

DATA CLEANING GUIDELINES – LIBERIA NATIONAL FOREST INVENTORY

NFI DATA CLEANING GUIDELINES

FORESTRY DEVELOPMENT AUTHORITY OF LIBERIA

Technical Support Provided by the Food and Agricultural Organization of the United Nations

April 2019, Monrovia

Project symbol: UTF/LIR/022/LIR

Budget code: TF/FOADD/TF5C350011682

Activity 1.3.: Data cleansing, entering, development data analysis methodology & analysis

©2019 by Forestry Development Authority of Liberia. Data Cleaning Guidelines – Liberia National Forest Inventory. Wein Town, Mt. Barclay, Montserrado County – Liberia, West Africa

Contents

Introduction	3
Data Preparation.....	4
Importing data into Collect.....	4
Manual Data Entry.....	6
Step 1 – Validation report analysis	9
Effective use of the validation report	9
Errors of interest.....	11
Carbon related errors.....	13
Step 2 – Additional survey assessment.....	18
Access route and photo assessment.....	18
Reference photos and prominent structures	20
Land use assessment.....	23
Harmonize Non-timber Forest Products.....	23
Exporting cleaned data	26
Step 3 – Tree Species Assessment	28
Data preparation.....	28
Species review and cleaning	32
Step 4 – Quantitative analysis of biophysical data	42
Preparing your checklist.....	42
Prepare raw analysis data.....	43
Analysis implementation – Single variables.....	46
Mean canopy closure.....	47
Diameter at breast height.....	49
Tree Height.....	51
Bole Height.....	53
Trees per hectare \ Basal area per hectare	53
Analysis implementation – Two variables	55
Mean canopy closure vs. Basal area per hectare	55
Diameter at breast height vs. Tree height – scatter plot.....	56
Diameter at breast height vs. Tree height – z-score analysis	58
Reporting and decisions regarding errors.....	65
Annex 1 – Error record sheets	66

Introduction

The following guidelines have been prepared to aid the data management personnel in the cleaning of the NFI data captured as part of the biophysical field work campaign conducted between June 2018 and March 2019. This manual / SoP has been updated to include additional cleaning activities that extend beyond the carbon based cleaning undertaken in the previous rounds of data cleaning activities. Three data cleaning officers have had significant experience using this manual as well as the online version of COLLECT which is used to temporarily host the database while FDA decides where the data should be stored.

The cleaning is split into three phases with an initial input data phase. Much of the data has been loaded onto the online server save for the clusters from the non-priority landscape. Experienced data cleaners will upload this data and make the clusters available to all for cleaning. Step1 of the data cleaning phase involves a review of a database validation report, which is an initial review of the errors identified by the built-in data survey validation tools (warnings and errors). Step 2 involves reviewing a number of non-carbon survey attributes that help to harmonize the database and facilitate data analysis. Step 3 involves a detailed assessment of the species identified by the field teams focusing on those species listed as *Unlisted sp.* Analysts will make use of online resources to verify and update those species identified as unlisted thereby improving the overall species data. The final phase (step 4) seeks to identify outliers using graphical tools as well as a z-score analysis to identify outliers within dbh – height relationships.

The data cleaning activities should take no longer than that 15 days with each data cleaning office expected to clean at least 2 clusters per day. The URL for the online version of the survey is:

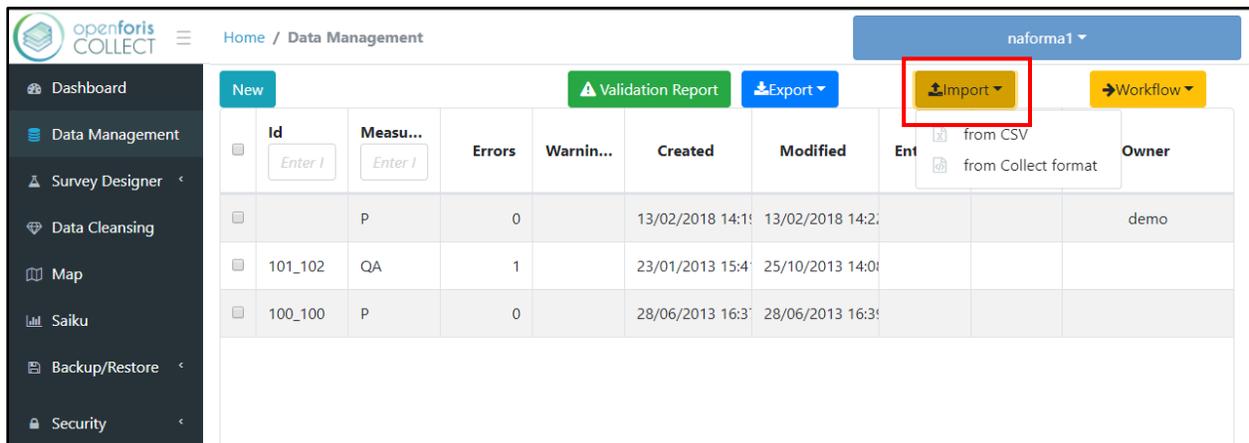
<http://www.openforis.org/wecollect>

User names have been setup for each of the data cleaning officers, passwords will be shared separately when the data cleaning begins

Data Preparation

Importing data into Collect

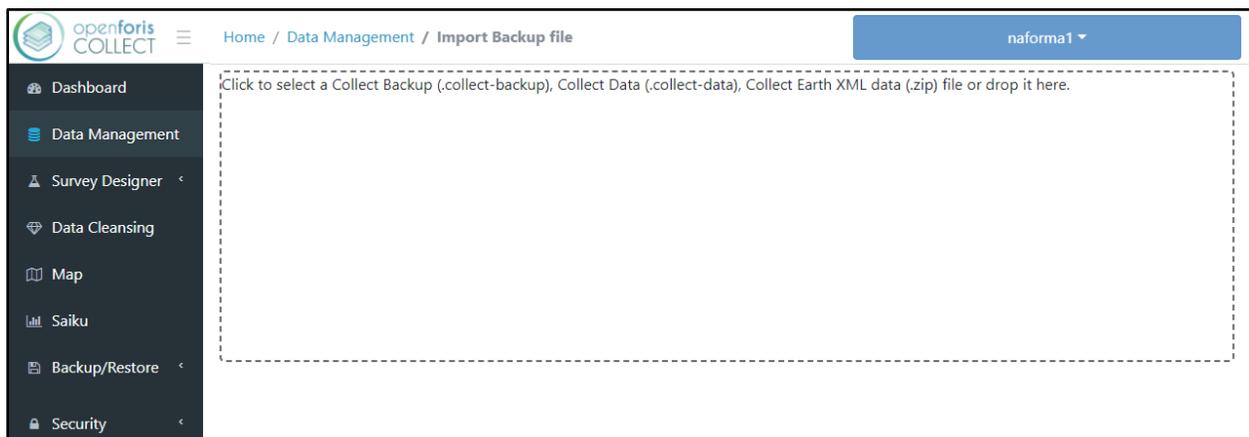
Importing data into COLLECT using exports from the COLLECT mobile software is covered on page 14 of the COLLECT Manual. These instructions are repeated here but users are encouraged to read the manual thoroughly. At the top of the Data Management screen, click on the Import yellow button and choose a format (CSV or COLLECT format).



The screenshot shows the 'Data Management' screen in the Openforis Collect web application. The breadcrumb navigation is 'Home / Data Management'. The current survey is 'naforma1'. The interface includes a sidebar with navigation options: Dashboard, Data Management, Survey Designer, Data Cleansing, Map, Saiku, Backup/Restore, and Security. The main content area has a 'New' button, a 'Validation Report' button, an 'Export' button, and a highlighted 'Import' button. The 'Import' dropdown menu is open, showing two options: 'from CSV' and 'from Collect format'. Below the menu is a table with columns: Id, Measu..., Errors, Warnin..., Created, Modified, Ent, and Owner. The table contains three rows of data.

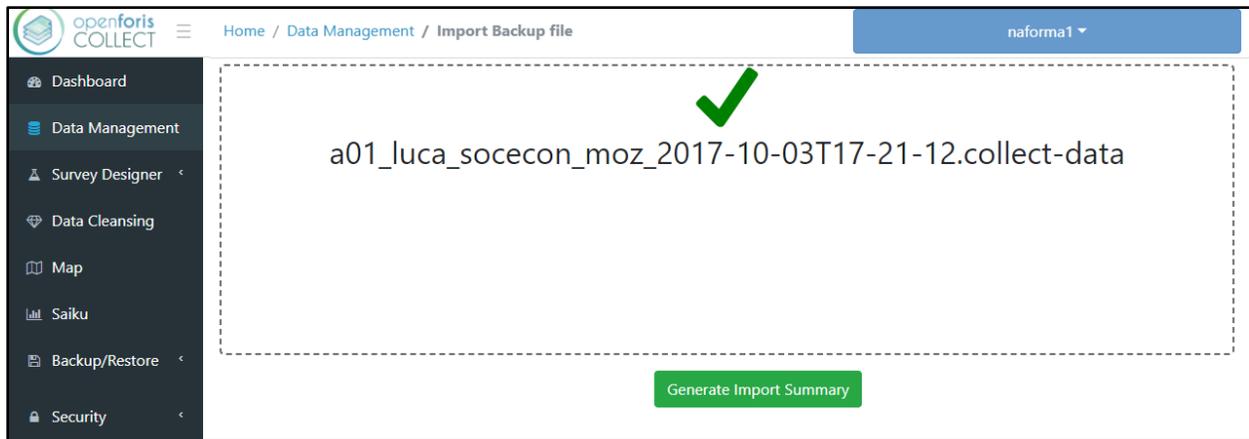
Id	Measu...	Errors	Warnin...	Created	Modified	Ent	Owner
<input type="text" value="Enter I"/>	<input type="text" value="Enter I"/>						
	P	0		13/02/2018 14:1:	13/02/2018 14:2:		demo
101_102	QA	1		23/01/2013 15:4:	25/10/2013 14:0:		
100_100	P	0		28/06/2013 16:3:	28/06/2013 16:3:		

The user will be prompted to click on an empty box to select a file to import, or to drop it inside. The data collected will be a .collect-data file

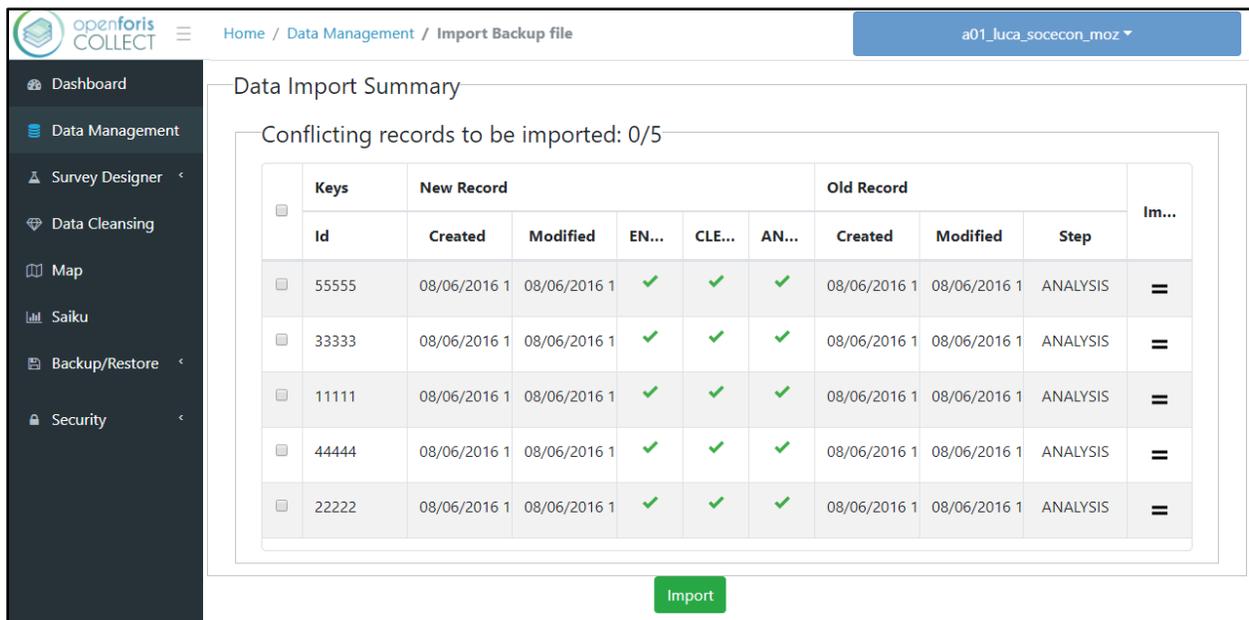


The screenshot shows the 'Import Backup file' screen in the Openforis Collect web application. The breadcrumb navigation is 'Home / Data Management / Import Backup file'. The current survey is 'naforma1'. The main content area features a large dashed box with the text: 'Click to select a Collect Backup (.collect-backup), Collect Data (.collect-data), Collect Earth XML data (.zip) file or drop it here.'

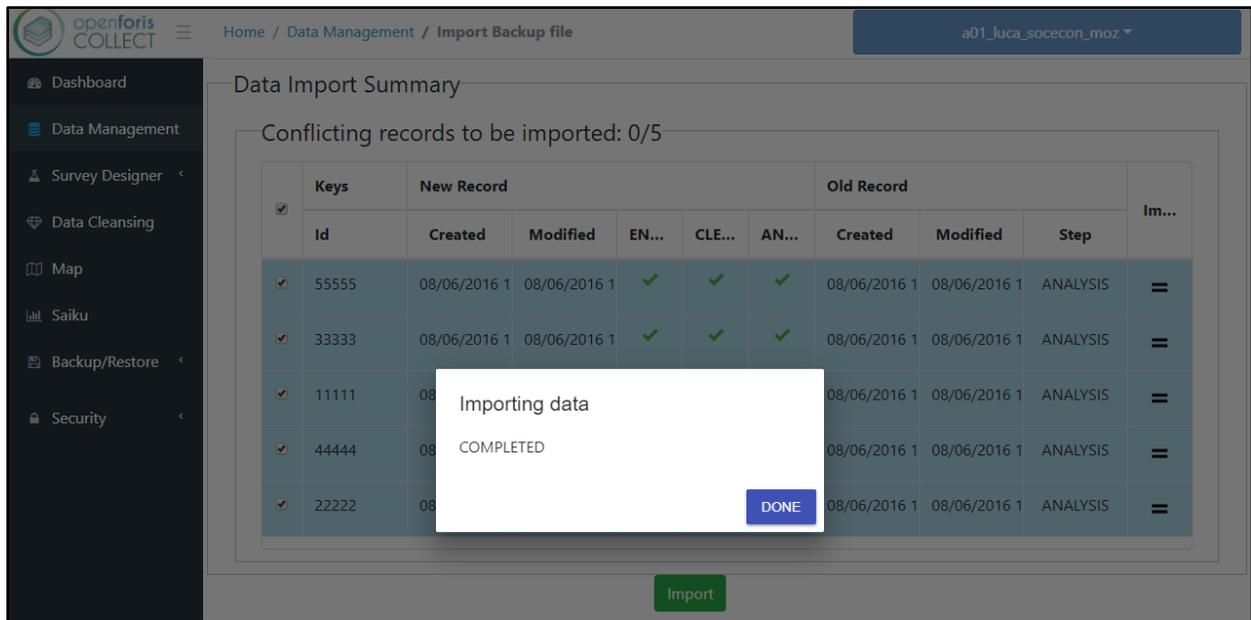
Click on the button Generate a complete summary, that will let you know if the data you are importing is compatible with that specific survey. If the data is not compatible with the survey we may need to load the data into a separate survey. Call the CTA if the data does not upload correctly.



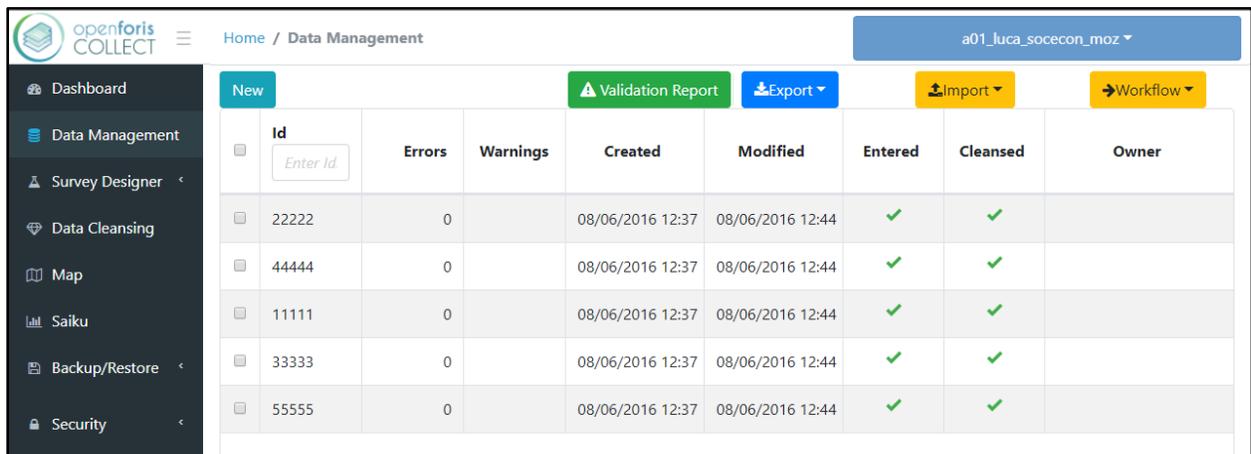
You will see the list of completed records with few metadata (Id, date of creation, step, etc.). From this list, you can select one, many, or all the records.



Once you selected the records you want to import, click on the button Import. About column Importability: if the records you are importing are newer than the existing records, importability will be green, otherwise red. If the records have exactly the same date and time labels, the equal symbol (=) is shown, and it's not necessary to import those records.



The screen will then display the current status, listing the records that have been added successfully to the database, as shown below. In case of Errors or Warnings, they will be indicated and can be dealt with during data cleansing.



Manual Data Entry

To add data manually, click on the New blue button at the top left of the screen. The user can start to enter data moving from field to field using 'Tab'. The behavior of each record field depends on its parameters (see examples in the chapter 'Schema').

As shown in the image below, Errors will be highlighted in red, warnings in yellow. If no value was present in the paper form, the data entry user can leave the field blank and specify the reason (by right-click on the field and choosing one of the options).

OPENFORIS COLLECT

FAO
FIAT PANIS

Back to list | Data Entry Cluster 2 | Submit

Cluster | Plot | Informant Interviews | Household Surveys | QA

Quality assurance field

Task	Person	Date
Form filled out	JOE	09 / 06 / 2014
Form checked	ADMIN	13 / 06 / 2014
Data entered	JANE	16 / 07 / 2014
Data cleaned	ADMIN	23 / 07 / 2014

Id 2

Measurement [Red box]

Region 008 Lindi

District 004 Liwale

Crew no. 4

Map sheet [Red box]

Accessibility [Red box]

Vehicle location

SRS

GPS Y

GPS X

GPS model

High precision GPS

Form Version: BP 28.12.2010 - SE 26.12.2010
Application version: 3.2.2-a2

Autosave Changes not saved Logged as: demo

Once all the fields have been filled-in for a record, the user can Submit it (top-right corner). The record will be added to the log and available for the next steps in the data work flow. In the present context manual data entry will be undertaken when users wish to correct errors made in the initial data entry.

openforis COLLECT Home / Data Management / Record a01_luca_socecon_moz

Data Entry change_it_to_your_sampling_unit [id]

Individual

ID	[id]
Age Range	1 31-45 y.o.
Activity	3 Honey production/collection
Genre	0 Male
Family Members	76
Location	
SRS	EPSG:4326
GPS Y	66.58
GPS X	87.65
Date Of Record	07 / 03 / 2018
Time Of Record	12 : 00

Confirm

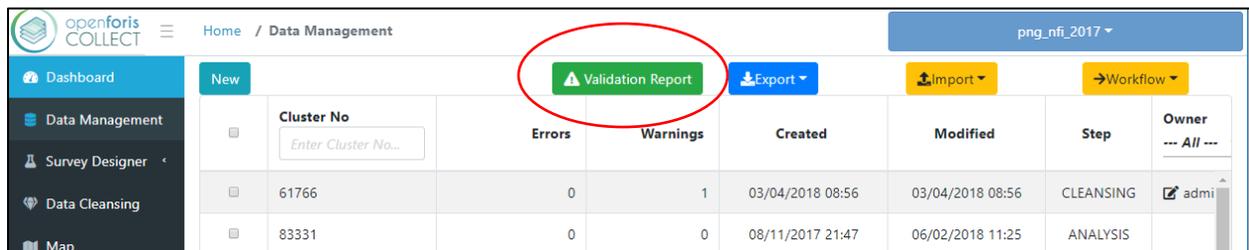
Submit this record for data cleansing?
No further changes will be allowed
in the entry phase.

