the new shape file as "*ISO3*_out_*date*.shp" saving it under the "Work" folder. This new shape file contains the records outside the country border selected in point 2.
4) Select the *ISO3_working_*shape*_date*.shp, open the attribute table and revert the selection in order to have all the points inside the country being selected.
5) Go back to the View, verify that the theme *ISO3_working_*shape*_date*.shp is selected, go to the Theme Menu and select "Convert to shape file" naming the new shape file as "*ISO3*_in_*date*.shp" and saving it under the "Work" folder (Figure 3). This file contains the GPS coordinates which are located within the country international border.

2.3.3 Analysis of the records for which the GPS coordinates are outside the country border

The objective of this section is to identify possible sources of errors for the GPS coordinates located outside the country's border.

For that, open the *ISO3*_out_*date*.shp in ArcView, put the attribute table in editing mode and follow these steps:

0) Add a new field ("out", string, 50) in the table of attribute where the various remarks observed in the following steps have to be entered.
1) Identify the records for which no coordinates are reported => enter the mention "no coordinates" in the field "out" as the ID of these records have to be sent to the survey institution.
2) Sort the record by cluster number. Let the cluster for which there are only one or two records and concentrate on the cluster for which there is an important number of records outside the country.
3) Add the cluster and administrative boundaries map in the view (if available).
4) Focus on the first cluster presenting an important number of records outside the country
5) For that cluster select all the records concerned and compare their GPS coordinates with:
    - the coordinates of the other records from the same cluster that are located within the country (in the *ISO3*_in_*date*.shp file)
    - the location of the cluster if available in a digital map
6) Try to see if an explanation can be found (shift, typing mistake, different projection)
7) Enter the possible explanation in the field "out" as well as if there is no explanation, as this has to be mentioned to the survey institution.

8) Move to the next cluster showing an important number of points outside the country and restart from point 5).
9) At the end of the process stop the editing mode and save the changes.

To create the file summarizing the situation with possible explanation for the errors by cluster follow these steps:

1) Open the file *ISO3*_out_*date*.dbf into Excel and save it as *ISO3*_out_*date*.xls under ISO3\1_Premilimary Check\Files Sent to the country.
2) Sort the records outside the country by ID.
3) Delete the columns not related to GPS ("d", "country", and "casemiss").
4) Delete the records that do not need corrections/explanation.
5) Add the header and legend based on the reference file called "reference_header" in the "Annexes_GI_Cleaning.zip" file.
6) Save the changes in the SO3_out_date.xls file.

2.3.4 Identification of clusters presenting a particular shape in the country

This section is used in order to identify the clouds of points that are presenting a particular shape or that are shifted regarding the rest of the clouds of points. The reasons for this type of errors are generally:

- a wrong setting for the GPS units
- not enough satellites signals received when taking the reading

The following steps have to be followed in order to identify these cases:

1) In ArcView select the theme *ISO3_in_date*.shp (theme saved in the "Work" folder) and open its attribute table.
2) Put the attribute table in editing mode and add a new field ("shape", string, 50) in the table of attribute where the various remarks observed in the following steps have to be entered.
3) Cluster by cluster select all the records located in the same cluster by using the query builder.
4) Zoom on the selected cluster and see if there is:
   - an horizontal or vertical alignments of the points that could be due to a lack of precision in the GPS reading (number of decimals or a data entry mistake
   - an important dispersion of the points that could be due to a problem with the units or projection setup of the GPS device (position format and map datum). This observation is done by comparing the shape of this particular cluster cloud of points with the clouds of points from the neighbouring clusters.

   Do not care about the outliners for the moment as long as there are no more than one or two.
5) For the clusters where one of the problem mentioned in point 4 is observed, enter the indication of its possible source in the field "shape".
6) Stop the editing mode and save the changes.

At the end of this process, in the following steps a file is created containing the summary of the situation with possible explanation for the errors found by cluster to be sent to the country for action, following these steps:

1) Open the file *ISO3*_in_*date*.dbf into Excel and save it as *ISO3*_shape_*date*.xls under ISO3\1_Premilimary Check\Files Sent to the country.

2) Sort the records by ID.
3) Delete the columns not related to GPS ("d", "country", and "casemiss").
4) Delete the records that do not need corrections/explanation.
5) Add the header and legend based on the reference file called "reference_header" in the "Annexes_GI_Cleaning.zip" file.
6) Save this file.

2.4. Other Geographic Data Preliminary Check

In the context of the WHO WHS apart from the check the GPS reading the percentage of missing data was automatically calculated and a report prepared for the country in order to make the survey institution aware of particular variables for which information was missing.

2.5. Send the Emergency Email to the Country

Once the preliminary check finalized (process mentioned in the sections 2.3 and 2.4) an email has to be directly sent to the survey institution with copy to the Country Officer. This email contains the files saved under ISO3\1_Premilimary Check\Files Sent to the country:

1) the file *ISO3*_ out _*date*.xls listing the GPS coordinates located outside the country (section 2.3.1.3)
2) the *ISO3*_shape_ *date*.xls listing the clusters presenting a particular shape (section 2.3.1.3)
3) the report of the percentage of missing data by variable (section 2.4.)

It is also important to mention in this email if other data seem to have been reported in the wrong field due for example to the fact that not enough fields were at disposal for a particular section (sampling information for example).

3. Deep check

The Deep check of the geographic component of the WHO WHS consist in a detailed systematic check of all the geographic variables.

This Deep check has to be applied once all the test and missing cases (non respondent) have been received for the concerned country (GPS and other geographic information) or when the only missing data are the ones that have been explained and cannot be completed.

This implies that the information regarding the number of Households visited in the field has been received.

The Deep check is separated in two main phases:

- first the sampling information's check, cleaning and update;
- then the GPS coordinates' check, cleaning and update.

3.1 Sampling Information Check

The first information for which the consistency need to be checked in the dataset is the sampling codes. This is due to the fact that this information has an impact on all the other geographic fields during the rest of the cleaning process.

3.1.1 Data Preparation

This Sample check is based on the whole completed dataset with the help of sampling information provided by the country.

3.1.1.1. Database

The next steps are followed in order to prepare the dataset:

1) Convert the version of the Stata file containing all the records for the section 0100 and 0200 of the questionnaire (Figure 1 and 2) into an Excel file following the steps mentioned in the section 2.2.1 "Transfer from Stata to Excel".
2) Save the resulting Excel file under *ISO3*\2_Deep check\file received\Stata with the name *ISO3*_complete_*date*.xls.
3) Make sure the names of the columns are as follow:
   - id
   - country
   - casemiss
   - q0100
   - q0101
   - q0102
   - q0103
   - q0104
   - q0105s
   - q0200_1
   - q0200_2
   - q0200_3
   - q0201_1
   - q0201_2
   - q0201_3
   - q0202
4) Create a file containing only the sampling data from the data set received (id, q0100, q0101, q0102, q0103) and save it as ISO3_sample_*date*.xls under the "Work" folder.