Yes No

Step 1 – Validation report analysis

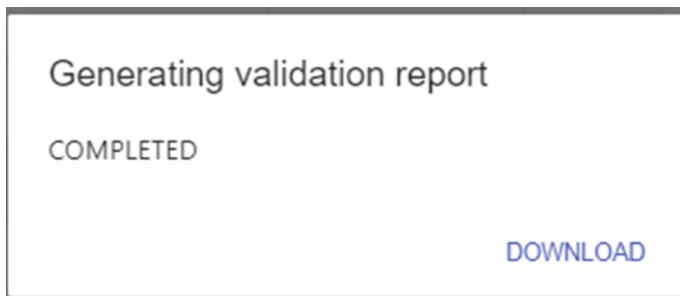
Effective use of the validation report

In the Data Management section of OpenForis Collect, you can run the validation report to analyze the nature of the errors displayed in the list of records. Click on the 'Validation Report' button, wait a moment and then click Download. Save the Validation Report to the same folder you are storing the error recording sheet. The result will be a .csv file with the details of the errors present in the database.



The screenshot shows the OpenForis Collect interface. The top navigation bar includes 'Home / Data Management' and a user profile 'png_nfi_2017'. The left sidebar contains 'Dashboard', 'Data Management', 'Survey Designer', 'Data Cleansing', and 'Map'. The main content area has a 'New' button and a table with columns: Cluster No, Errors, Warnings, Created, Modified, Step, and Owner. A green 'Validation Report' button is circled in red. Below the table, there are buttons for 'Export', 'Import', and 'Workflow'.

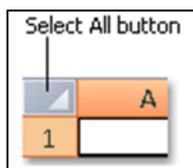
Cluster No	Errors	Warnings	Created	Modified	Step	Owner
<input type="text" value="Enter Cluster No..."/>						--- All ---
61766	0	1	03/04/2018 08:56	03/04/2018 08:56	CLEANSING	admi
83331	0	0	08/11/2017 21:47	06/02/2018 11:25	ANALYSIS	



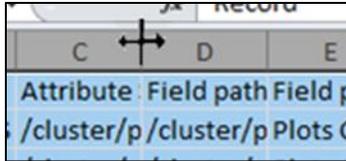
The validation report file is now saved to your default download folder.

Then open the .csv file with Excel. In Excel, first adjust the columns so that they are wider (or narrower) depending on the values in the column. The fastest way to do this is:

1. Select all in the Excel spreadsheet by clicking the Select All button, or by pressing CTRL + A,



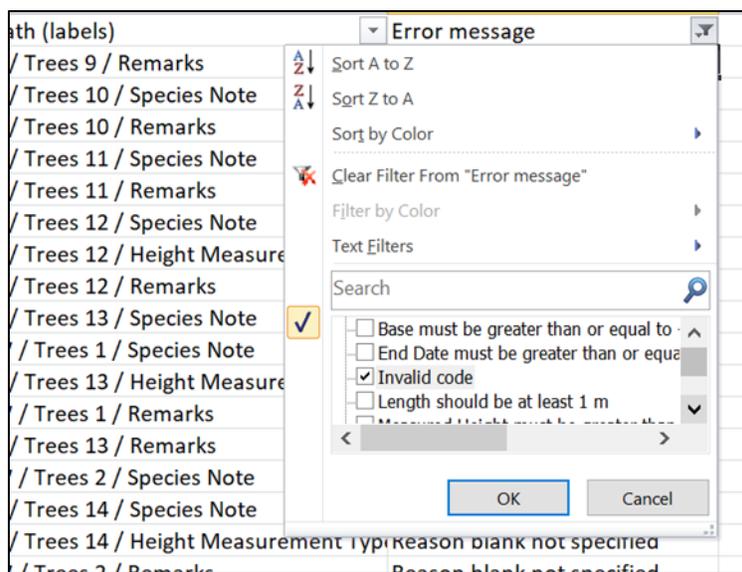
2. Double-click on a limit of a column header.



Then go to cell A2, and freeze the first row (View, freeze pane) so you can always see the header line. Then select the filter (Data, Filter).

Record	Phase	Attribute Schema Path	Field path	Field path (labels)	Error message
108924	ENTRY	/cluster/plot/tree/tree_remarks	/cluster/plot[3]/tree[9]/tree_remarks	Plots E / Trees 9 / Remarks	Reason blank not specified
108924	ENTRY	/cluster/plot/tree/species_note	/cluster/plot[3]/tree[10]/species_note	Plots E / Trees 10 / Species Note	Reason blank not specified
108924	ENTRY	/cluster/plot/tree/tree_remarks	/cluster/plot[3]/tree[10]/tree_remarks	Plots E / Trees 10 / Remarks	Reason blank not specified
108924	ENTRY	/cluster/plot/tree/species_note	/cluster/plot[3]/tree[11]/species_note	Plots E / Trees 11 / Species Note	Reason blank not specified

You can now filter your file using the "Error message" column. In this column, you will see error messages generated by your validation rules and some other checks that COLLECT performed by default.

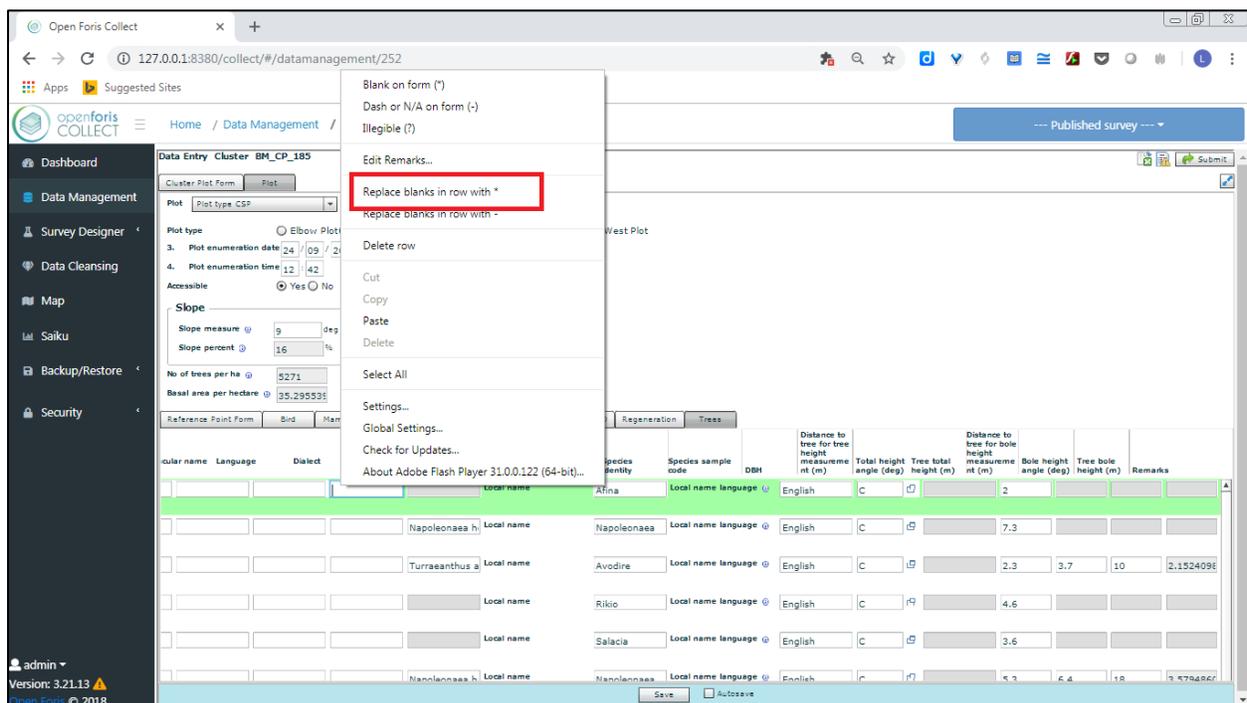


Please note that the plot and tree numbers in this report refer to the index (position) of the record in the database, and may be different from the actual attribute ('tree_no' value) entered in the database.

Field path	Field path (labels)
/cluster/plot[4]/tree[59]/tree_height_me	Plots C / Trees 59 / Height Measurement Typ
/cluster/plot[1]/tree[36]/tree_remarks	Plots W / Trees 36 / Remarks
/cluster/plot[3]/tree[24]/tree_remarks	Plots N / Trees 24 / Remarks

If your data was manually entered from paper forms using the COLLECT Desktop version, you may have several errors called "Reason blank not specified". To address this issue, see the tips at <http://www.openforis.org/support/questions/895/how-to-solve-error-reason-blank-not-specified>

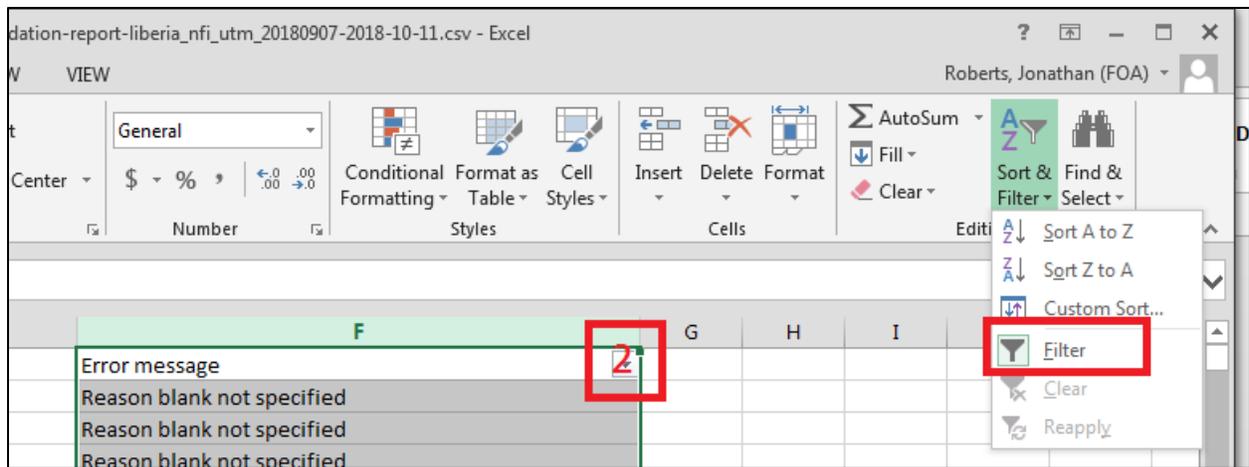
This particular error within the Liberian database is caused by enumerators not entering data into a field that specified data should be entered. Without the knowledge of why the value was left empty it is difficult for us to enter data that is relevant. In the present data cleaning exercise we will be filling the errors / omissions with either a * or a -. This correction can be achieved by right clicking in the empty entry slot and selecting “*”. See below



Errors of interest

The previous section outlined the process for correcting the “Reason blank not specified”. The following section will outline how the data cleaning officers are expected to deal with other errors identified during the validation process. Before you filter the error column first discuss with your colleagues which clusters each of you have been assigned and only work on those clusters within your area. Consider filtering the cluster column and selecting the clusters you are working on (see above). Filter to your clusters and then copy and paste the validation information into another sheet and work from that sheet.

In your validation error file, confirm that you have assigned a filter to the Error Message column. It should be column F. Highlight the column and select Sort & Filter, then select filter.



Once the filter is set we can now start looking at the types of errors present in the database. Below is a list of the types of errors we are likely to find. Reason blank not specified has been outlined above. Access times can be corrected / updated at a later date, we will make a list of these using error record sheet and ask the teams to update them.

Error message
Reason blank not specified
You have written an end access time in the future
You have written and end enumeration time in the future
stump diameter between 10 and 39.9 cm only in 7 m. subplot
dbh between 10 and 39.9 cm cm only in 7 m. subplot
Incomplete coordinate
The distance from the expected location is 995.7026481759515m but it must not exceed 60m
This field is required
dbh between 2 and 9.9 cm only in the smaller, 2 m. subplot
The distance from the expected location is 80022.26960861485m but it must not exceed 60m
The distance from the expected location is 499937.2237060204m but it must not exceed 60m
The distance from the expected location is 132.58279680303627m but it must not exceed 60m
The distance from the expected location is 1747.6822126139466m but it must not exceed 60m
You must specify at least 4 item(s)
The distance from the expected location is 120.92677148835058m but it must not exceed 60m
The distance from the expected location is 109.2352774530057m but it must not exceed 60m
The distance from the expected location is 104.21364799416175m but it must not exceed 60m
The distance from the expected location is 120.36197241487592m but it must not exceed 60m
The distance from the expected location is 120.8577103443011m but it must not exceed 60m

Carbon related errors

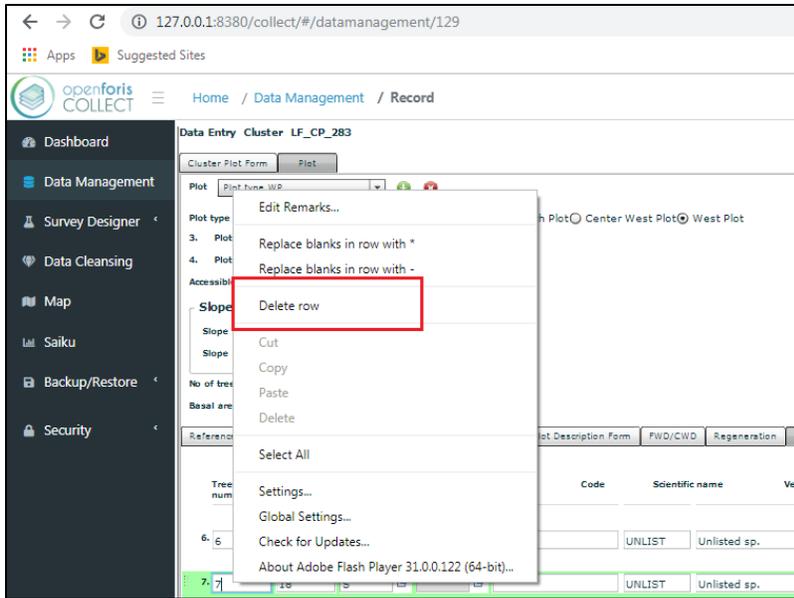
Errors that may have a negative impact on the estimation of carbon are of most importance to us at present. As such the first phase of the data cleaning will focus on correcting / updating these types of errors. In many cases trees have been recorded in the incorrect subplot, see these types of errors below

H	
Error message	
stump diameter between 10 and 39.9 cm only in 7 m. subplot	
dbh between 10 and 39.9 cm cm only in 7 m. subplot	
dbh between 2 and 9.9 cm only in the smaller, 2 m. subplot	

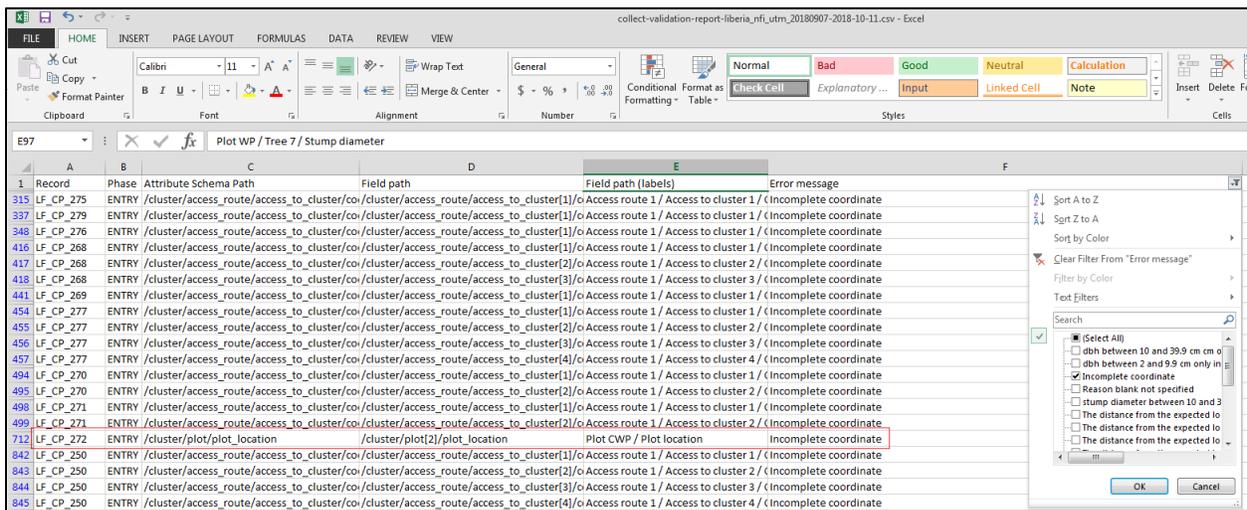
Use the data in the validation file to identify the relevant cluster, plot and tree. Filter for these errors in column F and observe the cluster number in column A, column E will then provide the plot type. The example below shows an error in cluster LF_CP_283, Plot WP tree 7.

Record	Phase	Attribute Schema Path	Field path	Field path (labels)	Error message
697	LF_CP_283	ENTRY /cluster/plot/tree/stump_diameter	/cluster/plot[1]/tree[7]/stump_diameter	Plot WP / Tree 7 / Stump diameter	stump diameter between 10 and 39.9 cm only in 7 m. subplot
698	LF_CP_283	ENTRY /cluster/plot/tree/tree_dbh	/cluster/plot[1]/tree[9]/tree_dbh	Plot WP / Tree 9 / DBH	dbh between 10 and 39.9 cm cm only in 7 m. subplot
711	LF_CP_272	ENTRY /cluster/plot/tree/tree_dbh	/cluster/plot[1]/tree[5]/tree_dbh	Plot WP / Tree 5 / DBH	dbh between 2 and 9.9 cm only in the smaller, 2 m. subplot

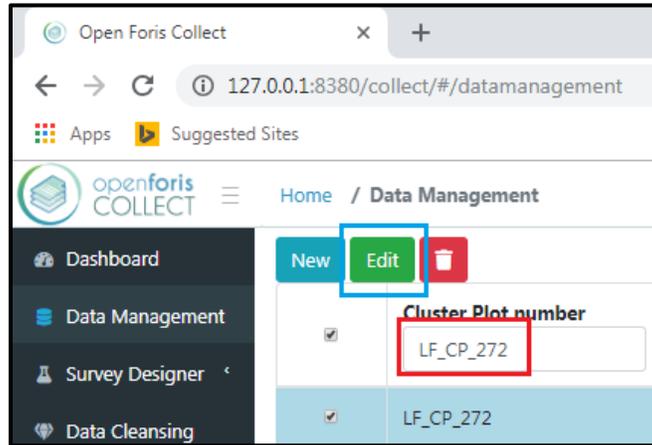
These types of errors need to be removed from the database. To do this navigate to the error in the online COLLECT installation using the data in the validation file. Find the tree (look in the trees tab). Confirm the error and delete the row. Deleting the row is very simple, select a row entry and right click, in the example below I have selected the tree number, applied a right click and will select the “Delete row” option. Don’t forget to save once you have corrected the error and removed the tree. Remove all trees that display this kind of error. Remember to make a note of the correction in the Carbon Related Errors and Corrections section. If you have removed a tree record the type of error, cluster number, plot name, tree number and the action taken, in this case the action taken is TREE DELETED. Repeat this process for all tree related errors in the validation report for the clusters you have been assigned. Once you have completed all the trees, move on to the next error type.



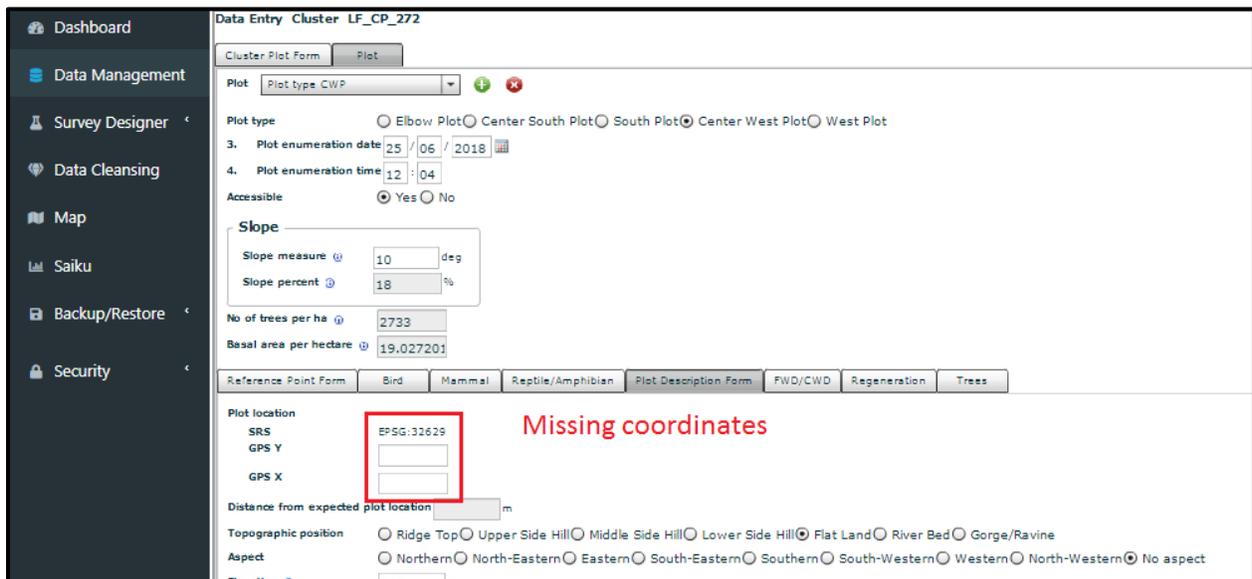
Another error which could affect the accurate estimation of Carbon content is the plot location as recorded by the field teams in the tablet. Filtering for these errors is relatively easy using the method described above. In this case we will be using the original coordinates provided to the teams to investigate why the errors or incomplete coordinates are present. You should have received a csv file with the original plot coordinates and cluster names to use as reference (LiberiaNFI_PlotLocationsCleaning_V1.csv). Using the filter option in your validation error file, filter for incomplete coordinates. You may notice that there are incomplete coordinates for both access routes and plots. For now we will focus on plot locations. Once the plot locations have been finalized and corrected we will then use the reports prepared by the teams to input the correct GPS coordinates for access points.