3.1.1.2. Sampling Key Correspondence Table

In order to check the information reported in the sampling information variables the Sampling key correspondence table needs to be ready.

This table should be under *ISO3*\Sampling folder and should contain the codes and labels for each level of the sampling as well as the information linking each cluster to the first and second administrative level units.

If this table is not ready it has to be completed before being able to start the deep check and then saved in the *ISO3*\Sampling folder.

3.1.2. Sampling Information Consistency Check

The aim of this sampling check is to compare the sampling codes entered in the dataset under the variables q0100, q0101, q0102 and q0103 with the Sampling Key correspondence table.

In order to automate the check of the data a Stata code (do file) has been created by Emese Verdese and Agnès Prudhomme. This code called "general.ado" is part of the "Annnexes_GI_Cleaning.zip" file and its text can be found in the Annexe 2 of this document. The use of this do file is possible through the following steps:

1) Check if the codes at the lowest level are unique. If not, create a new variable with unique codes (for example by merging all the sampling levels into one code);
2) Create a Stata file containing the list of sampling codes (q0100, q0101, q0102, q0103, new variable if necessary) based on the key correspondence table: transfer the data from Excel to Stata using StatTransfer. Save the Stata file as sample_key_*ISO3*.dta under the folder "Sampling";
3) Transfer the ISO3_sample_*date*.xls into Stata using StatTransfer. Save it as sample_data_*ISO3*.dta under the "Sampling" folder;
4) Open Stata and open the program "general.ado". In the program replace the mention "FOLDER NAME" (cd "FOLDER NAME") by the path and name of the folder "*C:**ISO3*\Sampling"
5) Run the program in order to check the last stage of sampling:
    - Open the ado file called "general.ado";
    - Run the ado file;
    - Write in the "Stata command" window: "general *ISO3 q0103*" (where q103 corresponds to the last stage of selection);
    - Open the output file "sample_PSUreport_*ISO3*.dta" which lists the IDs and the codes which need to be corrected (the codes entered in the data set which do not match the codes from the correspondence table);
    - Transfer "sample_PSUreport_*ISO3*.dta" into Excel and save it as "sample_PSUreport_*ISO3*.xls" still under the folder "Sampling".
6) Run the program in order to check the consistence between the codes of the different stages (PSU SSU TSU QSU):
    - Write in the "Stata command" window: "general *ISO3* q0103 q0100 q0101 q0102 " (the last stage of selection should be written first);

- Open the output file called "sample_report_*ISO3*.dta" which list the IDs and q0103, q0100, q0101, q0102 which need to be corrected and the proposition of correction: q0100_cor q0101_cor q0102_cor q0103_cor.
- Transfer "sample_report_*ISO3*.dta" into Excel and save it as "sample_report_*ISO3*.xls" still under the folder "Sampling"

7) Check the propositions of corrections manually in order to double check the corrections proposed by the program:
8) Open the Excel file "sample_key_ISO3.xls" and verify that the sampling corrections proposed by the program are correct based the Sampling Key table.

At the end of this process, a file to be sent to the country listing the sampling codes that need verification and the correction proposed is created following these steps:

1) Open the files sample_PSUreport_*ISO3*.xls and sample_report_*ISO3*.xls files.
2) Make one file out of these two files and save it as *ISO3*_sampling_need_cor_*date*.xls under ISO3\2_Deep Check\Files Sent to the country.
3) Sort the records by ID.
4) Add the header and the legend based on the reference file called "reference_header" from the "Annexes_GI_Cleaning.zip" file.
5) Save the changes done to this file.

3.1.3. Send the Sampling Check Email to the Country

Once the sampling check finalized (process mentioned in the section 3.1.2) an email has to be directly sent to the survey institution with copy to the Country Officer. This email contains the *ISO3*_sampling_need_cor_*date*.xls file as attachment.

In the email the survey institution is asked to report their own corrections or their acceptation of the proposed change **in the same Excel file that has been sent** in the following way:

- If they accept the proposed corrections ask them to put the corrected cells ("_cor" cells) in green.
- If they do not accept the proposed correction ask them to report their own correction in the "_cor" column and to put the concerned cells in purple.
- Regarding the missing (cells in orange) or uncorrected (cells in yellow) cases, ask them to report their own correction in the "_cor" column and to put the concerned cells in purple.

In the email the following comment are also added:

- the indication of possible source of the problems or pointing to elements that would need to be changed.
- the reasons of particular problems like the lack of sampling codes for a lot of cases.

3.1.4. Receiving the Sampling Correction from the Country

The steps reported here concern the corrections provided by the survey institution in answer to the sampling check email sent at the end of the previous section (3.1.3).

If the corrections are not reported in the same file than the one sent (see section 3.1.3) it is necessary to integrate them in the corresponding file.

Then save the received corrections under ISO3\2_Deep Check\file_received\Corec as ISO3_samp_correc_*date*.xls

3.1.4.1. Correction Received Check

Open the last version of the file *ISO3*_sampling_need_cor_*date*.xls sent to the country as well as the file *ISO3*_samp_correc_*date*.xls received from them with their correction and proceed as follows:

1) Make sure that the cells mentioned in green by the country in the *ISO3*_samp_correc_*date*.xls concerns cells for which a correction was proposed or cells for which no correction was needed. If this is the case these records are considered as corrected, if this is not the case put these cells in purple.
2) Make sure that the survey institution has proposed figures for the missing data (orange cells) or the records for which a problem has been identified without being able to find a correction (cells in yellow). These cells should normally appear as purple in the file received. If this is not the case, an explanation should at least be given by the survey institution.
3) Check the consistency of the newly proposed figures. Put the correct cells in green in the *ISO3*_samp_correc_*date*.xls. Use the same colour codes for the cells that still need corrections. If necessary (in case not all the cells are in green at the end of the process) prepare a new *ISO3*_samp_need_cor_*date*.xls and repeat the process from the section 3.1.3.
4) Make sure that the records for which the sampling codes could not be corrected are updated as empty values during the process. To do so the following code "NA" should be entered in the q0100 to q0103 variables for each one of the concerned records.
5) This loop has to be applied until all the records for which a sampling correction is needed have been corrected or explained. Once this is the case, it is possible to integrate the corrected figures into the main dataset as explained in the next section.