In the COLLECT database navigate to LF_CP_272. This can be done by typing in the cluster name in the Cluster Plot number box, see below. Select the cluster and click on the green edit button. The above table indicates to us that the error exists in the pot location coordinate for plot CWP.



In the present example the team failed to record the plot location. Make a note of the missing plot location including the cluster number, plot number, and the team who enumerated the plot. This information will be used to follow up with the relevant teams to identify if they captured the point using the handheld GPS. You will not be able to fix this error without the information from the team leader. Record this in error record sheet as COORDINATES TO BE UPDATED in the action taken column.



The other type of location error which may also negatively impact carbon estimates occurs when the plot location differs significantly from the GPS location inserted into the survey. The above error could be considered an omission error, while the errors below indicate incorrect data entry or some other issue.

Filtering for the errors we see that the distance, if above 60m is highlighted as an error. In the present analysis, these errors range from $\pm 100\text{m}$ to over 80,000m. All errors are of interest but the most important are obviously the plots with the highest error. Start with the plots that have the highest error, in the present case this is cluster GP_CP_223 which has an error of over 80,000m which is essentially saying that the plot is over 80kms from where it should be. This is obviously a serious error which needs additional investigation.

Record	Phase	Attribute Schema Path	Field path	Field path (labels)	Error message
346	GP_CP_226	ENTRY /cluster/plot/plot_location	/cluster/plot[3]/plot_location	Plot EP / Plot location	The distance from the expected location is 395.712698612959237m but it must not exceed 60m
1347	GP_CP_223	ENTRY /cluster/plot/plot_location	/cluster/plot[2]/plot_location	Plot CSP / Plot location	The distance from the expected location is 80022.26960861485m but it must not exceed 60m
1348	GP_CP_223	ENTRY /cluster/plot/plot_location	/cluster/plot[4]/plot_location	Plot CWP / Plot location	The distance from the expected location is 498927.3327060204m but it must not exceed 60m
1377	GP_CP_224	ENTRY /cluster/plot/plot_location	/cluster/plot[5]/plot_location	Plot CSP / Plot location	The distance from the expected location is 132.58279680303627m but it must not exceed 60m
1733	GP_CP_245	ENTRY /cluster/plot/plot_location	/cluster/plot[4]/plot_location	Plot CWP / Plot location	The distance from the expected location is 1747.6822126139466m but it must not exceed 60m
2859	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot[11]/plot_location	Plot EP / Plot location	The distance from the expected location is 120.92677148835058m but it must not exceed 60m
2860	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot[2]/plot_location	Plot CSP / Plot location	The distance from the expected location is 109.2352774530057m but it must not exceed 60m
2861	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot[3]/plot_location	Plot SP / Plot location	The distance from the expected location is 104.21364799416175m but it must not exceed 60m
2862	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot[4]/plot_location	Plot CWP / Plot location	The distance from the expected location is 120.36197241487592m but it must not exceed 60m
2863	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot[5]/plot_location	Plot WP / Plot location	The distance from the expected location is 120.8577103443011m but it must not exceed 60m

Navigate to the cluster in COLLECT and select Plot CSP. See below, you will notice that the error is indeed large.

Slope

Slope measure: 10 deg
 Slope percent: 18 %
 No of trees per ha: 1762
 Basal area per hectare: 19.693598

Reference Point Form: Bird, Mammal, Reptile/Amphibian, **Plot Description Form**, FWD/CWD, Regeneration, Trees

Plot location

SRS: EPSG:32629
 GPS Y: 882123
 GPS X: 392761
 Distance from expected plot location: 80022.26 m
 Topographic position: Ridge Top, Upper Side Hill, Middle Side Hill, Lower Side Hill, Flat Land, River Bed, Gorge/Ravine
 Aspect: Northern, North-Eastern, Eastern, South-Eastern, Southern, South-Western, Western, North-Western, No asp
 Elevation: 201 m

The field teams should have identified this error in the field and corrected for it. There are two options here. We can first look at the coordinates for the other plots to see if they are also erroneous. Both CSP and CWP have major errors in their locations. If we look at plot SP, we see that the X and Y coordinates entered have an error of only 8.9m. If we compare them to CSP and CWP we see a pattern in the errors.

SP
✓

CWP
✗

CSP
✗

Plot location	Plot location	Plot location
SRS EPSSG:32629	SRS EPSSG:32629	SRS EPSSG:32629
GPS Y 802062	GPS Y 802176	GPS Y 882123
GPS X 392764	GPS X 892705	GPS X 392761
Distance from expected plot location 8.997444m	Distance from expected plot location 499937.2m	Distance from expected plot location 80022.26m

The coordinate for SP is clearly correct, compare this coordinate pair CWP. For CWP we see that the GPS Y coordinate appears to match SP, however, the GPS X coordinate is slightly different. Looking carefully we can see that for CWP GPS X coordinate the only significant difference is that the first number is 8 while for SP GPS X this value is 3. It appears then that booker, or data manager entered the GPS X coordinate incorrectly. Replace the 8 with a 3 in CWP and save the cluster. Making this change you will notice that the error is now reduced to 9.6m which is completely within our margin of error.

Plot location
SRS EPSSG:32629
GPS Y 802176
GPS X 392705
Distance from expected plot location 9.619751m

Use the same method for the rest of the location errors in your data set. The errors may be more subtle so be careful with how you update the coordinates. If the error gets bigger after an update then the change should be reversed. Remember to note the cluster, plot and type of change applied in the Error Recording sheet. Keeping good records of the changes made will help to understand how to improve the data collection activities. We have only covered a couple of errors in this sheet but throughout the error update process the data cleaning team is encouraged to interpret the errors and to investigate the cause of these. If changes are made they should be recorded in the error record sheet. Once all the errors have been addressed and captured please save the database and continue to the next section.

Record	Phase	Attribute Schema Path	Field path	Field path (labels)	Error message
1347	GP_CP_223	ENTRY /cluster/plot/plot_location	/cluster/plot3/plot_location	Plot EP / Plot location	The distance from the expected location is 995.7026481759515m but it must not exceed 60m
1348	GP_CP_223	ENTRY /cluster/plot/plot_location	/cluster/plot2/plot_location	Plot CSP / Plot location	The distance from the expected location is 80022.26960861485m but it must not exceed 60m
1377	GP_CP_224	ENTRY /cluster/plot/plot_location	/cluster/plot4/plot_location	Plot CWP / Plot location	The distance from the expected location is 499937.2237060204m but it must not exceed 60m
1733	GP_CP_245	ENTRY /cluster/plot/plot_location	/cluster/plot5/plot_location	Plot CSP / Plot location	The distance from the expected location is 132.58279680303627m but it must not exceed 60m
2859	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot4/plot_location	Plot CWP / Plot location	The distance from the expected location is 1747.6822126139466m but it must not exceed 60m
2860	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot1/plot_location	Plot EP / Plot location	The distance from the expected location is 120.92677148830058m but it must not exceed 60m
2861	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot2/plot_location	Plot CSP / Plot location	The distance from the expected location is 109.23527745930057m but it must not exceed 60m
2862	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot3/plot_location	Plot SP / Plot location	The distance from the expected location is 104.21364799416175m but it must not exceed 60m
2863	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot4/plot_location	Plot CWP / Plot location	The distance from the expected location is 120.36197241487592m but it must not exceed 60m
2863	BM_CP_151	ENTRY /cluster/plot/plot_location	/cluster/plot5/plot_location	Plot WP / Plot location	The distance from the expected location is 120.857103443011m but it must not exceed 60m

Step 2 – Additional survey assessment

Access route and photo assessment

All surveys should contain information regarding the route field officers took to access a plot, the COLLECT mobile survey form provides tools to record the geographic location of points of interest along the way and to also record a photo name of the feature of interest, teams were instructed to record the photos and their names. This information is useful for teams who will return to the field to monitor the plots as part of the MRV activities in Liberia. During this assessment we will run through all of the clusters and check to see that teams have recorded access route GPS coordinates as well as photo names. The photo names will be compared to the photos provided by the teams to make sure that these match and can be used for visual confirmation of routes taken. .

Each data cleaning analyst will receive a list of Clusters to assess. In the example below we will use survey information from Sinoe County. The steps for assessing access route coordinates and photo names will be the same for all clusters. In the following assessment we will review the presence of access route points, the coordinates associated with these points and photo file names listed in the survey. The information will be compared to the photo data submitted to the field team data managers. The purpose of this exercise is to confirm that teams have captured access points (geographic coordinates) as well as photos at these points. We will also be checking for permanent structure points as well as reference points in the plot data.

To proceed with the cleaning and analysis, in COLLECT select data management, then select the survey you are using from the drop down menu in the top right. In the example below we will be assessing SN_CP_024, select this cluster and you should see an edit button available. Click on the edit button and you should see the imported data for the selected cluster. We will now review the cluster access coordinates as well as the photo name records. The photo name records will also be compared to the actual names given to the photo file names. We will begin reviewing the Access route data.

Cluster Plot number	Errors	Warnings	Created	Modified	Step
NB_CP_099	0	0	27/11/2018 14:38	21/12/2018 11:44	CLEANSING
SN_CP_024	5	6	27/11/2018 14:38	21/12/2018 10:56	CLEANSING
GK_CP_003	2	0	15/01/2010 04:44	21/12/2018 10:14	CLEANSING
GG_CP_120	3	0	12/12/2018 16:32	21/12/2018 09:53	CLEANSING

Once you have opened the cluster you are reviewing you should see the Cluster Plot Form followed by the cluster number, start access date and time and then the Access route information table. The figure below shows that for this specific cluster two access route points were captured. One for Weatuzon Town and one for Dugbeh River. Next check that the GPS coordinates have been inserted. Make sure that both GPS X and GPS Y are present for both points. When checking the coordinates themselves, make sure that the range of values for GPS X are between 223770.83 and 681168.56 while for GPS Y are between 481258.56 and 945603.90. If the values stored in the database are outside these ranges then please make a note of this error in the data cleaning error sheet. Do not make any changes yet. We are now only seeking to identify the presence or absence of this information. Once you have checked the coordinates we will now compare the access photo code with the photos captured and shared by the teams.

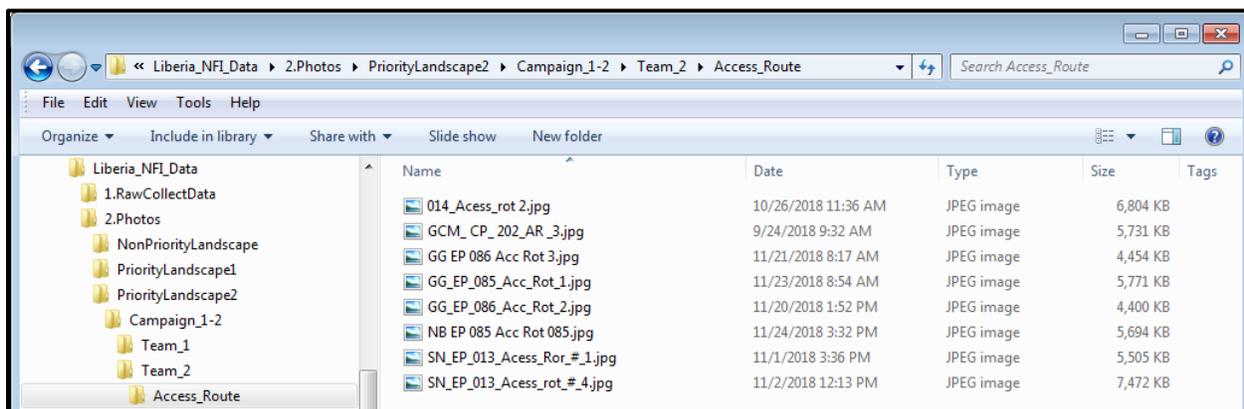
The screenshot shows the 'Data Cleansing Cluster SN_CP_024' form in the Open Foris Collect application. The 'Access route' section is highlighted with a red box and contains the following table:

Description	Coordinates SRS	GPS Y	GPS X	Access photo code	Bearing
Weatuzon Tow	EPSG:32629	554626	547133	SN_024/EP/Acc	149
Dugbeh River	EPSG:32629	555348	546701	SN_024/Ep/Acc	329

Other form fields include: Cluster Plot number (SN_CP_02), Start access date (28/10/2018), Start access time (09:18), End access date (28/10/2018), End access time (09:18), Team number (radio buttons 1-6, with 2 selected), Team Leader (Anthony J. Koig), and Cluster Plot description (Young secondary formation with fallow across).

The current cluster plot is located in Sinoe, we know this from the cluster name SN_CP_024, and we know that the Sinoe clusters were enumerated during the first campaign in the South East / Priority Landscape 2. Observing the additional information provided in the Team table we can see that this cluster was enumerated by Team 2. We will use this information to locate the photos in the data folders which have been shared with you at the start of the data cleaning. We know that data from campaign 2 in Priority

Landscape 2 were combined with campaign 1. In the data folder navigate to the Photos folder, select PriorityLandscape2, Campaign_1-2 and finally Team_2. In the folder you should see two sub-folders named Access_Route and ReferencePointPhotos. For now we are interested in Access_Route only. Double click on this folder. In our present example there are 8 photos in the folder. Photo naming conventions have not always been adhered to in the present inventory so you will need to use common sense to match the photos with the names recorded in the survey. The names in the survey are - SN_024/EP/Accs. Rt/ph_1 and SN_024/EP/Accs. Rt/ph_2. We can see below that none of the images in the Team 2 folder match or indicate they were captured at this cluster. Here we have a case of missing photos. If this is the case for your cluster, make sure to note the missing photos in the data cleaning error recording sheet, be sure to capture the cluster name and the number of missing photos. If you find a photo whose name is incorrect in the database replace the name in the database with the photo name in the respective photo folder. Make a note of this change in data cleaning error recording sheet. Repeat this process for all of the access route points in the clusters assigned to you. Be sure to accurately record the errors and or changes made to the database in your error recording sheet. We will now move on to the Reference Photos.

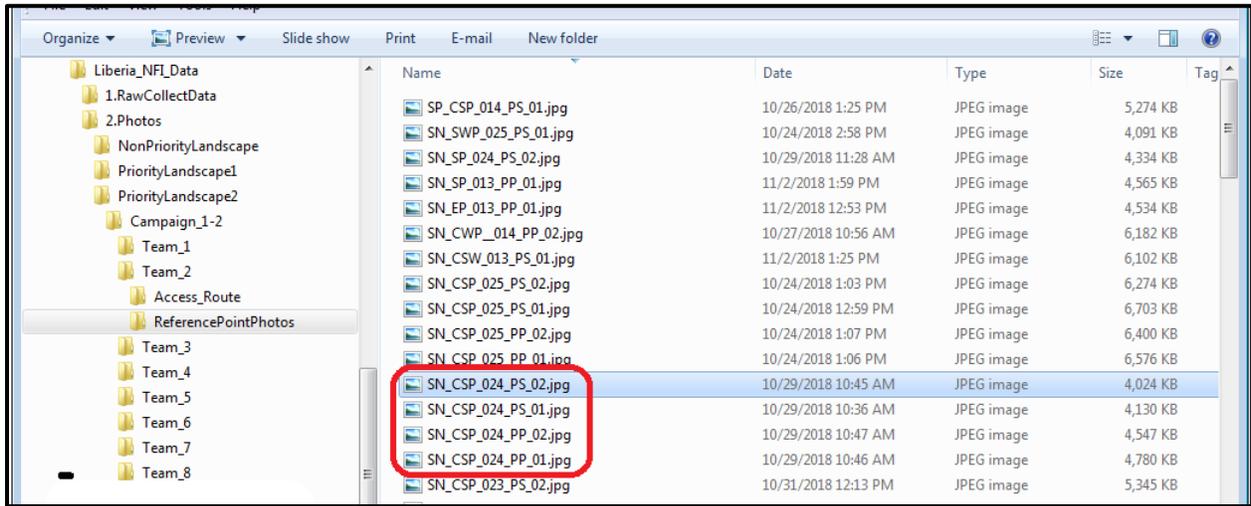


Reference photos and prominent structures

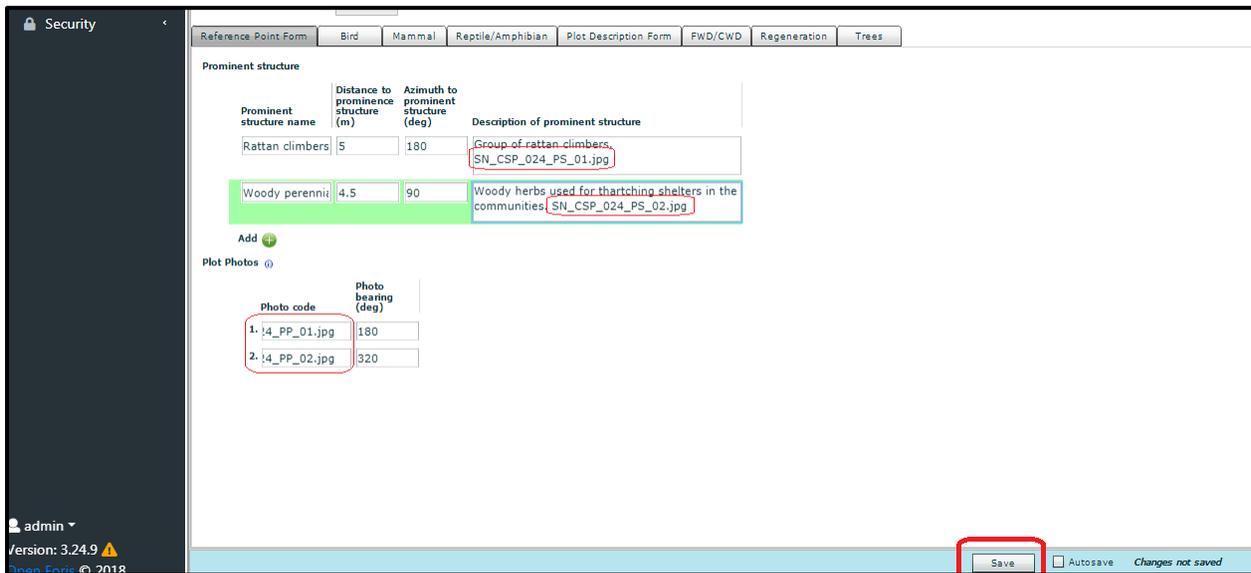
Reference photos were captured as part of the Plot data. As such select a plot from the plot tab, in this example we will be using Plot Type EP for the same SN_CP_024. Select the EP plot and then select the Reference Point Form. You will see that there are two prominent structures along with two plot photos. The prominent structures do not have an input location for the photo name, we will make use of the Description of prominent structure input dialogue to store the photo file name. We will also compare the photo code in the add photos section to the photo name stored in the photo database.

Navigate to the folder which contains the ReferencePointPhotos for the campaign one and two of Priority Landscape area 2. There are exactly 100 photos (this will vary depending on the team and campaign) in Team 2's folder, we now need to identify the photos captured for both the prominent structure and the plot photos. Sort the image files according to name and scroll down till you find the images with names that match the cluster you are assessing. In the present case we are looking for files with names that include SN_CP_024, we note that there are four files which have the relevant file naming convention with additional references to PP (plot photo) and PS (prominent structure). It should be noted that teams may have used different naming conventions, you will need to use common sense and or prior knowledge of the naming conventions to identify the relevant photos. The figure below highlights the four photos which team 2 captured at the elbow plot of SN_CP_024. The image files with the suffix PP refer to the plot photos recorded in the survey while the PF refer to the prominent structures. The PF images will need to be

visually examined to see which photo refers to the Rattan Climbers and which refers to the Woody perennial.



Once you have determined the correct names for the prominent structures, copy the file name of the respective image into the “Description of prominent structure” input, do this for all prominent structures. Now do the same for the Plot Photos, check if the image file has a different name to that which has been captured in the survey, if this is the case update the survey with the correct image file name.



Once you have made the changes do not forget to save the updates. Repeat this process for all clusters assigned to you. If you see that there are missing prominent structures in the survey (i.e. There is a photo captured but no record in the survey make a note of the omission in the error recording sheet. Please review each plot in each cluster assigned to you.