3.1.4.2. WHS Dataset Sampling Update

The following process has to be applied in order to update the main WHS dataset with the corrections made to the sampling variables:

1) From the last file *ISO3*_samp_correc_*date*.xls only the following columns are kept:
    - id
    - q0100_cor
    - q0101_cor
    - q0102 _cor
    - q0103_cor
2) Save the file as *ISO3*_sampling_update_*date*.xls under *ISO3*\2_Deep Check\Final File for the database.

3) Rename the header in order to fit the original variable name (delete the mention "_cor").
4) Use the file reference_update.xls from the "Annexes_GI_Cleaning.zip" file and copy the cells from V2 to AJ2 and paste them at the same place in the *ISO3*_sampling_update_*date*.xls file.
5) Copy these cells and paste them in the whole corresponding columns. Make sure that all the values are appearing without codes.
6) Select all these new cells except the first line corresponding to the header. Paste them into Word as unformatted txt.
7) In the whole document replace:
    - ^t by ^p
    - NA by $SYSMIS ($SYSMIS is needed for the blank items if they are not text).
8) Save the changes done to the file.
9) Send the file to the person in charge of the datasets in order to update the main database.
10) Convert the updated version of the Stata file using the steps mentioned in the section 2.2.1 "Transfer from Stata to Excel".
11) Save the resulting Excel file under *ISO3*\2_Deep check\file received with the name *ISO3*_sampling_corrected_*date*.xls

3.1.4.3. Conclusion Sampling Information Check

At the end of this whole sample check the main WHS data set have corrected sampling codes.

Nevertheless the correction of the sampling codes is rechecked during the coordinates check. Indeed the coordinates check allows the confirmation of the sampling cleaning process and may allow identifying few cases that had not been highlighted during the Sampling Information check. These cases being corrected and updated at the same time than the GPS coordinates.

3.2 GPS Coordinates Check

The Deep check of the GPS coordinates is based on the whole completed dataset with the help of external data such as the sampling information and the maps available.

This part of the process is very close to what has already been applied during the preliminary check (see section 2); the major difference is that this time corrections are proposed and that only the Test cases are concerned as the Retest should not be corrected.

The process is applied once the sampling variables have been cleaned and updated in the main data set. Nevertheless as mention earlier this step also allows to double check the sampling information and eventually highlight wrong sampling attribution that have not been corrected (i.e. if a coordinates fall within another cluster it may be necessary to recheck the sampling codes coherency). In that case the sampling codes are also corrected and updated following the same process than for the coordinates.

The process consists in checking the whole set of coordinates into ArcView and enter the corrections in an Excel file containing the original values and the proposed corrected ones.

3.2.1 Data Preparation

This section concerns the preparation of the different files used during the coordinates check process.

3.2.1.1 Lat/Long Coordinates Display in ArcView

The next steps have to be followed to prepare the data set in order to keep only the Test cases and display the coordinates into ArcView:

1) Open the *ISO3*_sampling_corrected_*date*.xls file and make sure the names of the columns are as follow:
   - id
   - country
   - casemiss
   - q0100
   - q0101
   - q0102
   - q0103
   - q0104
   - q0105s
   - q0200_1
   - q0200_2
   - q0200_3
   - q0201_1
   - q0201_2
   - q0201_3
   - q0202
2) The steps mentioned in the section 2.2.2 "Standardization of the data in Excel" have to be applied to this file in order to create the "d", "LAT" and "LONG" columns.
3) Proceed as follows in order to keep only the Test cases:

Sort the file by "d" descending, select all the cases where d=2 and delete them.

4) Save the file as *ISO3*_cor_GPS_test_*date*.xls under ISO3\Work.
5) Make sure that the column LAT/LONG format is numeric with 5 digits and then save the file as *ISO3*_cor_GPS_test_*date*.dbf
6) Import the ISO3_cor_GPS_test_*date*.dbf file in ArcView following the steps mentioned in the section 2.3.1 "Display of the lat/long coordinates in ArcView" and save the resulting shape file as ISO3_cor_GPS_test_shape_*date*.shp. This shape file is used to visualize the coordinates display.

3.2.1.2 Working File Preparation

In the following step a working Excel file is created where the variables' values are kept as original and the corrections proposed directly entered in new columns:

1) In the *ISO3*_cor_GPS_test_*date*.xls file copy the following columns:
   - q0100
   - q0101
   - q0102
   - q0103
   - q0200_1
   - q0200_2
   - q0200_3
   - LAT
   - q0201_1
   - q0201_2
   - q0201_3
   - LONG
2) Paste these columns on the right of all the existing columns, put all the figures in blue and rename the columns as:

- q0100_c
- q0101_c
- q0102_c
- q0103_c
- q0200_1_c
- q0200_2_c
- q0200_3_c
- LAT_c
- q0201_1_c
- q0201_2_c
- q0201_3_c
- LONG_c

3) In these new columns are entered directly all the proposed corrections while the original columns are not changed and are used as reference. All the corrections mentioned in the following steps are entered in these blue "_c" columns.
4) Create a new column called "**cor_GPS**" and another one called "**comments**" where the proposed comments and corrections' explanations are entered.
5) Save the changes done in the *ISO3*_cor_GPS_test_*date*.xls file.

3.2.1.3 Digital Maps Preparation

In the context of the WHS the data collection process included the search for:
- a digital map delimiting the clusters
- a digital map delimiting the 1st and 2nd administrative level units in the context of the SALB project

If no map has been found, the data cleaning has to be done without this source of information.

If digital maps delimiting the clusters and/or the 1st and 2nd administrative level units exist and are available it is first important to homogenize them in order to be able overlaying them with the cluster location in ArcView by looking at the map's representativity and format as follows:

1) The representativity of the administrative information entered in the key correspondence table should be identified in order to find the map corresponding to the administrative list used in the context of the survey. The administrative units' list used in the context of the survey is then compared with the SALB tables of codes and historic changes in order to determine its representativity.
   In case only a part of the administrative units have been surveyed the full list would need to be asked to the survey institution.
2) The reference format is the ArcView shape file. If the maps at disposal are not in this format, please make the necessary conversion.
3) The GPS units should have been setup in such a way that the latitude and longitude coordinates are given in decimal degrees. This allows us to use the geographic projection as reference for the overlay in ArcView. It is therefore important to make sure that all the vector digital maps are unprojected before starting the process (ask the necessary information to the institution that provided the map if necessary).
   The only limitation in this choice may be linked to the use of some satellite images that may be projected but in any case it is better to start from the geographic projection with all the vector layers and to project them if necessary.