Land use assessment

The plot description form of the biophysical survey captured land use information that is vital for the accurate analysis of the inventory data. This section of the SoP will review the assessment of the land use information captured as part of the stand description. Once again the manual will make use of the SN_CP_024 cluster, select the plot type EP, and select the tab labelled Plot Description Form and scroll down to the section titled Stand Description. Check that information has been captured for the Land Use class, Land Use subdivision, Successional status and finally forest type. If this information is missing make a note of the omission in the error recording sheet provided, taking special care to capture the correct cluster and plot type in the sheet along with the type of omission identified.

Repeat this process for all plots and clusters assigned to you making a note of any and all omissions identified.

Harmonize Non-timber Forest Products

Non-timber forest products are an important source of food and income for communities in Liberia, the NFI survey provided teams with an opportunity to record non-timber forest products (NTFPs) present at each of the clusters sub-plots. Unfortunately, at the start of the inventory a concise list of NTFPs was not available and as such the field teams were asked to input the NTFPs using free text. The following section will explain how data cleaning officers will harmonize these inputs making use of a lookup table prepared to support this activity. Once the data has been harmonized it will be easier to run data analysis on the information. We will use an example from the lookup table to explain why this is necessary. The figure

below shows a screen shot from the lookup table created. You will see that because teams were free to enter the names of the NTFPs, they sometimes used incorrect spelling, this is expected as in the field and indeed has been planned for. The table provided to data cleaning teams contains a column named **Survey NTFP** as well as a column called **Harmonized NTFP**. Each of the entries in the database are included in **Survey NTFP** column, in the **Harmonized NTFP** column we see what the values in the database should look like. In the example below we see that there are a number of incorrect spellings in the database for *Aframomum melegueta*, as well as the correct spelling in the **Harmonized NTFP** column. In the present data cleaning exercise you will be expected to navigate to the Forest Resources section of each cluster and sub-plot assigned to you and to update the NTFP name recorded using the **Harmonized NTFP** name provided in the excel table. A practical example is given below.

Number	Survey NTFP	Harmonized NTFP	NTFP Use
1	Abura	Abura	Folder, Medicinal
2	Aframomum melegueta		
3	Aframomum melegueta		
4	Aframomum melegueta.		
5	Aframomun melegueta	Aframomum melegueta	For infection, stomach, menstruation, and food fruit
6	Aframomun melegueta		
7	Aframomun melegueta		
8	Aframomum malaguitta		
9	Aframomum maliguetta		
10	Aframomum mekgueta		
11	African walnut	African walnut	Fruit for food
12	Alchonea codifolia		
13	Alchonia		
14	Alchonia chlodipholia		
15	Alchonia cordifonia	Alchonea cordifolia	Medicinal
16	Alchordea cordifolia		
17	Alchornea Cordifolia		
18	Alchornea cordifonia		
19	allanblackia	Allanblackia	used for cosmetics and medicals
20	Annika polycarpa tree	Annika polycarpa	Traditional medicine to cured or prevent Yellow fever and malaria.

The example presented below is taken from the Cluster GCM_CP_166. Open COLLECT and navigate to data management and select the Liberia_nfi_utm_20180907 survey. In cluster plot number, insert GCM_CP_166. Select this cluster and select the edit button to open the survey. Select the plot tab, once it is loaded select the Plot Description Form. At this point you should see the Forest Resources section title with a sub title called Non Timber Forest Products (NTFP). Directly below this you will see a table with two columns, one titled NTFP name and the other NTFP use. We will focus on updating the NTFP name (red box) with the harmonized name from the look up table.

SRS: EPSG:32629
 GPS Y: 751176
 GPS X: 264051
 Distance from expected plot location: 1,840503m

Topographic position: Ridge Top Upper Side Hill Middle Side Hill Lower Side Hill Flat Land River Bed Gorge/Ravine

Aspect: Northern North-Eastern Eastern South-Eastern Southern South-Western Western North-Western No aspect

Elevation: 17 m

Forest Resources

Non Timber Forest Products (NTFP)

NTFP name	NTFP use
Palm tree	Food and Medic
Palisota hirsuta	Medicinal
Mareya micranth	Medicinal
Costus dubius	Medicinal
Xylopia evensii	Medicinal
Tetracera affinis	Medicinal

Snag: 1-5 snags 5-10 snags More than 10 No Snags
 Fallen Trees: 1-5 Trees 5-10 Trees More than 10 No fallen tree

The graphic above indicates that for GCM_CP_166 Elbow plot, the team recorded six NTFP present. We will now go through the process of updating each of these NTFP names with the harmonized name from the lookup table shared with you. The first entry in the list is Palm tree, open NTFP_LookupTable.xlsx and scroll down or search for the term Palm tree, you should find it at number 99. You will see that the harmonized name for all NTFPs labelled Palm tree is *Elaeis guineensis*.

NTFP_LookupTable.xlsx - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Normal Page Break Preview Page Custom Layout Views Ruler Formula Bar Gridlines Headings Zoom 100% Zoom to Selection New Window Arrange All Freeze Panes Hide Split View Side by Side Synchronous Scrolling Reset Window Position Switch Windows Macros

B100 : X ✓ fx Plan tree

	A	B	C	D
1	Number	Survey NTFP	Harmonized NTFP	NTFP Use
89	88	Eleais guineensis		
90	89	Eliaes guineansis	Elaeis guineensis	Oil palm production, construction, palm butter, furniture
91	90	Eliaes guineensis		
92	91	Eliaes guineensis		
93	92	Eliaes guineensis		
94	93	Palm		
95	94	palm tree		
96	95	palm trees		
97	96	palm trees, coffee africana		
98	97	Palms		
99	98	Plam		
100	99	Plam tree		
101	100	Plam trees		
102	101	Eliaes guineensis		
103	102	Eliaes guineensis		

Replace Palm tree in the survey with *Elaeis guineensis*. Now repeat the process for all NTFPs present. For the present example we will not actually make any changes to the database, rather the example is here to guide your own cleaning activities. It may be the case that the data inserted into the database is in fact correct and does not require updating. You have the opportunity to check each of the NTFPs listed against the lookup table provided. It is very important that all NTFPs have names that are common with those **Harmonized NTFPs** in the excel spreadsheet. In the example below we have updated the first and the last

NTFP name using the harmonized names from the spreadsheet. Once you have completed your own updates do not forget to update and save the data.

Reference Point Form | Bird | Mammal | Reptile/Amphibian | Plot Description Form | FWD/CWD | Regeneration | Trees

Elevation ⓘ m

Forest Resources

Non Timber Forest Products (NTFP)

NTFP name	NTFP use
<input type="text" value="Elaeis guineens"/>	<input type="text" value="Food and Medic"/>
<input type="text" value="Palisota hirsuta"/>	<input type="text" value="Medicinal"/>
<input type="text" value="Mareya micrant"/>	<input type="text" value="Medicinal"/>
<input type="text" value="Costus dubius"/>	<input type="text" value="Medicinal"/>
<input type="text" value="Xylopia evensii"/>	<input type="text" value="Medicinal"/>
<input type="text" value="Tetracera affinis"/>	<input type="text" value="Medicinal"/>

Snag 1-5 snags 5-10 snags More than 10 No Snags

Fallen Trees ⓘ 1-5 Trees 5-10 Trees More than 10 No fallen tree

Cycle through all of the clusters assigned to you and make sure to check and or update the NTFPs using the spreadsheet provided. If you come across any NTFPs that don't have a harmonized name, please make a note in the relevant error recording sheet. We will follow up with these errors at a later stage.

Exporting cleaned data

Once the online database has been cleaned and everyone has completed their allotted clusters, it will be necessary to download the data to your local machine. To do this make sure you have saved all of your changes to the database. Return to Data Management, make sure you are in the correct survey (you will only have access to one), click on Export and select CSV/Excel. Once prepared the software will give you the option to save a *.zip file similar to this one (collect-csv-data-export-liberia_nfi_utm_20180907-2018-10-12T10_33_31.zip). Save it to your local drive. We will be using this data for the remainder of the data cleaning activities.

Open Foris Collect | 127.0.0.1:8380/collect/#/datamanagement

Home / Data Management | liberia_nfi_utm_20180907

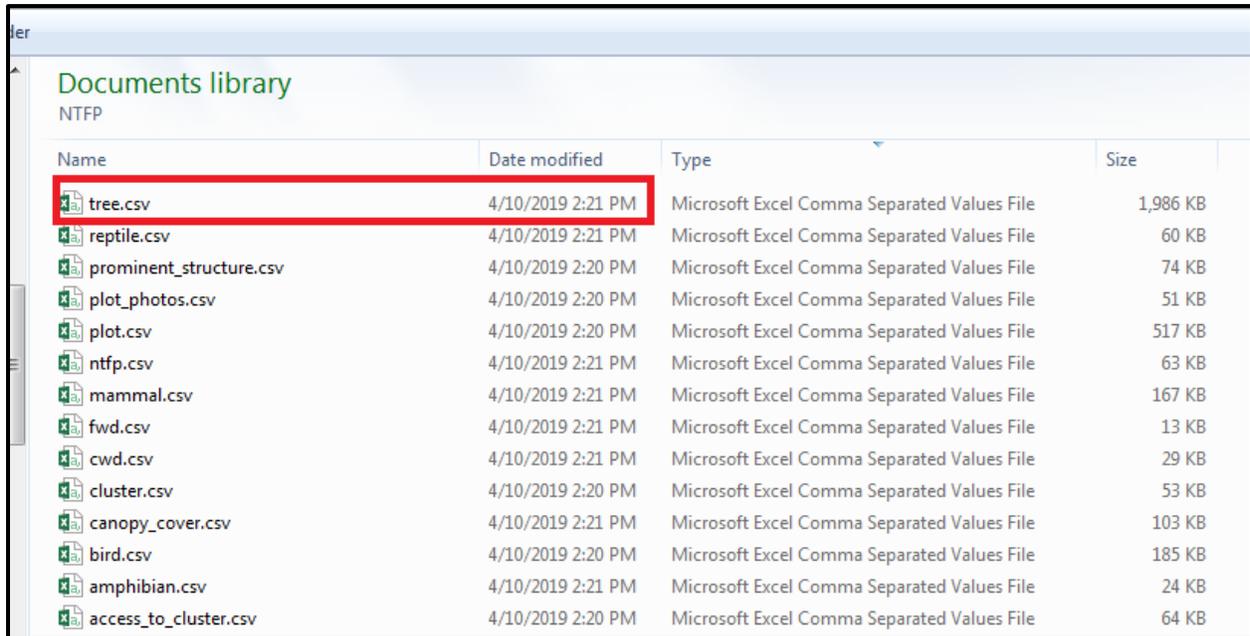
[New](#)
[Validation Report](#)
[Export](#)
[Import](#)
[Workflow](#)

Cluster Plot number	Errors	Warnings	Validated	Step	Owner
<input type="text" value="Enter Cluster Plot number"/>					All
GP_CP_223	10	1	11/09/2018 16:00	11/10/2018 16:59	ENTRY admin
LF_CP_282	72	1	11/09/2018 11:17	11/10/2018 11:53	ENTRY admin
BM_CP_185	0	0	08/10/2018 16:25	08/10/2018 16:25	ENTRY admin
GCM_CP_165	0	0	08/10/2018 16:25	08/10/2018 16:25	ENTRY admin

Step 3 – Tree Species Assessment

Data preparation

Step 3 of the data cleaning exercise will make use of the data you exported from the previous exercise. Navigate to the folder where you stored the exported data and unzip this folder. You should see something similar to image below. The tree species assessment will make use of the **tree.csv** file.



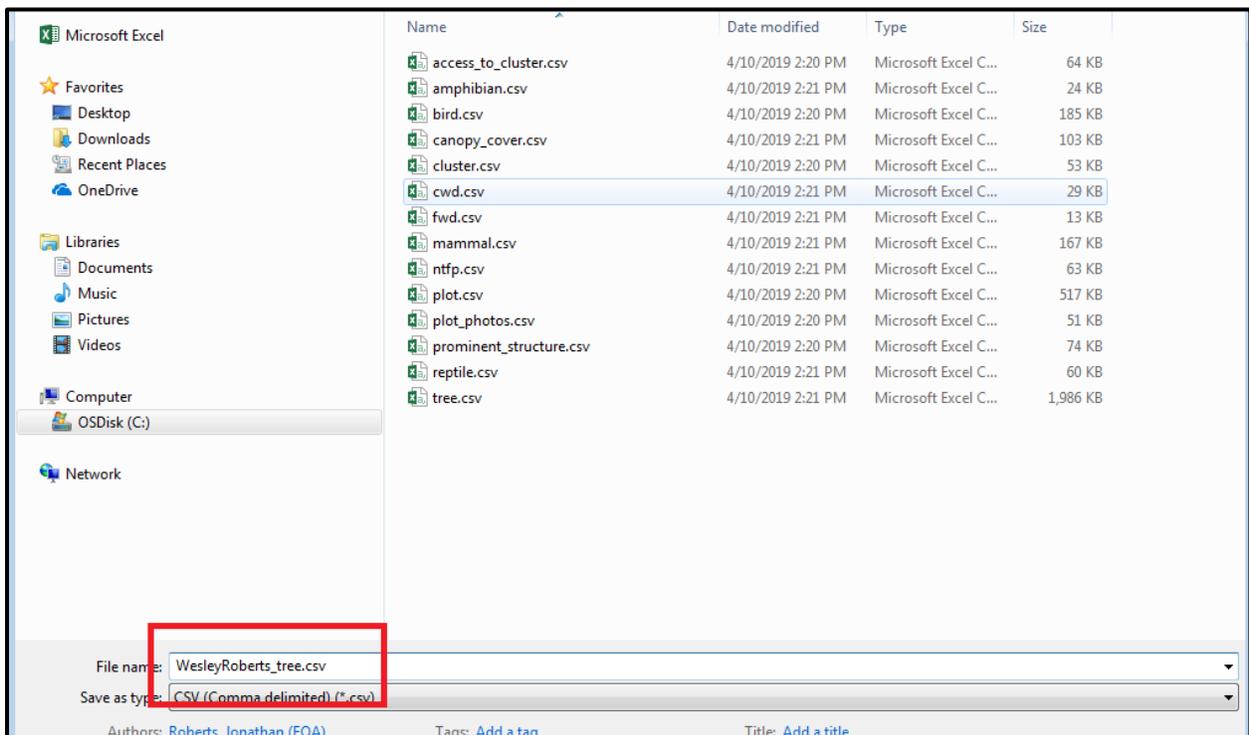
Name	Date modified	Type	Size
tree.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	1,986 KB
reptile.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	60 KB
prominent_structure.csv	4/10/2019 2:20 PM	Microsoft Excel Comma Separated Values File	74 KB
plot_photos.csv	4/10/2019 2:20 PM	Microsoft Excel Comma Separated Values File	51 KB
plot.csv	4/10/2019 2:20 PM	Microsoft Excel Comma Separated Values File	517 KB
ntfp.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	63 KB
mammal.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	167 KB
fwd.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	13 KB
cwd.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	29 KB
cluster.csv	4/10/2019 2:20 PM	Microsoft Excel Comma Separated Values File	53 KB
canopy_cover.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	103 KB
bird.csv	4/10/2019 2:20 PM	Microsoft Excel Comma Separated Values File	185 KB
amphibian.csv	4/10/2019 2:21 PM	Microsoft Excel Comma Separated Values File	24 KB
access_to_cluster.csv	4/10/2019 2:20 PM	Microsoft Excel Comma Separated Values File	64 KB

Although the survey already contains a long list of official scientific names for tree (and animal) species, the crews during the survey are given two possibilities that do not include scientific names:

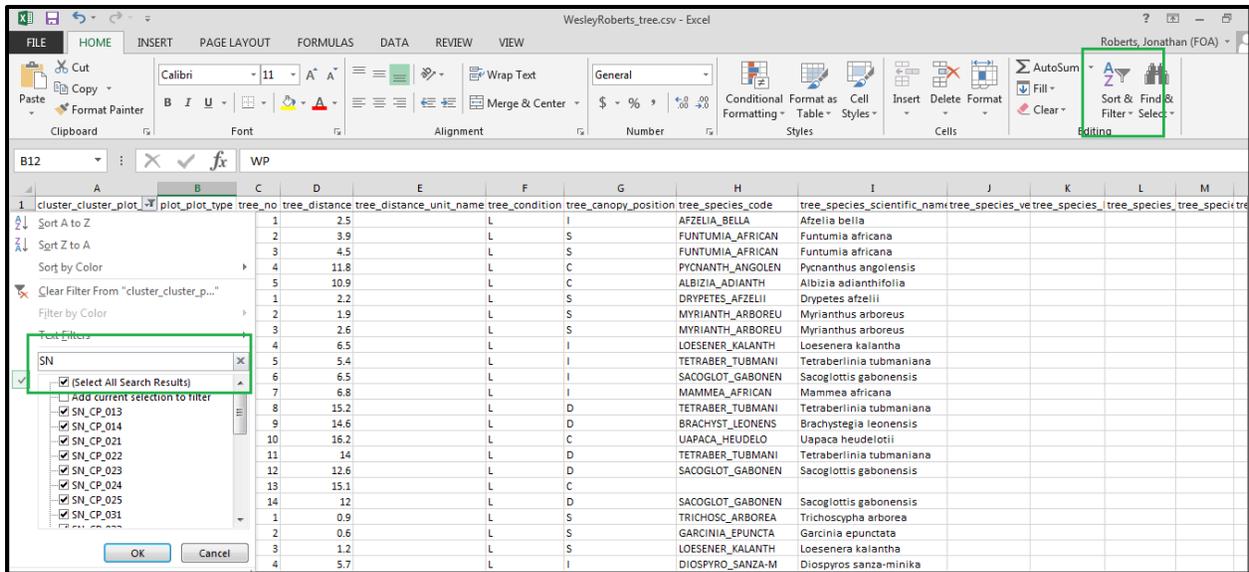
- *Unknown sp.* Reserved in principle for those species that are unknown, which in theory should have been collected (as leaves, flowers, fruits... or other components that might help recognize the plant), put into a bag, codified under the column **species_sample** and sent to the laboratory for identification.
- *Unlisted sp.* Reserved for species that might not have been included in the initial survey list of Liberian species but are known to the botanist. In this case, while the code in the column **tree_species_scientific_name** should be *Unlisted_sp.*, the proposed known scientific name should be included under the column **unlisted_tree_species** and, if the local name is known, it should have been included under the column **local_name_local_name** in the exported csv file (this could be tree.csv, amphibian.csv, bird.csv, mammal.csv, ntfp.csv, or reptile.csv).'

Yet, the current database contains also *Blank cells*, given the fact that the variable **tree_species_scientific_name** was optional.

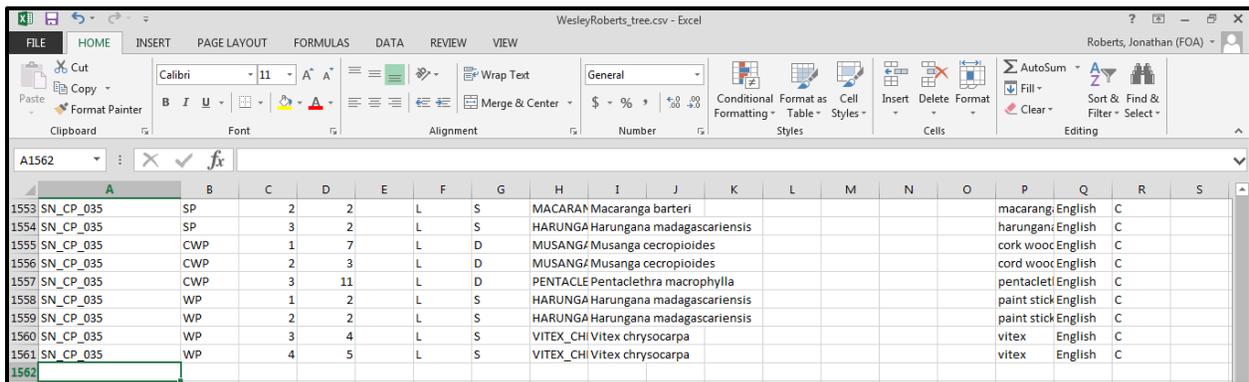
The following cleaning exercise will involve interrogating the unlisted species using online tools and updating the tree.csv file. The tree.csv file that currently resides on your hard drive includes all trees from all clusters in the data set. Each of you have been assigned a specific set of clusters to clean and correct. The first step in the species assessment process is to filter for your specific list of clusters. Open tree.csv and immediately re-save it using the following naming convention <youname>.tree.csv. We will now do all of our work in the new file.



In your new file add a new sheet and call it “corrected_tree”. Go back to the previous sheet, we will now select only the clusters that have been assigned to you for this exercise. This will be done by preparing a filter and copying and pasting the selected data into the corrected_tree sheet. Click on column A to select the column, column A contains the cluster number, we will use this column to select our clusters of interest. With the column selected, click on Sort & Filter on the right (see below). In the example below I have selected all cluster from Sinoe County using the first two letters from the Sinoe counties clusters, “SN” in the search box. You may have clusters from multiple counties, if so you will need to make multiple selections to include all clusters. It is really important that you select only those clusters which are assigned to you, we do not want there to be any omissions and would like to avoid duplication of efforts.