3.2.2 GPS Coordinates Consistency Check

This section concerns the visual check of the GPS coordinates in ArcView including the calculation of each cluster's Weighted Center of Gravity and the comparison with digital maps when available.

First get all the files ready to start the process:

1) Open the ISO3_cor_GPS_test_date.xls (in the "Work" folder).
2) Identify the records for which no coordinates are reported by sorting the table by the LAT and LONG column. Enter the mention "missing" in the cor_GPS column for these records and put the corresponding cells of the "_c" columns in orange.
3) Sort the full Excel table using the id field and Filter the file by cluster codes.
4) Open the sampling key correspondence table (*ISO3*_Samp_Key_Table_data under the folder Sampling)
5) Open the reference_deep_check.xls from the "Annexes_GI_Cleaning.zip" file as it may be useful to convert latitude/longitude.
6) In ArcView display the *ISO3*_cor_GPS_test_shape_*date*.shp.
7) Add the cluster location and administrative boundaries digital maps in the view (if available)
8) Put the distance unit to meters in the view properties.
9) Save the project.

3.2.2.1. GPS Coordinates Analysis for each Cluster

This check consists in individualising the records for which the coordinates are far from the rest of the cluster they belong to.

The steps performed are similar to the ones reported in the section 2.3.1.3 for the preliminary check.

1) In ArcView, select the *ISO3*_cor_GPS_test_shape_*date*.shp and using the "Query Builder" tool, select all the records part of the first cluster and zoom on the selected points in the view.
2) Identify the possible outliners appearing as points outside the densest part of the cloud. The "Measure distance" tool can be used to get a better sense of the possibility of having a point outside the rest of the group (distance more than 2 km for example). If this is the case:
   - identify the record id
   - search for the record in the *ISO3*_cor_GPS_test_*date*.xls file (filtered by the corresponding cluster code)
   - make the necessary corrections modifying directly the concerned value in red in the corresponding "_c" columns of the *ISO3*_cor_GPS_test_*date*.xls file. Mention in the "cor_GPS column" which variable has been modified using the following codes:
     - Q0200_1: 201
     - Q0200_2: 202
     - Q0200_3: 203
     - Q0201_1: 211
     - Q0201_2: 212
     - Q0201_3: 213

   And eventually in case of sampling correction:

- Q0100: 100
- Q0101: 101
- Q0102: 102
- Q0103: 103

In case of multiple corrections, just collate the two numbers together ordering them in an ascending way (i.e. 201212).

3) Put the cells in yellow in the case of an error that could not be explained and put a comment in the "comments" column. If the outliners is falling into another cluster it is important to make sure that this is a problem of GPS coordinates and not a wrong sampling attribution. In case of doubt, put the corresponding cells in yellow and indicate the problem in the “comments” column.
4) Measure the size of the cloud of point in order to make sure that there is no problem of projection.
5) If no corrections are necessary for this cluster, put all the "_c" cells in green in order to indicate that it has been checked.
6) In ArcView, using the Query Builder, select all the records for the next cluster, zoom on the selected points in the view and start again from point 2 until all the clusters have been checked.
7) At the end of the process save the changes done in the *ISO3*_cor_GPS_test_*date*.xls file.

3.2.2.2 Weighted Center of Gravity (WCG) Check

Once the cluster homogeneity has been checked (Section 3.2.2.1) the Weighted Center of Gravity (WCG) of each Cluster can be calculated in order to verify that no outliners have been omitted and to double check the cluster homogeneity.

3.2.2.2.1 File Preparation

1) Save the *ISO3*_cor_GPS_test_*date*.xls file as *ISO3*_cor_GPS_test_*date***.dbf**.
2) Import *ISO3*_cor_GPS_test_*date*.dbf in ArcView. Add it as event theme in the View using the LAT_c and LONG_c columns and convert it as Theme naming it "*ISO3*_cor_GPS_test_*date*_***cg***.shp".
3) From the Theme's Table of attributes delete all the records that could not be corrected or with problems that could not be solved.

3.2.2.2.2 Cluster Individualisation

The following steps allow giving a unique ID from 1 to X (where X is the total number of surveyed cluster) to each cluster.

1) Still in ArcView, open the *ISO3*_cor_GPS_test_*date*_cg.shp theme attribute table.
2) Sum the field corresponding to the cluster code and save the sum shape file as clust_sum.dbf in the "Work" folder
3) Open the clust_ sum.dbf file into Excel. Sort the file by ascending cluster code.
4) Insert a column on the right of the cluster code with the following formula: "= *cell above* + 1". Copy the formula into the whole column. Copy this column and paste only the values. Name the column GI_ID_I.
5) Save the file as clust_giid.dbf under Work.
6) Open the table clust_giid.dbf in ArcView and join it to the *ISO3*_cor_GPS_test_*date*_cg.shp using the cluster code field as the common field.
7) Add a new Field (number) in the *ISO3*_cor_GPS_test_*date*_cg.shp table of attribute and name it "GI_ID"

8) Use the ArcView calculator to copy the values of the GI_ID_I field into the GI_ID field
9) Remove all joins and save the project.

Now all the records have a GI_ID from 1 to X (where X is the total number of cluster).

3.2.2.2.3 Weighted Center of Gravity (WCG) Calculation

The following operations allow the calculation of the clusters' Weighted Center of Gravity (WCG) and the creation of a corresponding point shape file.

1) Two ArcView extensions are needed during the process (Centroids and XTools Extension). Copy the "centroid.avx" and "XTOOLSMH.avx" files from the "Annexes_GI_Cleaning.zip" file to the C:\ESRI\AV_GIS30\ARCVIEW\EXT32 folder.
2) Then the extensions must be activated proceeding as follows:
    - In ArcView, go to File/Extensions.
    - Check the boxes in front of :
    - Polygons -- > Centroids
    - XTools Extension - Metric
3) In the View, select the theme: *ISO3*_cor_GPS_test_*date_cg*.shp. Display the attribute table, and sort the GI_ID column by descending values and identify the highest value.
4) From the View click the Icon “Points to Weighted Centre of Gravity”: 

Two windows successively appear:
- first enter the ISO3 code of the selected country,
- when asking “Enter the GI_ID class number” specify the highest GI_ID value,

This script creates a shape file *ISO3*_w_cg.shp representing the WCG of each cluster including the following fields in its attribute table:
    - Weight_field (GI_ID) = Field used to calculate the WCG
    - GI_ID
    - Num_pts = Number of points in each Cluster
    - Y_wc = Latitude of the WCG
    - X_wc = Longitude of the WCG
    - Sourcethm = Field used to calculate the WCG.
5) Save the project.

3.2.2.2.4 WCG Comparison with the Cluster Clouds of Points

By overlapping the *ISO3*_w_cg.shp on the *ISO3*_cor_GPS_test_*date_cg*.shp it is possible to verify the cluster homogeneity.

A WCG appearing outside of the cluster's cloud of points would for example highlight a record for which the coordinates are far from the rest of the cluster and that has not been corrected so far.

Review all WCG and clusters' clouds of points entering eventual corrections in the *ISO3*_cor_GPS_test_*date*.xls file.

3.2.2.2.5 WCG Recalculation

In case corrections have been reported redo the steps 3.3.2.2.1 to 3.3.2.2.3 in order to obtain the corrected WCG shape file.

3.2.2.3 Map Check

This section concerns only the case for which a digital map is available and can be used to verify:

- that the cluster is displayed in the correct map unit,
- the projection used is the correct one as it may happen that a same country used different GPS units set up to measure the coordinates.

In order to ease the check the *ISO3*_w_cg.shp is used to perform the steps mentioned in the following sections.

3.2.2.3.1 Data Preparation

1) Firstly, the cluster code is reintegrated into the WCG shape file. In the *ISO3*_w_cg.shp table of attribute add 2 new fields:
   - "cluster_code", number, 16
   - "check map", string, 50
2) Join the *ISO3*_w_cg.shp table of attribute to the clust_giid.dbf using the GI_ID and use the calculator to copy the values corresponding to the cluster code into the "cluster_code" field. Remove all joins and save the project.
3) In order to verify if the correct GPS units setting has been used a new WCG's shape file is created by changing its coordinates' projection. The most probable setting that would have been used if not the recommended one (hddd.ddddd°) is the GPS device setting by default (hddd°mm.mmm):
   - Open the *ISO3*_w_cg.dbf in Excel.
   - Using the reference_deep_check.xls file (in the "Annexes_GI_Cleaning.zip" file) convert the coordinates in hddd.ddddd° (Deg Conversion (hddd°mm.mmm to hddd.ddddd°)).
4) Save the file as *ISO3*_w_cg_proj2.dbf in the "Work" folder, import it in ArcView, add it as event theme and convert it as shape file named *ISO3*_w_cg2.shp (to be saved in the "Work" folder).

3.2.2.3.2 WCG Comparison with the Map

1) In ArcView display the map and overlay the *ISO3*_w_cg.shp and the *ISO3*_w_cg2.shp with two distinct colours.
2) Going through all the WCG (*ISO3*_w_cg.shp) and using the "Identify Tool" verify that:
   - they are in the correct map unit (comparing the sampling information with the map units.)
   - the WCGs corresponding to the second projection (*ISO3*_w_cg_proj2.shp) are or not in the same map unit.

Note: It can happen that the precision of the GPS is higher than the digital map at disposal or that there is a shift in the map. In that context points at less than 10km from the map's corresponding unit should be considered as potentially in the unit.

3) Enter the corresponding remarks in the field "check_map" of the *ISO3*_w_cg.shp table of attribute using the following convention:
   - "ok" if the point corresponding to *ISO3*_w_cg.shp is in the correct map unit but not the one from the second projection (*ISO3*_w_cg_proj2.shp).
   - "both proj ok" if both WCG are in the correct map unit
   - "chge proj" if only the second WCG is in the correct map unit
   - "not in correct map unit" if both WCG are not in the correct map unit.

4) In case both projections are correct, try to find out which one is the correct one with the help of other sources of information such as the city or other sampling units locations. If this could not be solved, prepare figures in ArcView showing both projections in order to ask to the country which one is the correct one (see one example for Ethiopia in Figure 6):
    - Zoom to the concerned area
    - label the 2 possible WCG points with their corresponding cluster code and indicate clearly in the legend which colour correspond to which projection.
    - Label the map's units.
    - Insert the figures in a word document and save it as "ISO3_projections_fig_*date*.doc" under *ISO3*\2_Deep Check\File sent to the country
    - Prepare an Excel table listing all the cases concerned and two columns where the institution should quote if the 1st or the 2nd projection is the correct one. Save the file as "ISO3_projections_*date*.xls" under *ISO3*\2_Deep Check\File sent to the country.