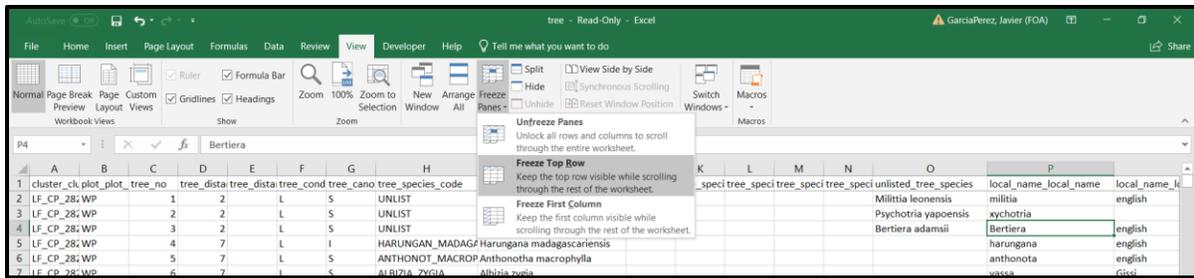


Once you are sure that you have selected all clusters that are assigned to you, please copy and paste the selection to your tree_corrected sheet. Once complete, turn off the filter and save your file. Switch to the corrected_tree sheet and scroll down to the bottom of your data. An unfiltered tree.csv file should have more than 10,000 entries, check that you have less than 10,000 entries. In the example below Sinoe has around 1,561 trees. Yours will certainly differ. We now have our assessment data sets prepared, before we continue it is important to look at the potential issues.

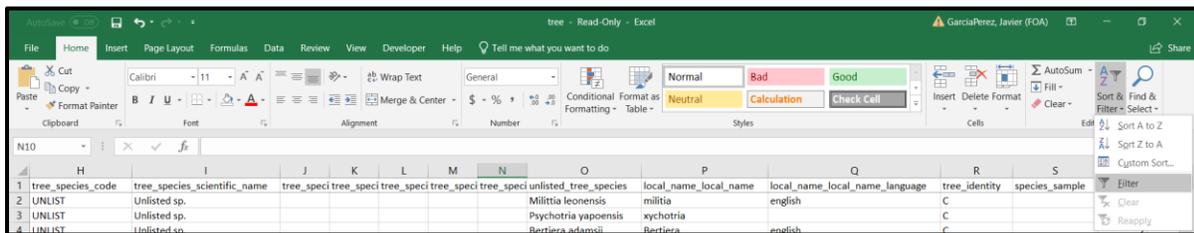


In order to see the potential issues with these different options, the steps would be:

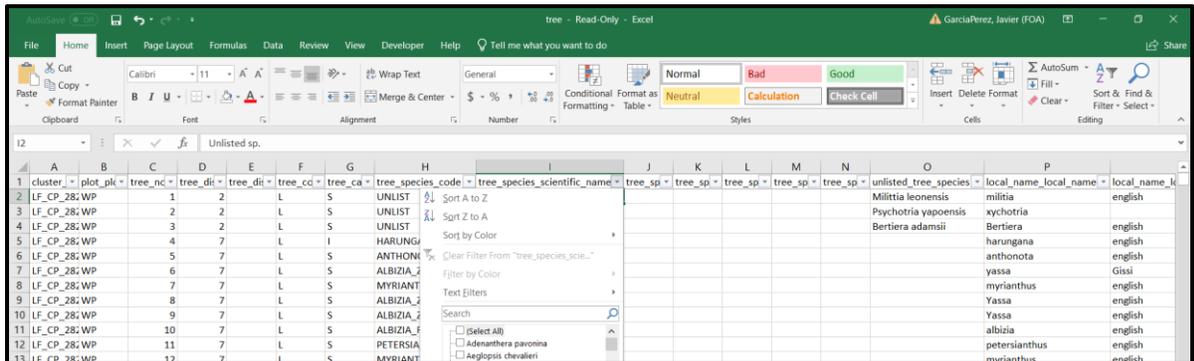
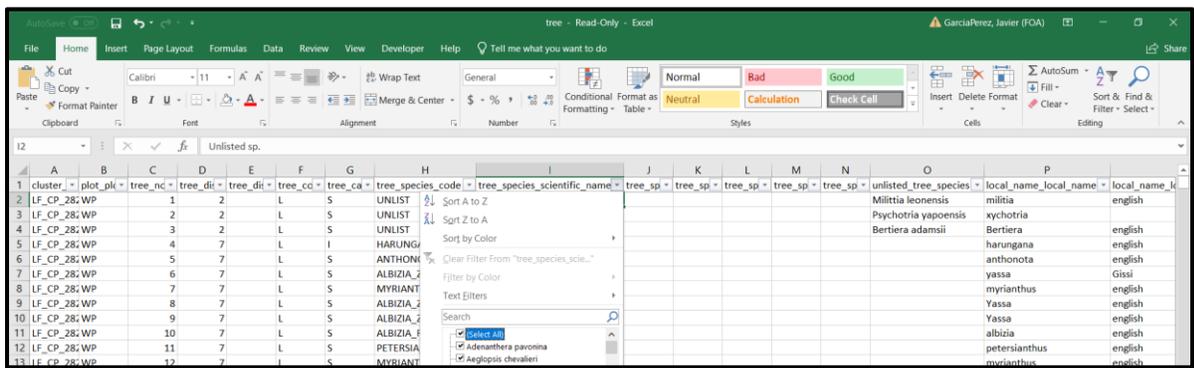
1. Open with Excel the <youname>.tree.csv file we have just created.
2. Navigate to the corrected tree sheet, freeze the Top Row containing the variable names



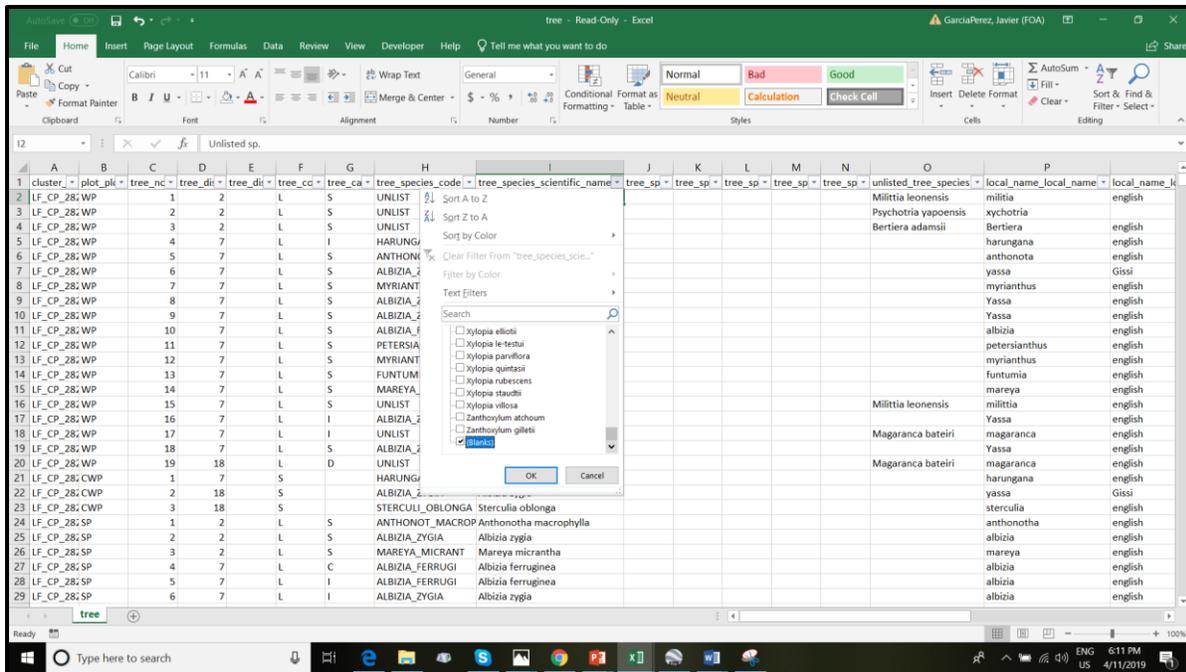
3. Locate cursor on a cell that is not blank and filter (here we want to apply a filter option to all columns in the data set)



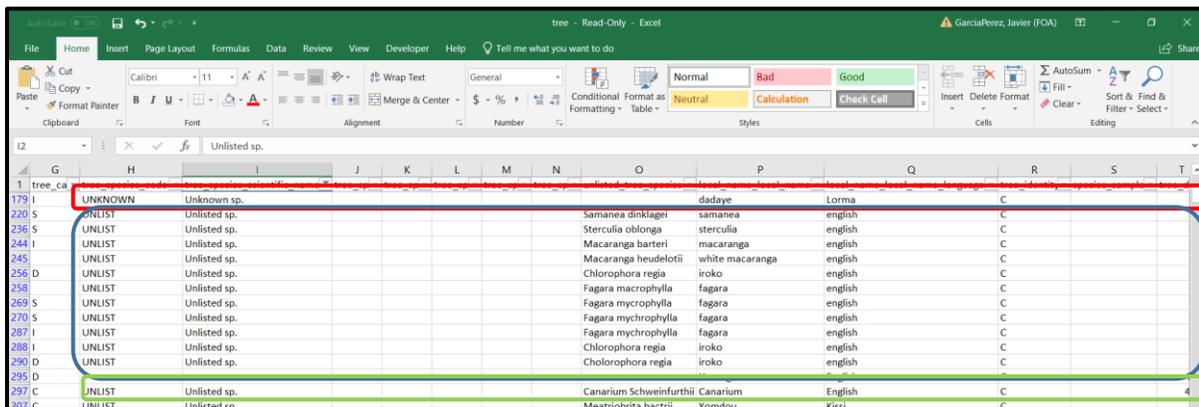
4. Click on the arrow on the right side of the tree_species_scientific_name (Column I) row name. A menu of options appears. Click on (Select all) to leave all options blank.



5. Scroll down on the list of options and select only the boxes for *Unlisted sp.*, *Unknown sp.* and Blank cells.



An example of the current status is shown below.



Species review and cleaning

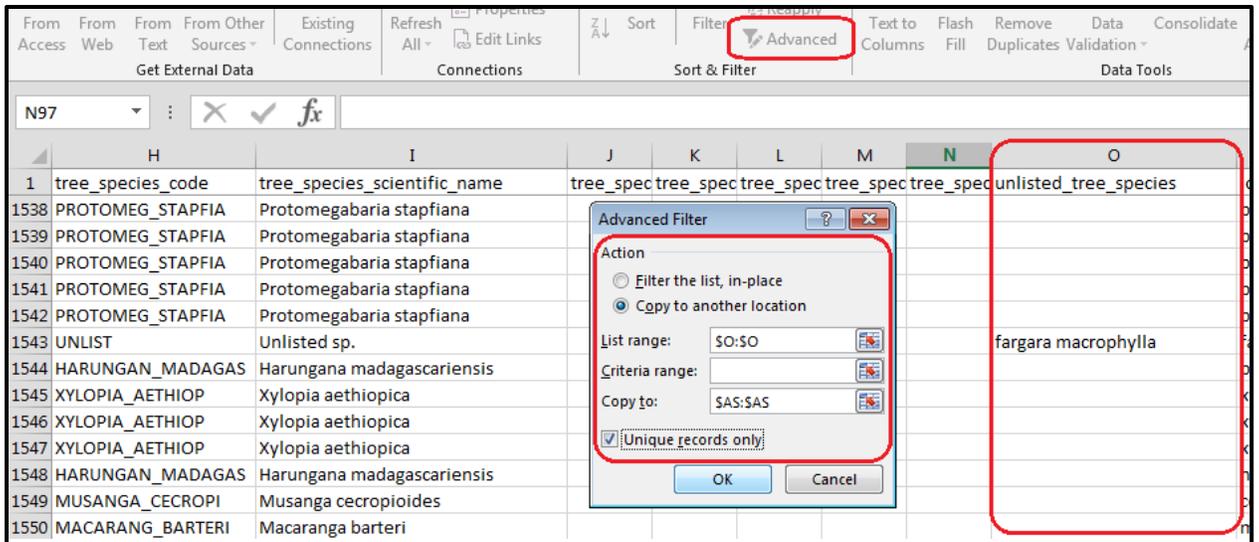
Examples for unknown/unlisted species and blanks are shown in red, blue and green, respectively.

Group by group a schematic process of cleaning is:

- Unknown sp.*: One can see that the category is wrongly inputted, since local names are provided for the species. If indeed these names are correct (that is, unless the crews did not know the species and decided to provide a ballpark local name suggestion), then the species should have been classified as *UNLIST* under *tree_species_code*, *UNLIST sp.* under *tree_species_scientific_name*, and the *unlisted_tree_species* name should be found by using a glossary lookup table once this glossary is finished. Yet, for the moment, we aim to leave

Unknown sp. as it is, until a glossary of local names, relating them to scientific names, is available in future developments.

- b) *Blank cell* under **tree_species_scientific_name**. They are errors since it is not mandatory to fill up this variable. However, there should be two potential solutions here:
- i) To consider it in the same way as *Unknown sp.* in the recommendation above. That is, if a local name has been provided (as in the figure), then **tree_species_scientific_name** should be corrected as *Unlisted sp.*, the **tree_species_code** as *UNLIST* and the **unlisted_tree_species** name should be found by using a glossary lookup table once this glossary is finished.
 - ii) To temporarily change the **tree_species_scientific_name** and **tree_species_code** to *Unknown sp.* and *UNKNOWN* respectively. Please pay attention to the period symbol in “sp.”!!
- c) *Unlisted sp.* This is the more important category to work with as a prioritizing one. One has to go cell by cell checking the name under **unlisted_tree_species**. These names are likely to contain orthographic mistakes and the person in the data management team will need to make use of several tools to double check names. Before we start this process we will need to create a list of unique unlisted tree species names based on those found in **unlisted_tree_species**.
- i) With the whole <youname>.tree.csv (unfiltered) data set, sheet corrected-tree, filter for unique values, click **Data > Filter > Advanced**. A tab appears. Click on “Copy to another location”. Then select the list range. In the above figure the **unlisted_tree_species** column is in column **O**. Then in the List range we input **\$O:\$O**. We will create the list of unique values in a new column. In my case I selected column **AS** at the right end of the csv file. Then I input **\$AS:\$AS**. Finally I check the box “**Unique records only**”. Press ok. A new list of unique species will be created in column **\$AS**. We will want to copy this into a new sheet as we will be filtering and updating **corrected_tree**. Create a third sheet in your worksheet and call it **unique_unlisted**. Cut column **AS** from **corrected_tree** to **Unique_unlisted** and paste it into column **B**.



- ii) In your third worksheet you can see column AS where the whole list of unique names will show up. First modify AS top row name from **unlisted_tree_species** to **unlisted_tree_species_unique**. In the top row of column AT (the next one to the right) create a new variable name: **unlisted_tree_species_corrected**. We will input corrected names on this column. So far we have something like this. Remember that your list of species will be different when compared to this one, it is meant to be that way.

AS	AT
unlisted_tree_species_unique	unlisted_tree_species_corrected
Millettia leonensis	
Psychotria yapoensis	
Bertiera adamsii	
Magaranca bateiri	
Millettia macrophylla	
Millettia liberica	
Samanea dinklagei	
Santaria trimera	
Hannoa klaineana	
Uapaca paludosa	
Parinari glabra	
Hallea ciliata	

- iii) First, observe that there is a blank in the fourth cell of column AS. This is not a problem. It should just represent those Unknown or blank species (which would not include anything under **unlisted_tree_species**). Select the whole list from column AS. In my case the list goes from \$AS\$2 to \$AS\$491. Then copy (Ctrl-C). Go to the Taxonomic Name Resolution Service (TNRS) webpage <http://tnrs.iplantcollaborative.org/TNRSapp.html> and paste the list inside the box **Enter List**. Then click on *Submit List* at the bottom right corner.

Enter List | Upload and Submit List | Retrieve Results

Enter scientific names to check

Milittia leonensis
 Psychotria yapoensis
 Bertiera adamsii

Magaranca bateiri
 Milittia macrophylla
 Milittia liberica
 Samanea dinklagei
 Santaria trimera
 Hannoa klaineana

[Click here for support](#) Clear Submit List

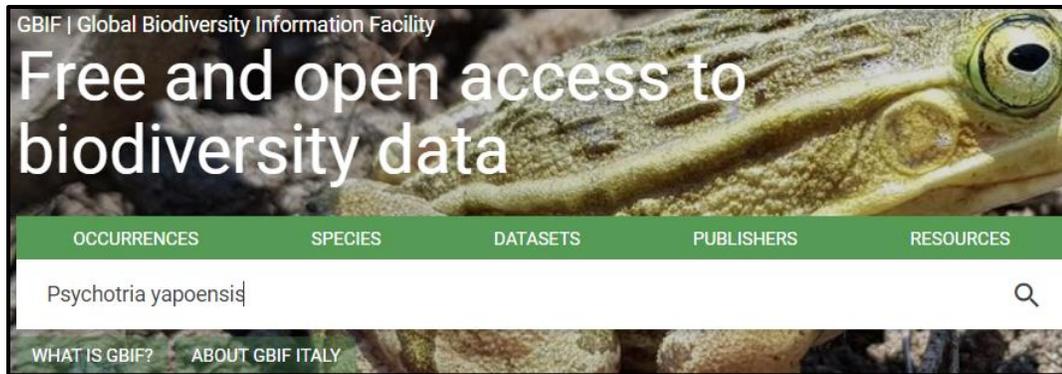
iv) The list will take a while to process. At the end a box below appears looking like:

Name Submitted	Name Matched	Name Source	Overall Score	Taxonomic Status	Accepted Name	Details
Milittia leonensis	Millettia leonensis Hepper (+2 more)	ILDIS TROPICOS	.95	Accepted	Millettia leonensis Hepper	Details
Psychotria yapoensis	Psychotria yapoensis (Schnell) Verdc.	TPL TROPICOS	1.00	Accepted	Psychotria yapoensis (Schnell) Verc	Details
Bertiera adamsii	Bertiera adamsii (Hepper) N.Hallé (+1 more)	TPL	1.00	Accepted	Bertiera adamsii (Hepper) N.Hallé	Details
Magaranca bateiri	Magadania (+1 more)	TPL	.47	No opinion		Details
Milittia macrophylla	Millettia macrophylla Benth. (+2 more)	ILDIS TROPICOS	.95	Accepted	Millettia macrophylla Benth.	Details
Milittia liberica	Millettia liberica Jongkind	TPL TROPICOS	.95	Accepted	Millettia liberica Jongkind	Details
Samanea dinklagei	Samanea dinklagei (Harms)Keay (+1 more)	ILDIS	1.00	Accepted	Samanea dinklagei (Harms)Keay	Details
Santaria trimera	Santaria trimera (Oliv.) Aubrév.	TPL TROPICOS	.98	Accepted	Santaria trimera (Oliv.) Aubrév.	Details
Hannoa klaineana	Hannoa klaineana Pierre & Engl. (+1 more)	TPL	1.00	Synonym	Quassia gabonensis Pierre	Details

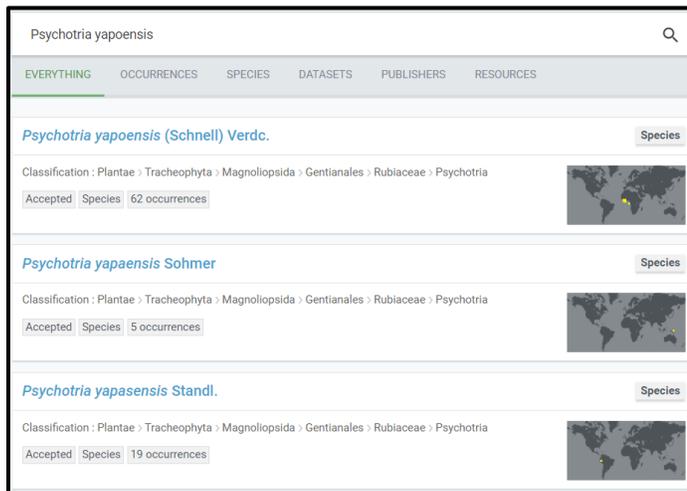
Page 1 of 5 | Displaying 1 - 100 of 490

The first column contains the name you entered. The fifth column is the taxonomic status (Accepted, Synonym, No opinion) of the potential closest official match found in the global taxonomic repository. For cleaning purposes we will be focusing initially on those species listed as Accepted or Synonym, No opinion will be addressed at a later date. As such those species that have an Accepted and Synonym status will be updated in the sheet corrected_tree. The sixth column contains the accepted name of the match. For example, in the first species introduced, *Milittia leonensis*, the database found a match: *Millettia leonensis*, with an *Accepted* status and 0.95 (95%) orthographic matching. This is a good match. Download these results from TNRS and use the Accepted taxonomic status for the second stage of the analysis. Update the sheet corrected_tree using methods described below in the second case example.