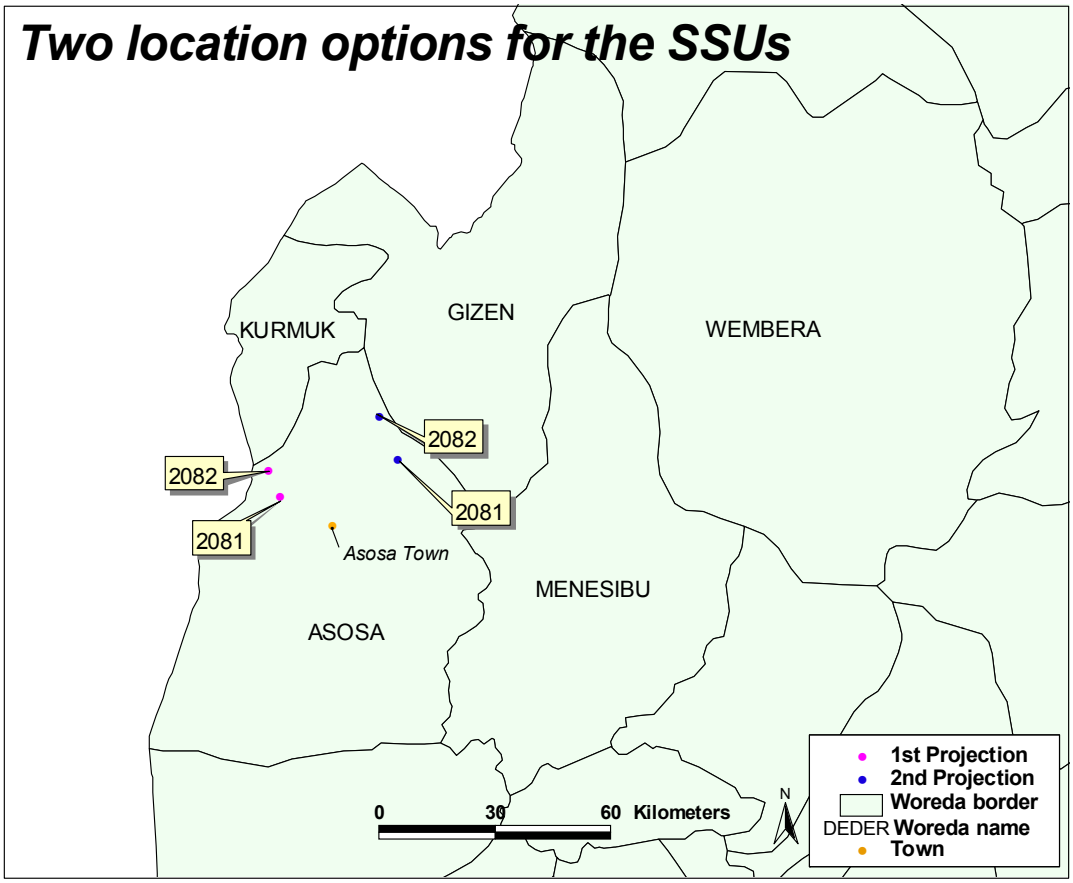


Figure 6 - Figure showing two possible projections in Ethiopia.

The same kind of figure can be created to illustrate WCG that are not appearing in the correct map unit.

5) Enter the corrections in the ISO3_cor_GPS_test_*date*.xls based on the comments entered in the field "Check map" of the *ISO3*_w_cg.dbf and in the following way:
    - change projections using the reference_deep_check.xls for the corresponding clusters and add a mention "changed projection" in the column "comments".
    - enter the comment "both projection ok" or "not in correct map unit" in that same column "comments".

6) Save the changes done to the file ISO3_cor_GPS_test_*date*.xls.

3.2.2.4 Coordinates Repetition

In the next steps clusters having more than 4 times the same coordinates collected are identified in order to highlight eventual data collection issues.

1) In ArcView open the table ISO3_cor_GPS_test_date.dbf.
2) Start editing the table and add a new field, name it “clust+coord”, String 50.
   Complete this field using the calculator function with the following formula:
   “clust+coord” = *Cluster ID*.AsString + “-” + lat_c + “/”+ long_c
3) Sum this field and save the file as Sum_Clust_coord.shp under the "Work" folder.
4) Open the Sum_Clust_coord.dbf into Excel, sort by count descending and select all the records with a count value higher than 4, delete all the others records.
   Thus all the clusters' IDs having more than 4 times the same coordinates reading are identified (the concerned coordinates are mentioned beside the cluster ID).
5) Select the first column and use Tool Data, Text to column in order to separate the coordinates in the same way than the original file.
6) Use the headers of the "reference_com_coord.xls" file from the "Annexes_GI_Cleaning.zip" file to create the file to be sent to the country. Save the file as *ISO3*_com_coord_*date*.xls under *ISO3*\2_Deep Check\File sent to the country.

3.2.2.5 Preparation of the File to be Sent to the Country

Once the whole Deep check of the GPS coordinates done a file is prepared to be sent to the country by following these steps:

1) Sort the *ISO3*_cor_GPS_test_*date*.xls by id.
2) Delete the records that do not need corrections (records without any value in the columns cor_GPS or comments).
3) Format the table and add headers and legend as in as in the "reference_header.xls" file in the "Annexes_GI_Cleaning.zip" file.
4) Save this file as *ISO3*_GPS_need_cor_*date*.xls under *ISO3*\2_Deep Check\File sent to the country.

3.2.3 Send GPS Coordinates Deep Check Email to the Country

Once the deep check of the GPS coordinates finalized (process mentioned in the sections 3.2.1 to 3.2.2) an email has to be directly sent to the survey institution with copy to the Country Officer. This email contains in attachment the following documents saved under *ISO3*\2_Deep Check\File sent to the country:

- the file ***ISO3*_GPS_need_cor_*date*.xls** containing the proposition of correction for the GPS coordinates and eventually the sampling codes.
- the two files illustrating the cases for which a confirmation regarding the projection used is needed: **ISO3_projections_fig_*date*.doc** and **ISO3_projections_*date*.xls**.
- the file ***ISO3*_com_coord_*date*.xls** listing the coordinates repetition in a same cluster.

In the email it is important to ask the survey institution to report their own corrections or their acceptation of the proposed change **in the same Excel files that have been sent** and in the following way:

- In case they accept the proposed corrections ask them to put the corrected cells ("_c" cells) in green.
- If they do not accept the proposed correction ask them to report their own correction in the "_c" column and to put the concerned cells in purple.
- Regarding the missing (cells in orange) or uncorrected (cells in yellow) cases, ask them to report their own correction in the "_c" column and to put the concerned cells in purple.

In the email also add comments on:

- the indication of possible source of the problems or pointing to elements that would need to be changed.
- the reasons of particular problems like lack of coordinates for a lot of cluster.

3.2.4. Receiving GPS Coordinates Corrections from the Country

The steps reported here concern the corrections provided by the survey institution in answer to the coordinates deep check email sent at the end of the previous section (3.3.3).

If the corrections are not reported in the same file than the one sent (see section 3.3.3) it is necessary to integrate them in the corresponding file.

Then save the received corrections under *ISO3*\2_Deep Check\file_received\Corec as:

- *ISO3*_GPS_correc_*date*.xls
- *ISO3*_com_coord_corec_*date*.xls
- *ISO3*_projections_corec_*date*.xls

3.2.4.1 Check of the GPS coordinates corrections received

The following steps allow verifying if the proposition of corrections sent to the country have been validated and if the different issues have been answered and could be solved in order to update the file *ISO3*_cor_GPS_test_*date*.xls.

3.2.4.2 Verification of the corrections proposed for the GPS coordinates consistency

1) Open the ***ISO3*_GPS_correc_*date*.xls** as well as the ***ISO3*_GPS_need_cor_*date*.xls** files.
2) Make sure that the cells mentioned in green by the country in the *ISO3*_GPS_correc_*date*.xls concerns cells for which a correction was proposed or cells for which no correction was needed.
    - If this is not the case put these cells in purple.
    - If this is the case these records have to be considered as corrected and the corresponding cells should be put in green in the *ISO3*_cor_GPS_test_*date*.xls file.
3) Make sure that the survey institution has proposed figures for the missing data (cells in orange) or the records for which a problem has been identified without being able to propose a correction (yellow cells) (these cells should normally appear as purple in the file received).
    - If no correction has been provided, an explanation should at least be given by the survey institution.

- If the GPS information cannot be provided or corrected, the mention "NA" should be entered in the corresponding cells (q0200_1 to q0201_3 and in the column"cor_GPS") and these cells should be put in green.