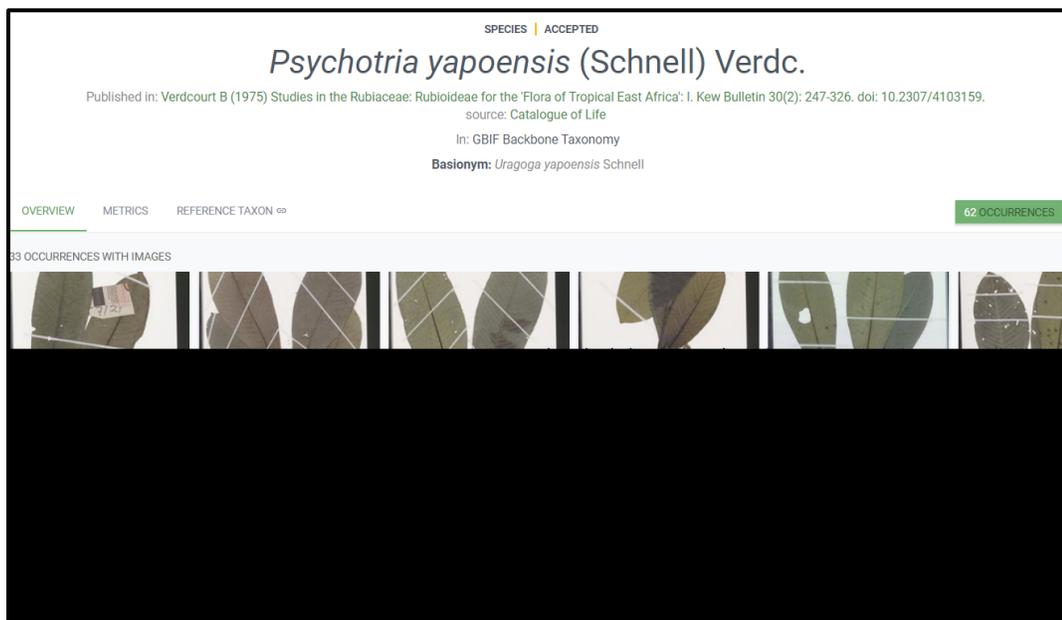
The second case is *Psychotria yapoensis*, with a 1.00 Overall score (100% match) and Accepted status. But it was inputted as Unlisted sp. and we should now double check by, for example going into the Global Biodiversity Information Facility (GBIF) website to look whether the species is in fact present around Liberia. All species that return an Accepted or Synonym status should be subjected to an additional analysis using the GBIF, use the data you downloaded from TNR for this part of the analysis. Go to <https://www.gbif.org/> and in the search box input the name (i.e., Ctrl-C, to copy from the list from TNRS and Ctrl-V to paste into the box in gbif.org)



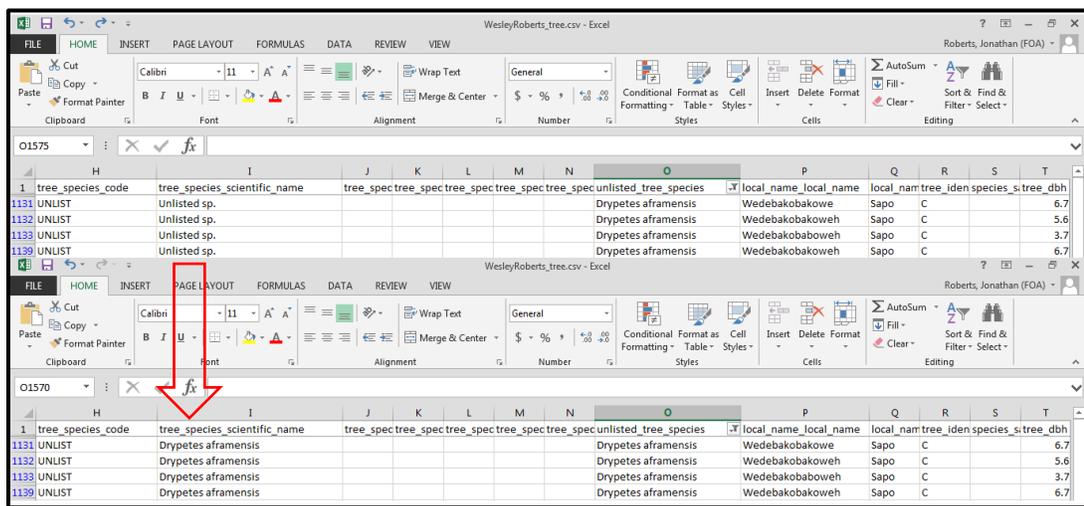
Then click in the lens on the right to search. The next screen is:



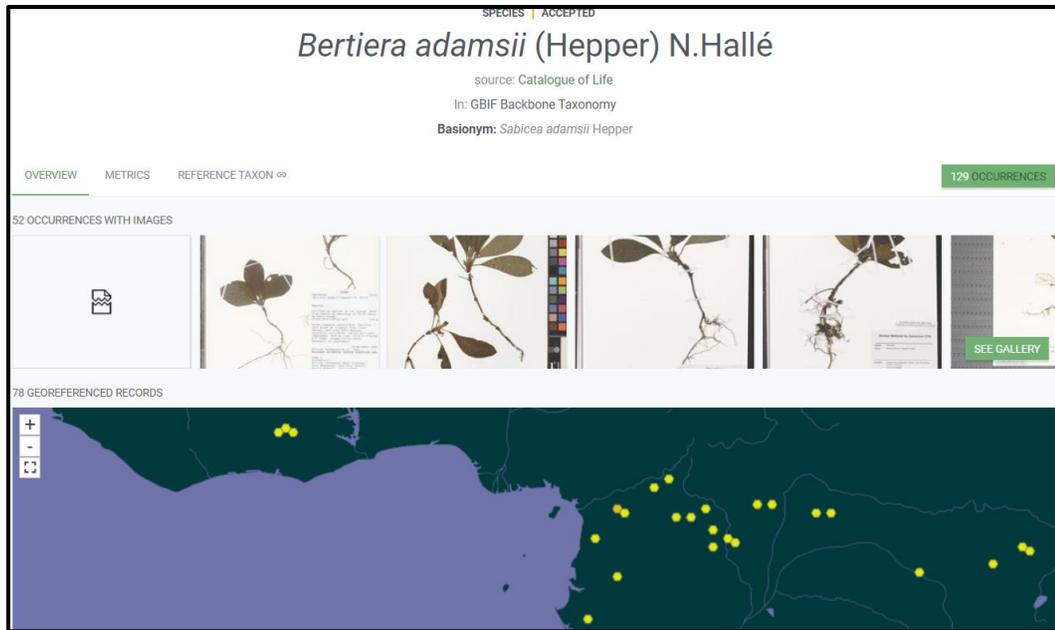
Then click on the first entry (click on the blue text) which reflects the exact name. A map will appear:



One can see that the species occurs in Liberia. This means that the entry in the COLLECT survey was correct. In this case we must go back to the sheet named corrected_tree and replace unlisted status with the correct species name. In the sheet corrected_tree apply a filter to column O. At this stage you will search for the unlisted specie you are analyzing in column O, filter for this species (remember to use the original name). Once you have found the unlisted_tree_species of interest you will need to update the tree_species_scientific_name column, changing the entry from *Unlisted sp.* to the entry in column O. Note that this should only be done when you have confirmed that the species exists in Liberia. See below an example where an unlisted species has been updated to *Drypetes aframensis*. All updates should be recorded in the error recording sheet.



The third case, *Bertiera adamsii*, is equally correct. However, the distribution map from gbif.org does not include Liberia. It is up to the data management team, in conjunction with botanists, to decide whether the identification is correct. In this case, the species has been previously found in Ghana. Given the proximity and ecological similarities between both Ghana and Liberia, it is quite possible that the identification in Liberia is correct. Hence the data management team should in this case maintain the identification and under **tree_specific_scientific_name** change from *Unlisted sp.* to *Bertiera adamsii*.



Once again data cleaning officers are encouraged to update corrected_tree sheet for the above instance.

The fourth species, written as *Magaranca bateiri*, is found in TNRS as *No opinion* under Taxonomic Status. One can make use of an extra tool at <http://plantminer.com/>. There, one must select **The Plant List** in the left black column. In the middle column one can leave the default options and tune the value (values between 0.7 and 1 are possible) for the name conservativeness. A lower value will try to find names like the name suggested but still quite different. A value of 1 will only try to find exact matches to the name provided. In the current case, leave it at 0.7. In the middle column below, one must paste the names of the species (one or several) in the box. Next one must click on **Process list**. In the third tab, on the right-hand side, an optional match will appear. In this case it is *Macaranga barteri*. One can again double-check its distribution in gbif.org. If the distribution seems correct, then it is obvious than the initial name was a misspelling. If you identify a species with a misspelt name remember to search using the misspelling but update the Unlisted sp. with the correct spelling from GBIF.

The Plantminer interface shows the following data in the 'Processed taxa' section:

id	family	genus	species	infraspecific rank	infraspecific epithet	authorship	taxonomic status in tpl	confidence level	source	accepted id	name	note
1	Euphorbiaceae	Macaranga	barteri			Müll.Arg.	Accepted	H	WCSP		Macaranga barteri	was misspelled

We must introduce the new match *Macaranga barteri* and remove *Unlisted sp.* under `tree_species_scientific_name`. In fact, *Macaranga barteri* was already present in the initial species list, so it has its own code. Remember to make all changes to the corrected_tree sheet and to save your work as go through the species list. Also remember to record all updates in the error recording sheet.

To check the species codes for the species in the original species list, do it by searching with Ctrl-F in Excel and inputting *Macaranga barteri*. The original species list is in COLLECT and can be exported to an csv file as seen below.

Rank	Num	Code	Scientific name
Family			Acanthaceae
Family			Achariaceae
Species		ACRIDOCA_PLAG	Acridocarpus plagiopterus
Genus			Acridocarpus sp.
Species		ADENANTH_PAVC	Adenanthera pavonina
Genus			Adenanthera sp.
Species		AEGLOPSI_CHEV	Aeglopsis chevalleri
Genus			Aeglopsis sp.
Species		AFROLICA_ELAEC	Afrolicania elaeosperma
Genus			Afrolicania sp.
Species		AFZELIA_BELLA	Afzelia bella
Species		AFZELIA_PARVIFI	Afzelia parviflora
Genus			Afzelia sp.
Species		AGANOPE_LEUCO	Aganope leucobotrya

Sometimes the status of an entry appears as a *Synonym*. This is the case with the last species in the original list returned by TNRS, *Hannoa klaineana*. One must be particularly careful with synonyms, since the species to report should have a taxonomix status as *Accepted*. In this case TNRS suggest *Quassia gabonensis*. One could check the distribution in gbif.org, but the closest specimens are found in Benin, which is far enough to be suspect of this species. However, one can directly input *Hannoa klaineana* in gbif.org. The result shows that the species is vastly present in Liberia. But also shows that *Hannoa klaineana* is in fact a synonym of *Quassia undulata* (red rectangle in figure below). One can later check in the original species list that *Quassia undulata* is in fact already present. Hence, we would correct the **tree_species_scientific_name** and **tree_species_code** according to the requirements for *Quassia undulata*.

SPECIES | SYNONYM

Hannoa klaineana Pierre & Engl.

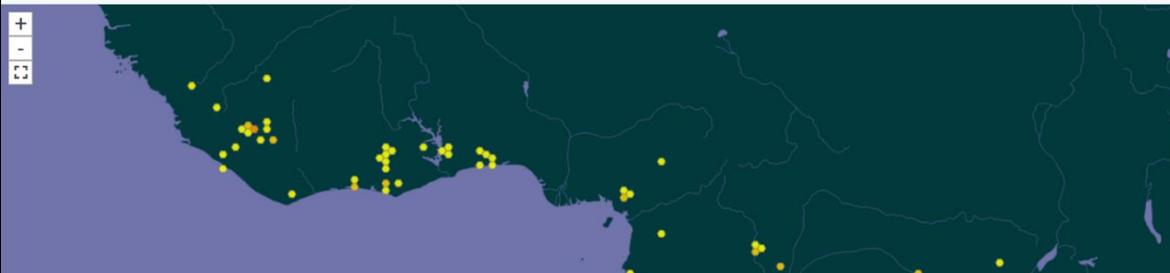
In: GBIF Backbone Taxonomy
 Synonym of Quassia undulata D.Dietr.

OVERVIEW METRICS 357 OCCURRENCES

101 OCCURRENCES WITH IMAGES



261 GEOREFERENCED RECORDS



It is very important to check all these Unlisted species because a large fraction of them have been classified as unlisted when they already exist in the original species list, but were either misspelled by the field crew person, or else are old synonyms, instead of the most currently accepted names, were used. However, it is also very important to check also for potentially new species that might not be present in Liberia according to gbif.org, and hence are not in the current original species list of the NFI.

All updates to the database must be implemented and saved in the corrected_tree sheet. Please run through all of the unlisted species listed in the Unique_unlisted sheet. You may want to delete the species once you have updated it elsewhere as this will help with keeping track of your work. For the clusters enumerated in Sinoe County the database has approximately 67 unlisted species to review. The clusters assigned to you may include more or less species. The excel file you update should be returned to the CTA (jonathan.roberts@fao.org) when you have completed the entire exercise along with the rest of the data from this exercise. A separate data management sheet will be provided to the data cleaning staff with specific instructions.

Step 4 – Quantitative analysis of biophysical data

Before moving ahead with the preparation activities it is important to have a quick look at the data we downloaded in the previous section. Navigate to the file location where you saved the cleaned database from Collect. You should see a list of *.csv files. These are the files which contain some of the information we will be visualizing and analyzing.

CSV File	Contents
access_to_cluster.csv	access photo names and locations
amphibian.csv	Amphibian information captured
bird.csv	Bird information captured
canopy_cover.csv	Canopy cover for each plot - data from densiometer
cluster.csv	Cluster data including access and remarks
cwd.csv	Coarse woody debris measured in the field
fwd.csv	Fine woody debris measured in the field
mammal.csv	Mammal information captured
ntfp.csv	Information concerning non-timber forest
plot.csv	Plot information including descriptive data and plot level estimates
plot_photos.csv	Plot photo names
prominent_structure.csv	Prominent structure names and photo
reptile.csv	Reptile information captured
tree.csv	Tree information for each plot

Preparing your checklist

It is important to start with a good understanding of the model data and the field guide that was used, in the form of a simple overview of the variables. The data management team should at this stage be familiar with the data collected in the field and have a good understanding of the probable ranges of values for all categorical and continuous variables collected during the field inventory. Below find a list of variables we will be focusing on for the present analysis as well as the csv file this data can be found in. This list may change in the future and will certainly be expanded upon when time allows. For now we will be focusing our attention on those variables which affect the calculation of carbon within the forest. We will be looking at individual variables as well as the relationships between key variables to identify potential errors in the data set. We will not be physically changing those values, we will only be making a note of their location. Following discussions with experts in HQ we will review the output list of potential issues and collectively decide on how to update the database. The table below provides a list of variables we will

be assessing in this data cleaning exercise including the range of expected values. The csv file which contains the data is also included.

Table 1 Inventory variables of interest

No.	Variable	Acceptable Ranges	*.csv file	data_analysis_<yourname> sheet
1.	Mean canopy closure	0 – 100%	plot.csv	plot
2.	Diameter at breast height	> 2cm	tree.csv	tree
3.	Tree Height	1.3m > 100	tree.csv	tree
4.	Bole Height	1.3m > 100	tree.csv	tree
5.	Basal area per hectare	XX	plot.csv	plot
6.	Trees per hectare	XX	plot.csv	plot

Following a review of single variables, we will also be looking at the relationships between variables to determine if there are any outliers in the data set. Here we will use simple scatter plots and z-scores to identify those records that require additional attention. Below please find a list of variables to be compared during your investigations.

Table 2 Relations between inventory variables

No.	Variable 1	Variable 2	Method
1.	Mean canopy closure	Trees per hectare	Scatterplot
2.	Mean canopy closure	Basal area per hectare	Scatterplot
3.	Diameter at breast height	Tree Height	Scatterplot / z-score analysis

Prepare raw analysis data

Table 1 provides an overview of the key variables we will be assessing as well as the files containing the relevant data. You will see that the data of interest are contained in two files, plot.csv and tree.csv. The present analysis will not be updating the raw csv files, rather the erroneous data points will simply be identified and recorded for additional updating at a later time (See Annex 1). The first step in the analysis is however to identify and select the clusters which have been assigned to you and to transfer this data to a separate data analysis file. In the directory containing the csv files discussed above create an excel file and call it data_analysis_<yourfirstname>.xlsx.

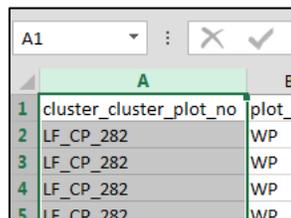
Documents library

collect-csv-data-export-liberia_nfi_utm_20180907-2018-10-12T10_33_31

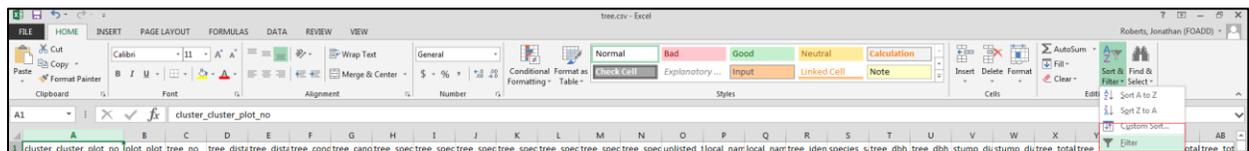
Arrange by: Folder ▾

Name	Date modified	Type	Size
access_to_cluster.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	33 KB
amphibian.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	14 KB
bird.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	101 KB
canopy_cover.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	49 KB
cluster.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	25 KB
cwd.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	18 KB
data_analysis_wesley.xlsx	10/12/2018 4:07 PM	Microsoft Excel Worksheet	7 KB
fwd.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	7 KB
mammal.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	87 KB
ntfp.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	34 KB
plot.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	266 KB
plot_photos.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	28 KB
prominent_structure.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	35 KB
reptile.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	29 KB
tree.csv	10/12/2018 10:33 AM	Microsoft Excel Comma Separated Values File	1,183 KB

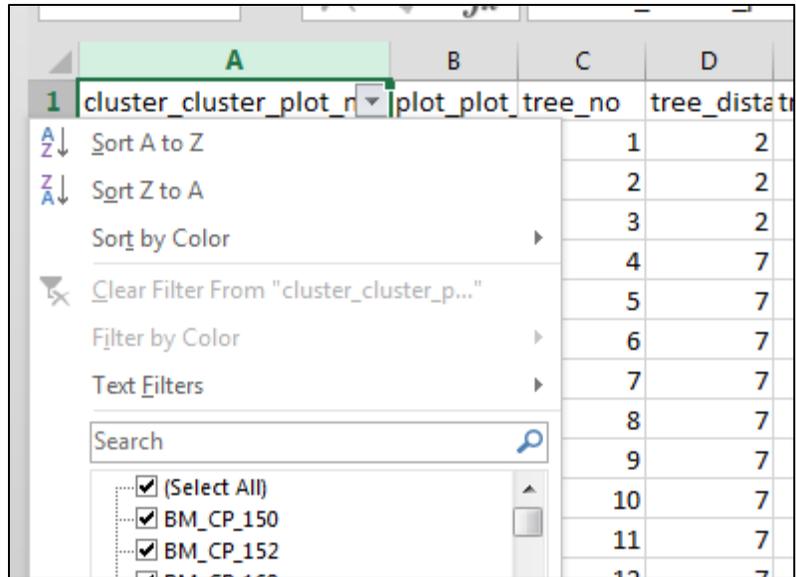
This is the file that will contain the data from your assigned clusters. Open the file and create two empty spreadsheets. Call one tree and the other plot, activate tree and got back to your windows explorer folder with the tree.csv and plot.csv files. Open the tree.csv file. Extend column A until you can see the full column header. Select the column by clicking directly on the A.



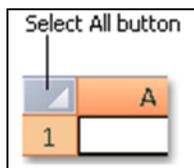
Once the column is fully selected, head over to the filter tab on the home ribbon in excel and apply the filter to the column.



Once we have applied the filter, click on the upside down arrow adjacent to the column name. A menu will appear containing all of the unique cluster names in the column. Using Liberia_NFI_<YourName.xlsx select the clusters which have been assigned to you in the list. We are now selecting the data which you will be analyzing for the bulk of the data cleaning period.



Be very sure to select only the clusters you are tasked with cleaning. This should be fairly easy as everyone has been assigned clusters by county. Make use of the search function to select county level data. For Gbarpolu and Lofa this is fairly easy as the clusters names begin with either LF_* or GP_*. You can even search for these by entering LF_* or GP_* in the search function. Apply the filter. Your tree.csv excel file should only present data for your assigned clusters. We are now going to copy this selection to the tree sheet in our data_analysis_<yourfirstname>.xlsx excel file.



Using the select all button, select the entire worksheet, the whole sheet should have a transparent grey colour. Copy the entire sheet and paste it into tree sheet in the data_analysis_<yourfirstname>.xlsx excel file. Save the data in this file.