4) Check the records for which new figures have been provided by following the steps of the section 3.3.2 "Check of the GPS coordinates consistency", this implies the creation of the LAT_c and LONG_c columns in order to visualize the location of the records in ArcView.
   - If the corrections proposed are correct put the corresponding cells in green in the *ISO3*_GPS_correc_*date*.xls file and enter these corrections in the *ISO3*_cor_GPS_test_*date*.xls file putting the corresponding cells in green.
   - If the corrections proposed are not correct, propose new corrections and prepare a new *ISO3*_GPS_need_cor_*date*.xls file to be sent to the country.

3.2.4.3 Verification of the corrections proposed for the Projections issues

1) Open the ***ISO3*_projections_correc_*date*.xls** as well as the ***ISO3*_projections_*date*.xls** files.
2) Make sure that the survey institution indicated the corresponding projection for all the cases. If this is not the case prepare a new *ISO3*_ projections_*date*.xls to be sent to the country.
3) Enter the country's feedback In the *ISO3*_cor_GPS_test_*date*.xls file following these steps:
   - For the clusters where the projection needs to be changed: use the reference_deep_check.xls file to change the projection and add a mention "changed projection" in the column "comments" and put the cells in green.
   - For the clusters where the first projection is the correct one: add the mention "proj ok" in the column "comments" and put the cells in green.

3.2.4.4 Verification of the corrections proposed for the Coordinates repetition

1) Open the ***ISO3*_com_coord_correc_*date*.xls** as well as the ***ISO3*_com_coord_*date*.xls** files.
3) Make sure that the survey institution has proposed explanation for all the clusters having more than 4 times the same coordinates collected. Put the cells in green.
4) In case explanations are missing prepare a new *ISO3*_com_coord_*date*.xls to be sent to the country.

3.2.4.5 Conclusion GPS Coordinates Deep Check

1) If all the records sent for correction has been corrected or explained and entered in the file *ISO3*_cor_GPS_test_*date*.xls all the records of that file should appear in green.
   - Save the file *ISO3*_cor_GPS_test_*date*.xls as *ISO3*_geo_corrected_*date*.xls under "Work".
   - Proceed to point 4 "Resulting file: the cleaned version of the Sampling and Geocoding sections".
2) If this is not the case, it is necessary to send a new email to the survey institution for the correction of the remaining records. To do so start again the process from point 3.3.3 "Send GPS coordinates deep check email to the country" in order to send the remaining cases in need of corrections.
3) This loop has to be applied until all the records for which a correction is needed have been corrected or explained and the corrections entered in the file *ISO3*_cor_GPS_test_*date*.xls.

3.2.5. WHS Dataset Update

In order to update the main dataset with the corrections made to the geographic components during the process the next steps have to be applied:

0) From the *ISO3*_geo_corrected_*date*.xls, keep only the records that have been corrected or modified. To do so select all the records having values in one of the 2 columns "cor_GPS" and "comments". Save the file as *ISO3*_geo_update_*date*.xls under *ISO3*\2_Deep Check\Final File for the database.
1) In the file *ISO3*_geo_update_*date*.xls keep only the following columns:
   - id
   - q0100_cor
   - q0101_cor
   - q0102_cor
   - q0103_cor
   - q0200_1_cor
   - q0200_2_cor
   - q0200_3_cor
   - q0201_1_cor
   - q0201_2_cor
   - q0201_3_cor

Rename the header in order to fit the original variables' names (delete the mention "cor").
2) Use the file reference_update.xls from the "Annexes_GI_Cleaning.zip" file and copy the cells from V2 to AJ2 and paste them in the same place in the file *ISO3*_geo_update_*date*.xls.
3)Copy these cells and paste them in the whole corresponding columns.
4) Make sure that all the values with letters (the variables q0200_1 and q0201_1) are appearing with codes (between '', i.e.: 'S'), and that all the numerical values are appearing without codes.
5) Select all these new cells except the first line corresponding to the header. Paste them into Word as unformatted txt.
6) In the whole document replace ^t by ^p (the process may be long if an high number of records are concerned).
7) In order that the records for which the coordinates have not been collected or that could not be corrected are updated as empty values during the process proceed as follows:

In the whole document replace:
   - **'NA'** (be careful not to forget the **' '**) by **''**
   - **NA** by **$SYSMIS** ($SYSMIS is needed for the blank items if they are not text).

i.e.
IF (ID=2920294721) Q0200_2=$SYSMIS.
IF (ID=2920294721) Q0200_3=$SYSMIS.)
8) Save the file as ***ISO3*_geo_ update _*date*.doc** and send it to the person in charge of the database to update the main dataset.

4. Resulting file: the cleaned version of the Sampling and Geocoding Sections

Once all the feedback and corrections from the survey institution regarding the deep check have been obtained, the final dataset can be prepared.

This final file represents all the corrected and completed version of the Sampling of Geocoding sections.

Once the main WHS dataset has been updated it should be transferred from Stata to Excel.

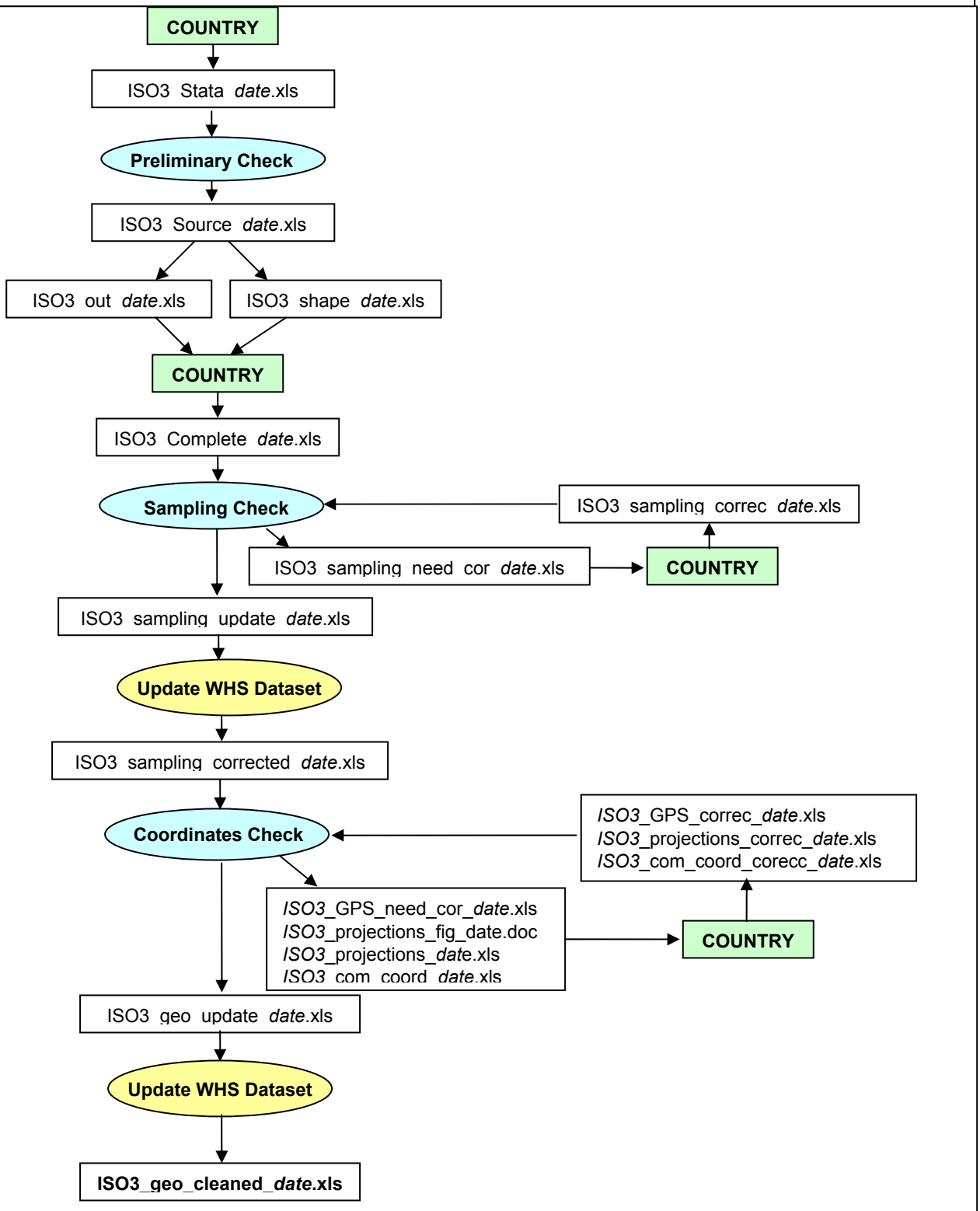
To do so, use the Stat Transfer Program, filling the Transfer window as follows:

1. **Input File Type**: Stata
2. **File Specification**: Enter the corresponding country updated sata file
3. **Output file type**: Excel
4. **File Specification**: ISO3\2_Deep Check\Final file for the database and save the file as ***ISO3*_geo_cleaned.xls**.

The file ***ISO3*_geo_cleaned.xls** represents the cleaned version of the Sampling and Geocoding Section of the WHS Dataset. It represents the **dataset of reference for the calculation of all the geographic variables and to create the "WHS geographic subset"**.

A specific protocol is available to perform these calculations. This protocol called "Generation of the GEO Subset " can be downloaded from the WHO WHS Web Site.

FLOW CHART
WHS - Cleaning Protocol - Geographic Component



ANNEXE 1: Flow Chart Cleaning Protocol - Geographic Component

ANNEXE 2: Stata code for the check of the sampling information consistency

```
program define general

*** initial settings, determination of the variables to be checked

      set more off
      clear
      capture log close
      cd  "FOLDER NAME"
      local i=1
      while "`1'"~="" {
            local arg`i'="`1'"
            local i=`i'+1
            mac shift
      }
      local checknum=`i'-2
      local country="`arg1'"
      forvalues j=1(1)`checknum' {
            local k=`j'+1
            local checkvar`j'="`arg`k''"
      }

*** checking q0100 (or some other variable being the PSU variable)
      use sample_data_`country', clear
      gen goodcase=0
      save, replace

      use sample_key_`country', clear
      preserve
      local rownum=_N
      forvalues i=1(1)`rownum' {
            local a=`checkvar1' in `i'
            use sample_data_`country', clear
            local rownumber=_N
            forvalues j=1(1)`rownumber' {
                  if `checkvar1'==`a' in `j' {
                        quietly replace goodcase=1 in `j'
                  }
            }
            quietly save, replace
            restore, preserve
      }
      use sample_data_`country', clear
      drop if goodcase==1
      keep id `checkvar1'
      save sample_PSUreport_`country', replace
      use sample_data_`country', clear
      drop goodcase
      save, replace

*** checking the additional variables using the PSU variable as reference

      restore, preserve
      if `checknum'>1 {
            tempname tab1
            if `checknum'==2 {
                  postfile `tab1' long id `checkvar1' `checkvar2'  `checkvar1'_cor
`checkvar2'_cor   using sample_report_`country', replace
            }
            if `checknum'==3 {
```

```
                  postfile `tab1' long id `checkvar1' `checkvar2' `checkvar3'
`checkvar1'_cor `checkvar2'_cor `checkvar3'_cor   using sample_report_`country',
replace
            }
            if `checknum'==4 {
                  postfile `tab1' long id `checkvar1' `checkvar2' `checkvar3'
`checkvar4' `checkvar1'_cor `checkvar2'_cor `checkvar3'_cor  `checkvar4'_cor using
sample_report_`country', replace
            }
            local rownum=_N
            forvalues i=1(1)`rownum' {
                  forvalues l=1(1)`checknum' {
                        local a`l'=`checkvar`l'' in `i'
                  }
                  use sample_data_`country', clear

                  quietly keep if `checkvar1'==`a1'
                  local rownumber=_N
                  forvalues j=1(1)`rownumber' {
                        local ide=id in `j'
                        forvalues l=1(1)`checknum' {
                              local er`l'=`a`l''
                        }
                        local ermes=0
                        forvalues l=2(1)`checknum' {
                              if `checkvar`l''~=`a`l'' in `j' {
                                    local er`l'=`checkvar`l'' in `j'
                                    local ermes=1
                                    quietly replace `checkvar`l''=`a`l'' in `j'
                              }
                        }

                        if `ermes'==1 {
                              if `checknum'==2 {
                                    quietly post `tab1' (`ide') (`er1') (`er2')
(`a1') (`a2')
                              }
                              if `checknum'==3 {
                                    quietly post `tab1' (`ide') (`er1') (`er2')
(`er3') (`a1') (`a2') (`a3')
                              }
                              if `checknum'==4 {
                                    quietly post `tab1' (`ide') (`er1') (`er2')
(`er3') (`er4') (`a1') (`a2') (`a3') (`a4')
                              }
                              quietly save sample_data_corrected, replace
                        }
                   }

                  restore, preserve
            }

            postclose `tab1'
      }

end
exit
```