Repeat the process for the plot.csv file and copy the resulting selection into the plot sheet of data_analysis_<yourfirstname>.xlsx. It is crucial that you make sure that you select the same clusters in both the tree.csv and plot.csv files, you must review the same clusters in both tree.csv and plot.csv. Save the file. We now have our data ready for analysis. This guide will use randomly selected clusters so the graphs and tables you see here will be different from those in your data set.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J
1	cluster_cluster_plot_no	plot_plot	tree_no	tree_dist	tree_dist	tree_cond	tree_cano	tree_spect	tree_spect	tree_spect
2	GP_CP_221	EP	1	1.2		L	S	DIOSPYRC	Diospyros gabu	
3	GP_CP_221	EP	2	3.5		L	S	DIDELOTI	Didelotia unifol	
4	GP_CP_221	EP	3	3.7		L	I	CARAPA_I	Carapa procera	
5	GP_CP_221	EP	4	5.9		L	I	CARAPA_I	Carapa procera	
6	GP_CP_221	EP	5	6		L	I	CARAPA_I	Carapa procera	
7	GP_CP_221	EP	6	6.1		L	I	CARAPA_I	Carapa procera	
8	GP_CP_221	EP	7	6.4		L	I	VEPRIS_TA	Vepris tabouen	
9	GP_CP_221	EP	8	6.8		L	I	NEWTONI	Newtonia aubre	
10	GP_CP_221	EP	9	3.5		L	D	VEPRIS_TA	Vepris tabouen	
11	GP_CP_221	EP	10	15.8		L	D	ERYTHROF	Erythrophleum	
12	GP_CP_221	EP	11	12.4		L	D	ERYTHROF	Erythrophleum	
13	GP_CP_221	EP	12	17.2		L	D	PYCNANTI	Pycnanthus ang	
14	GP_CP_221	EP	13	7.5		L	C	DIALIUM_I	Dialium aubre	
15	GP_CP_221	EP	14	5.2		L	C	VEPRIS_TA	Vepris tabouen	

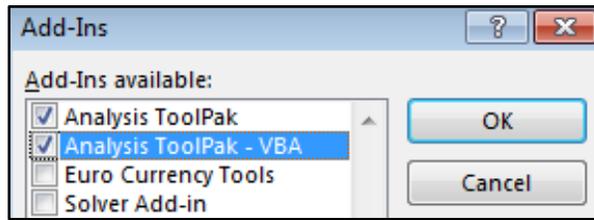
Analysis implementation – Single variables

Once you completed preparing your data_analysis file it is time to begin the analysis / investigative component of your work. We will use table 1 as a reference for each of the variables we will be analyzing. Before we start be sure to activate the Analysis ToolPak and Analysis ToolPak – VBA from excel Add-Ins. We will be using the Analysis ToolPak for the first part of our analysis.

To activate an Excel add-in

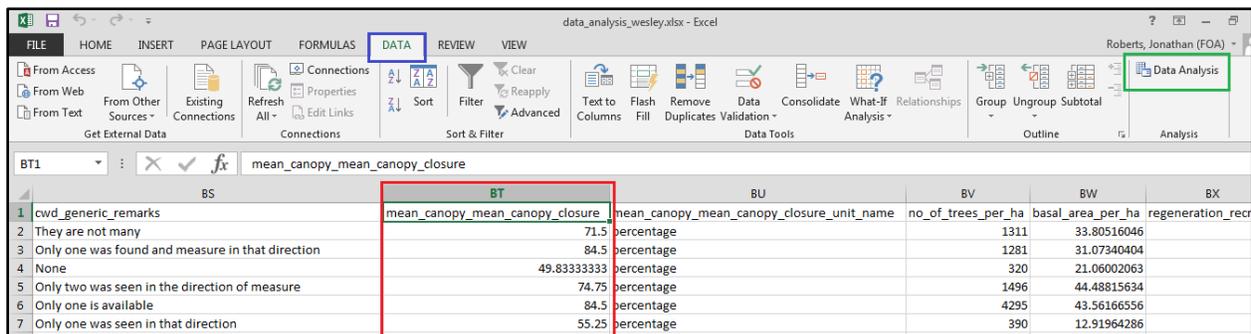
1. Click the **File** tab, click **Options**, and then click the **Add-Ins** category.
2. In the **Manage** box, click **Excel Add-ins**, and then click **Go**.
The **Add-Ins** dialog box appears.
3. In the **Add-Ins available** box, select the check box next to the add-in that you want to activate, and then click **OK**.

The Add-Ins dialog box should appear something like this. Select Analysis ToolPak and Analysis ToolPak – VBA.



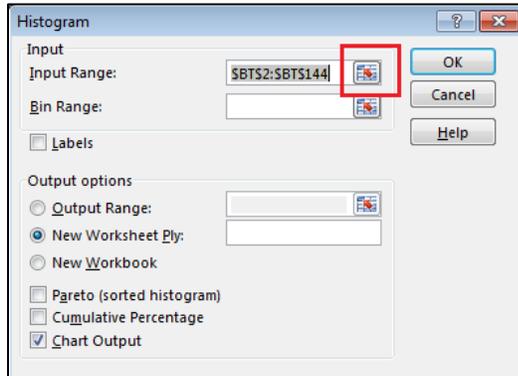
Mean canopy closure

Mean canopy closure is an important biophysical variable collected during the inventory and will potentially help to stratify the plots later on when the inventory is finalized. The information is captured at the plot level (1 value per plot) and thus here we will be using the plot sheet in the data analysis excel file. Canopy closure values are generated by the OpenForis COLLECT survey based on measurements collected in field. As such we expect the range of values currently stored in the data set to be relevant i.e. they should range from 0 - 100. For the present analysis we are simply going to graph the data and generate some exploratory statistics. We will be using similar methods for the rest of the variables in table 1. Navigate to the plot sheet, make sure you have activated the necessary Add-Ins. Column BT contains the mean canopy closure data we are interested in. Navigate to this column and select the Data tab (blue box below).

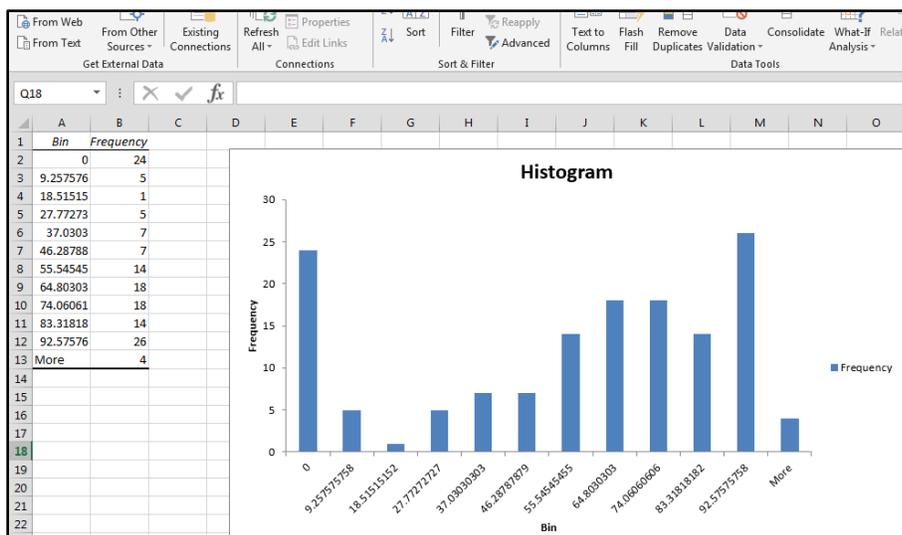


The DATA tab contains the analysis tools we will be using to analyse most of the single variable data, it can be found on the right hand side of the DATA tab (green box). Activate it, you should see the following. Scroll down and select the Histogram option. We are going to create a histogram to get a better idea of the underlying distribution of the data (outliers, skewness). For the present purposes we are simply getting used to the use of the tool and noting any strange or unexpected results. Once you have selected histogram, the following dialog opens. Select Chart Output, leave all other options as default. In Input

Range select the square box (red box). This now allows you to select the data you would like to use to create a histogram. Select the numerical data in column BT, do not select the top cell as this is not data and we will not be creating a graphic with a title.



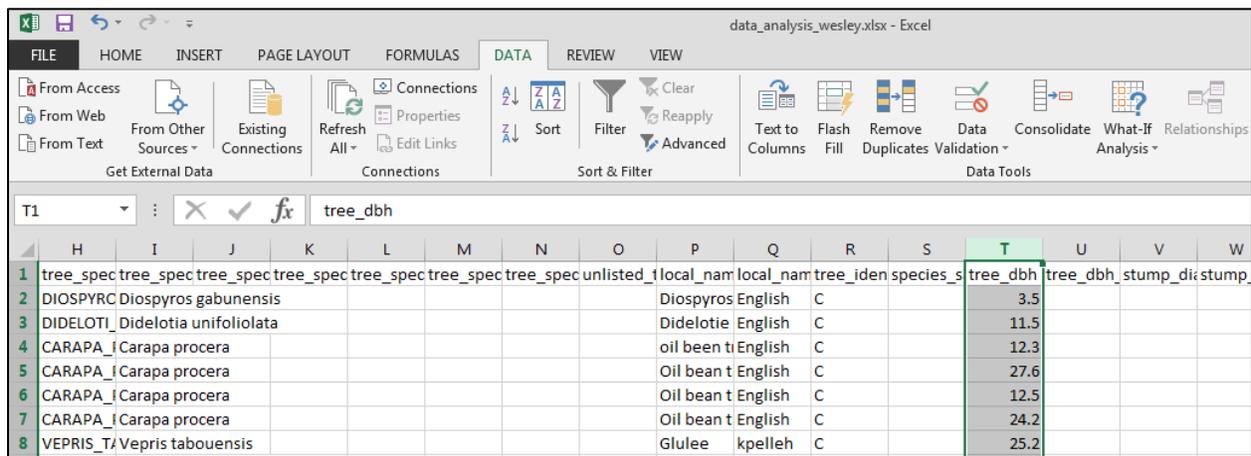
Once you have selected the data to graph, click once again on the square again and click OK (top right). A new sheet will be created with a frequency table as well as a histogram of the mean canopy closure. This is from the graphic you will be using for interpretation. See below a graphic of the histogram I created using my randomly selected data. You may notice on my graphic that there are almost 24 plots with 0 canopy closure and almost 25 with a value of 90% or larger. Between these two we seem to have a histogram with a slight negative skew. We will explore the nature of the histograms shape later when we compare the canopy closure with stems per hectare. For now make a note of the general shape of the histogram.



The histogram tool will open a new sheet and store the information in this sheet. Go ahead and change the name of this sheet to “Canopy closure histogram”. Save your data analysis worksheet and move on.

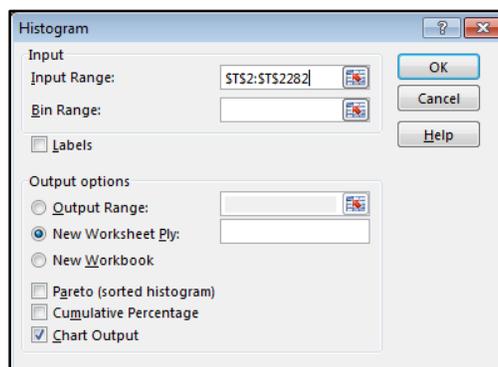
Diameter at breast height

We will now explore the diameter at breast height distribution using the same methods we used for canopy closure above. This time we will be using the tree sheet in your data_analysis worksheet. Recall that DBH is measured for all trees in the data set. The tree level DBH data is stored in column T with a heading called tree_dbh. Table 1 indicates to us that accepted values for this variable range from 2cm upwards, we do not specify a maximum DBH here but will make a note of any unusual values found in the data set. One potential error that might be prevalent is when those entering the data into the tablet enter the information using the incorrect units i.e. recording the data in meters instead of centimeters. This may not be evident in our single variable assessment, rather it will be explored when we compare DBH to tree height later on.



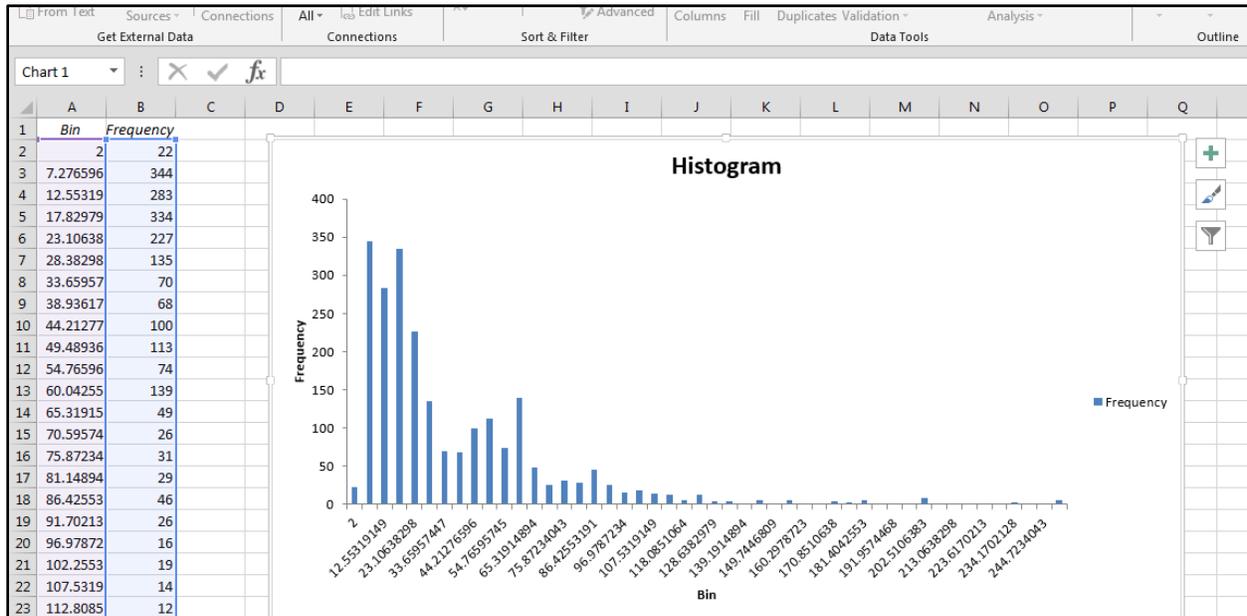
	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	tree_spec	tree_spec	tree_spec	tree_spec	tree_spec	tree_spec	tree_spec	unlisted_t	local_nam	local_nam	tree_iden	species_s	tree_dbh	tree_dbh	stump_di	stump
2	DIOSPYRC	Diospyros	gabunensis						Diospyros	English	C		3.5			
3	DIDELOTI	Didelotia	unifoliolata						Didelotie	English	C		11.5			
4	CARAPA_f	Carapa	procera						oil bean t	English	C		12.3			
5	CARAPA_f	Carapa	procera						Oil bean t	English	C		27.6			
6	CARAPA_f	Carapa	procera						Oil bean t	English	C		12.5			
7	CARAPA_f	Carapa	procera						Oil bean t	English	C		24.2			
8	VEPRIS_T	Vepris	tabouensis						Glulee	kpelleh	C		25.2			

Using the histogram function we utilized earlier, create a histogram of the DBH values in the clusters you have been assigned. Remember not to select the column title and to select graphic output.

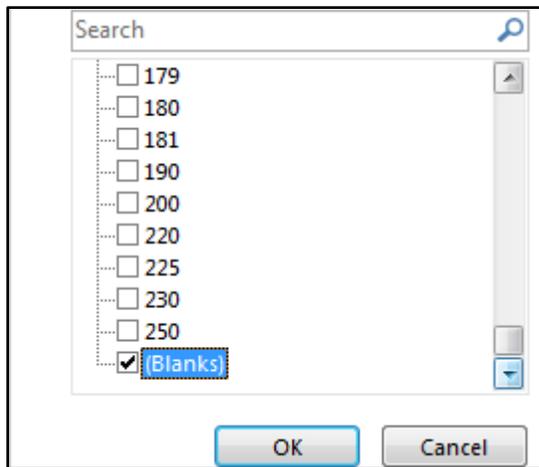


Go ahead and create the histogram for tree DBH, your graphic should look similar to the graphic below but recall that we will all be using different data so the shape of the histogram may be different. The

graphic below indicates that the distribution of DBH values in my data set is positively skewed indicating that the majority of the values in the data set are at the lower end. It is preferable to have a distribution which is normal in nature but this type of shape is not uncommon. We note here that there are a number of trees with DBH values greater than 150cm. These values may be outliers and may require additional investigation at a later date. Record your observations in the error recording data sheet.



For the present analysis we will also investigate if there are missing values in the data set. We see that there are a number of values that fall way above 200cm which should not be uncommon in Liberian forests. The shape of the histogram may be problematic but this can be fixed later. We now want to check to see if there are missing DBH values. Apply a filter to column T, deselect all values and scroll down until you find the (Blanks). We will now see how many DBH values are missing in the data set. Select OK and only those trees without DBH values will remain in the spreadsheet.



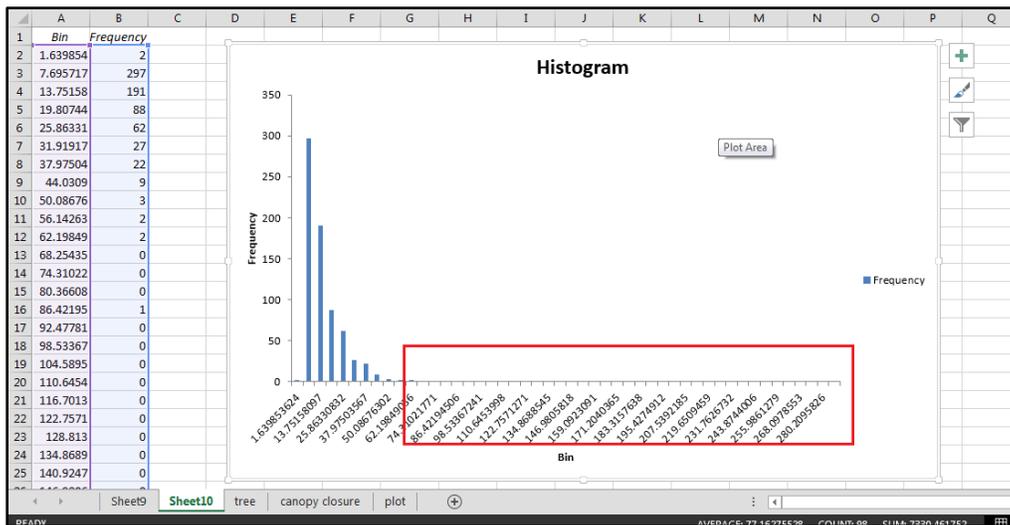
If your data has missing DBH values, scroll to the left and make a note of the clusters where the missing DBH values occur, capture these in the error recording sheet. In the present example we see that the missing DBH values all come from one cluster. If you have missing DBH values from multiple clusters make a note of these clusters names and we will follow up at a later date with the team leaders. Note that columns G to S are hidden in the graphic below to facilitate explanation.

	A	B	C	D	E	F	T	U	V	W	X	Y
1	cluster_cluster_plot_no	plot_plot_type	tree_no	tree_distance	tree_distance_unit_name	tree_condition	tree_di	tree_dbh	stump_di	stump_di	tree_total_height_distance	tree_total_height_distance_un
2254	GP_CP_187	EP	2	2		S				6		
2255	GP_CP_187	EP	4	7		S				23		
2256	GP_CP_187	EP	6	7		S				12		5
2257	GP_CP_187	EP	7	7		S				25		
2258	GP_CP_187	EP	8	7		S				30		
2259	GP_CP_187	EP	9	7		S				35		5
2260	GP_CP_187	EP	10	7		S				33		
2261	GP_CP_187	EP	11	7		S				30		
2262	GP_CP_187	EP	12	18		S				45		5

Tree Height

The next variable of interest is also present in the tree data sheet, navigate to column AB with header tree_total_height. This column contains the calculated tree heights based on the distance to the tree and the angle from this distance to the top of the tree. As such the tree height values are calculated by the survey. The method for measuring tree heights specifies that every third tree height is measured in the field. As such for us to create a histogram of tree heights we will need to remove the blank cells. As with the previous approach, filter the tree height and remove blanks from the data set, simply scroll down to the bottom of the filter list and unselect the (Blanks) option.

Create a histogram of the tree heights using the data analysis tool as before. Remember to leave all options default but to select a graphic output. As before once you click OK the application will create a new sheet with a frequency table as well as a histogram output. The example graphic below immediately highlights an error in the data set. The red box highlights the large positive skew, further we also note that within the data set one of the trees has a height of 280m. This is obviously an error which needs to be investigated further. The tallest tree in the world is less than 150m so this is obviously an error. Our valid data range in table 1 ends at 100m. As such if your data contains any trees above 100m these should be noted and recorded for further investigation. Use the error record sheet for this purpose.

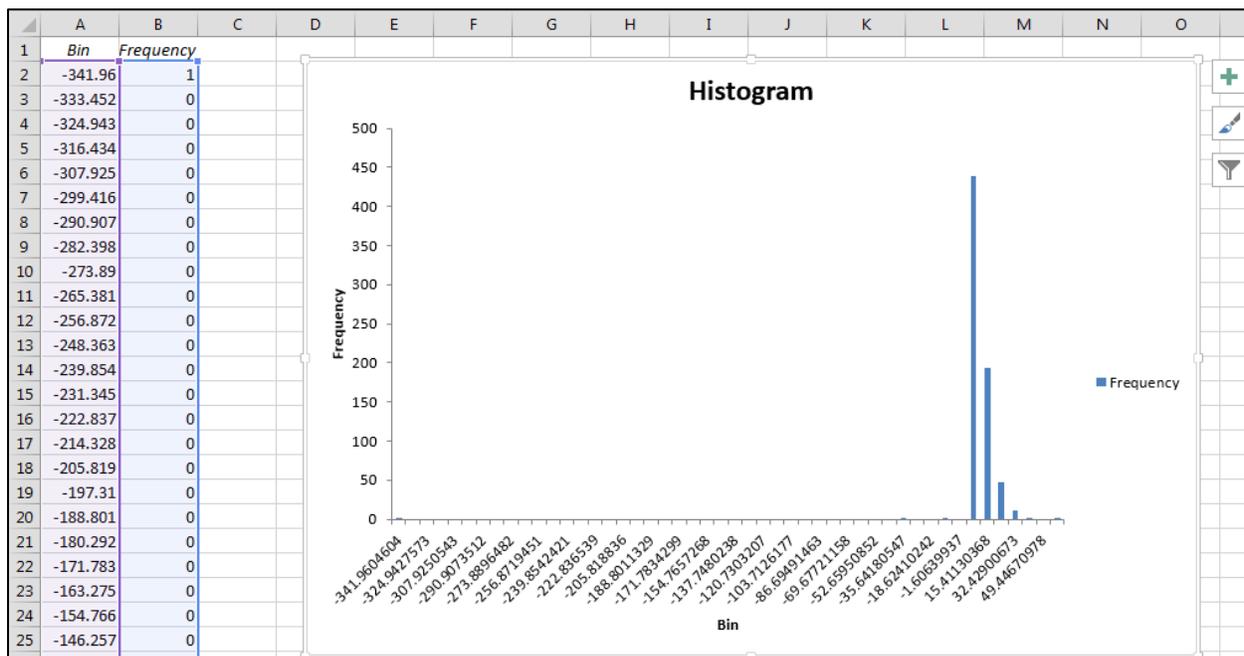


It may be useful to also sort the column from largest to smallest, this way you will be able to identify the particular tree which has returned the unusual tree height. Below we can see the cluster name as well as the plot and tree number of the 280m tree. Make a note of this information as we will be returning to this and other errors later on. If your data has multiple trees with heights above 100m note the cluster, plot and tree number on your error recording sheet.

cluster	cluster_plot_no	plot	plot_type	tree_no	tree_distance	tree_distance_unit_name	tree_condition	tree_canopy_position	tree_species_code	tree_species_scientific_name	tree_total_height	tree_total_height_unit_name
GP_CP_207	SP	6	17.4	L	C	PIPTADEN_AFRICAN	Piptadeniastrum africanum	286.2654463	met			
GP_CP_221	WP	12	15	L	D	PARINARI_EXCELSA	Parinari excelsa	80.7027991	met			
GCM_CP_233	CWP	12	14.1	L	D	PIPTADEN_AFRICAN	Piptadeniastrum africanum	58.37278258	met			
GCM_CP_233	SP	21	6.1	L	D	SACOGLOT_GABONEN	Sacoglottis gabonensis	57.65091935	met			
GCM_CP_218	SP	18	17.7	L	D	GILBERTI_PREUSSI	Gilbertiodendron preussii	55.68585255	met			
GCM_CP_217	CWP	9	6.5	L	D	PARINARI_EXCELSA	Parinari excelsa	50.95459355	met			

Bole Height

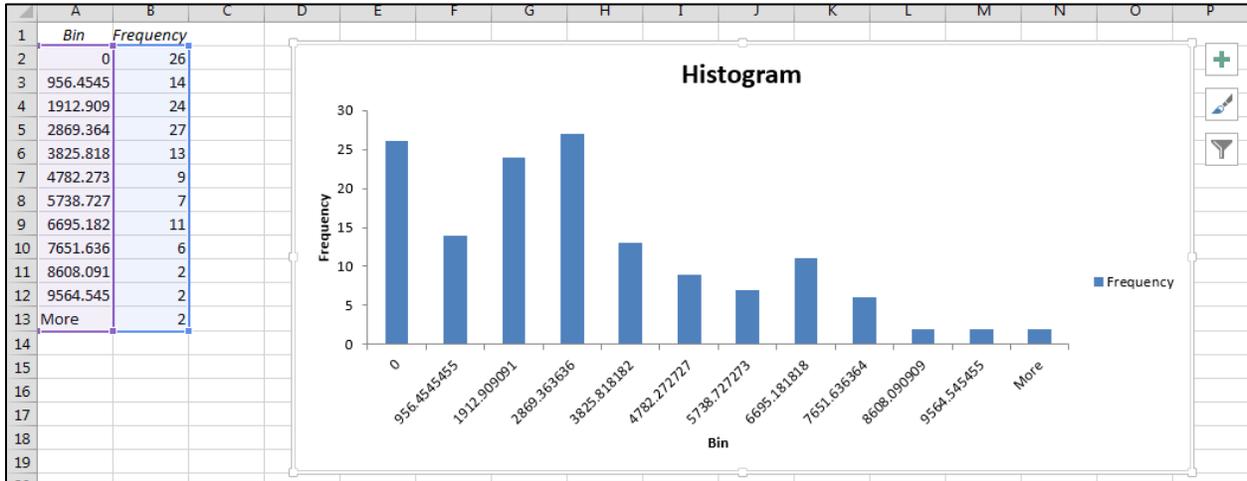
Bole height has similar data ranges compared to the tree height measure above. The survey once again identifies errors relating to the bole height measurement for each tree i.e. if the bole height is higher than the tree height there is an error. These errors should have been identified during the validation report prepared earlier. For now we will simply be looking at the distribution of the data and identifying any unusual values. As before prepare a histogram of the data using the analysis tool, the bole height data can be found in column AH. Below is the histogram from the randomly selected clusters used for illustration. The histogram below highlights an immediate error that needs to be recorded. The negative value is clearly wrong as it is impossible for any bole height to be negative. Use the filter function or sort column AH to identify the cluster, plot, and tree number of the erroneous value. It will be corrected later on following completion of the data cleaning activities. If there are multiple negative values make a note of each tree in the data error recording sheet.



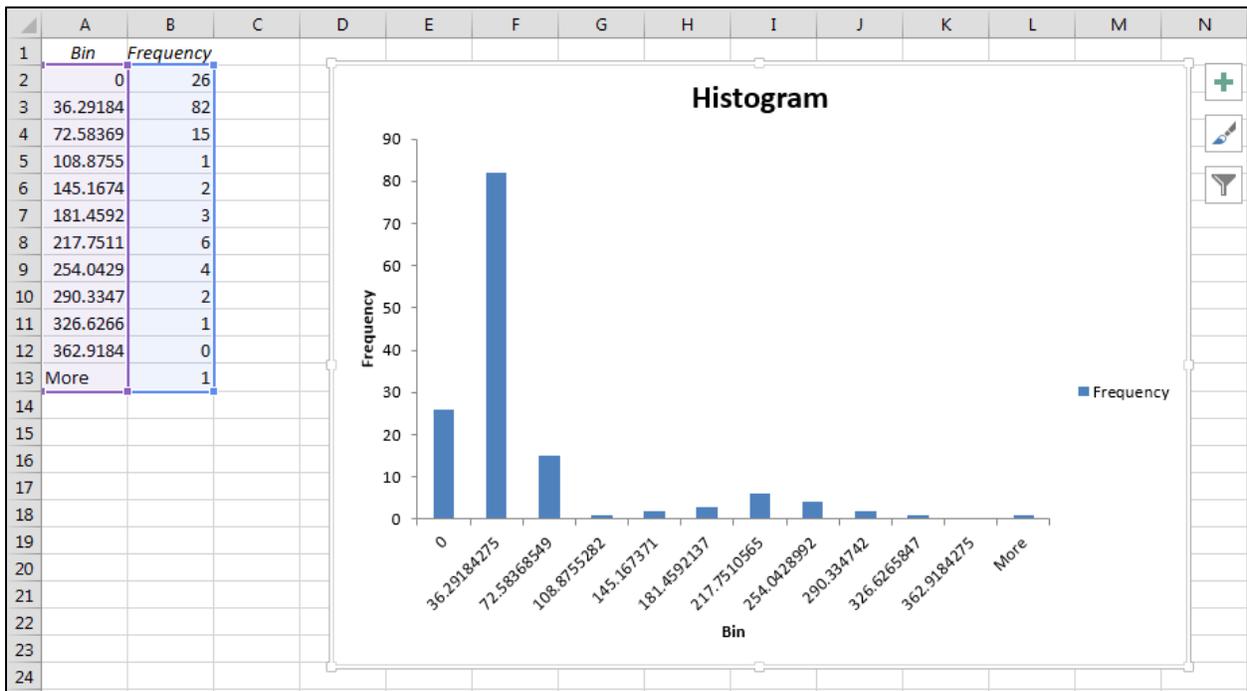
Trees per hectare \ Basal area per hectare

Both basal area per hectare and trees per hectare are plot level variables captured and calculated by the survey. Both variables are relevant for forest management activities but more importantly will be used for assessing the quality of the data through additional comparisons below. For now the purpose of the analysis is to verify that the values calculated by the survey are realistic and to identify any unusual patterns. Both calculated variables can be found in the plot sheet of your data and can be found adjacent to each other in columns BV and BW. Create histograms of both using the data analysis tool. Below find

the histogram for trees per hectare. It seems there are a number of plots with 0 trees per hectare, in fact in this data set 26 of the plots have 0 trees which is the second largest bin in the data set. Following your analysis, make a note of the plots with 0 stems per hectare.



Below find the histogram of the basal area per hectare. The distribution of basal area across the randomly selected plots is positively skewed but does show a slight bimodal distribution. This will warrant further analysis later, for now please make a note of the shape of the histogram and don't forget to save the worksheet in your data_analysis worksheet.

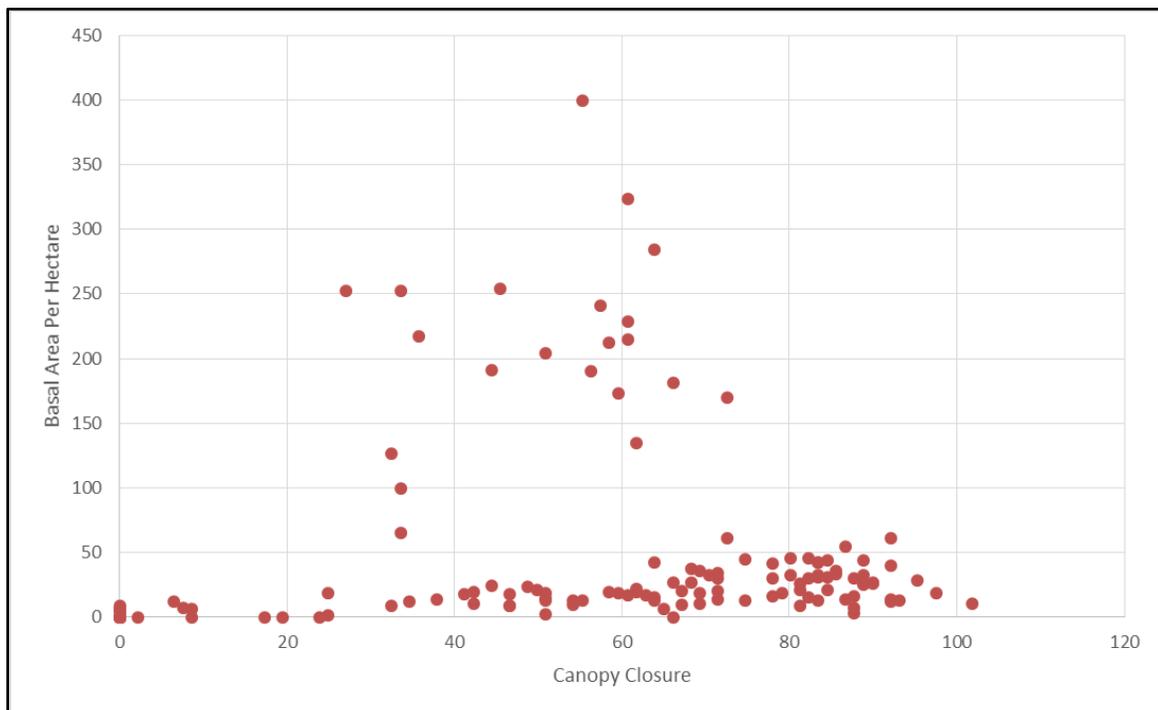


Analysis implementation – Two variables

The following analysis will compare two variables captured in the inventory with a view to identifying additional outliers and or erroneous results. The analysis will employ simple scatterplots as well as more advanced z-score analysis to compare the dbh and tree height measures from the inventory.

Mean canopy closure vs. Basal area per hectare

Both variables data can be found in the plot sheet of your data analysis excel worksheet. We will be comparing mean canopy closure to basal area as there should be a relationship between these two and if there are any unusual values in the data we will be able to identify these through visual interpretation. Navigate to the plot sheet in your data analysis worksheet, using the ctrl button select both mean canopy closure (column BT) and basal area per hectare (column BW). Once selected head to the insert tab and click on recommended charts. You should be given the option of creating a scatter plot as default. Click OK and the scatterplot should be created. As you can see from the graphic below, it appears that basal area per hectare has a number of entries that appear to be outliers when compared to the rest of the data. The closure data appears correct however the basal area values that are above 100 m² do not appear to conform to the general trend in this data set. You may observe a similar trend in your data set. Copy and paste the scatterplot into your error recording sheet.



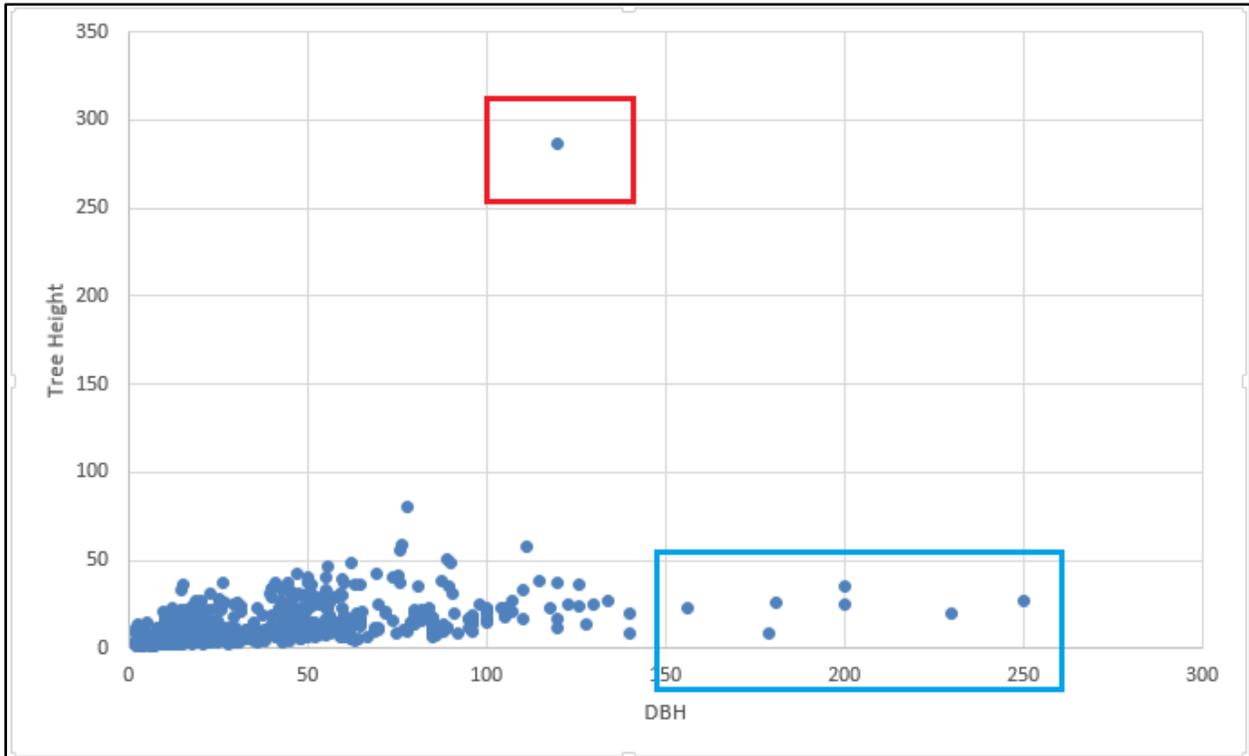
Based on the graphic above I now know that there are a number of basal area values in the data set that need to be checked. Head back to the spreadsheet and sort the basal area per hectare column from largest to smallest. Select all the values above 100, now move to the left in the spreadsheet to identify which clusters contain this information. Below we can see that the clusters with an unusually high basal area estimates are all located in Gbarpolu. If you have found similar patterns in your data set, please copy and paste the clusters into the error recording sheet such that additional investigation can be undertaken. As we know most teams are making use of a linear tape for dbh measurements, it may be that the field team who captured this data forgot to convert the linear circumference measurement to a diameter value. Be sure to keep a record of the values that you suspect of being outliers. Record their cluster and plot numbers in the error recording sheet.

	A	B	C	D	E			
1	cluster	cluster	plot_no	plot type	plot_enumeration date_year	plot_enumeration date_month	plot_enumeration date_day	plot enume
2	GP_CP_204			SP	2018	8	7	
3	GP_CP_204			EP	2018	8	6	
4	GP_CP_186			EP	2018	8	15	
5	GP_CP_205			CWP	2018	8	4	
6	GP_CP_205			WP	2018	8	4	
7	GP_CP_203			CSP	2018	8	20	
8	GP_CP_204			CSP	2018	8	7	
9	GP_CP_186			CWP	2018	8	16	
10	GP_CP_203			SP	2018	8	20	
11	GP_CP_205			SP	2018	8	3	
12	GP_CP_186			CSP	2018	8	17	
13	GP_CP_186			SP	2018	8	17	
14	GP_CP_203			CWP	2018	8	19	
15	GP_CP_204			WP	2018	8	8	
16	GP_CP_204			CWP	2018	8	8	
17	GP_CP_186			WP	2018	8	16	
18	GP_CP_205			EP	2018	8	2	
19	GP_CP_205			CSP	2018	8	3	
20	GP_CP_203			WP	2018	8	19	
21	GP_CP_203			EP	2018	8	18	

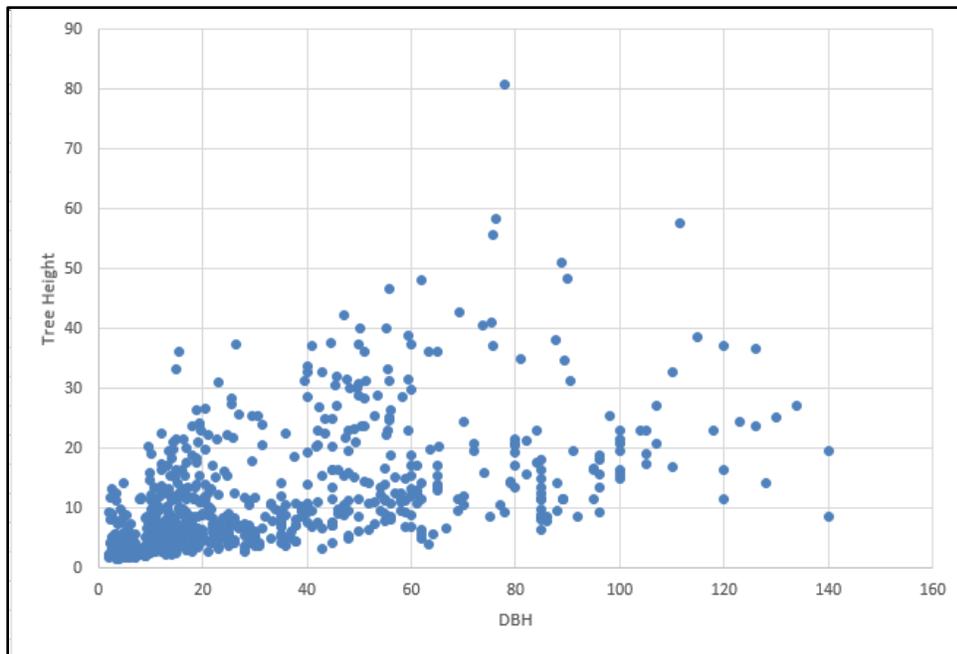
Diameter at breast height vs. Tree height – scatter plot

The relationship between diameter at breast height and tree height is well known throughout all forest ecological zones. The relationship is sometimes employed in Allometric equations used to estimate biomass and carbon stocks within forests. In the present data cleaning activities we will be using two methods to compare dbh to tree height. The first is graphical while the second is an advanced statistical method known as z-score analysis. Tree heights (column AB) and dbh (column T) data are both stored in the tree sheet of your data analysis file. Recall that only every third tree has its height measured, as such we must once again filter out the blanks in this column before we can prepare the graphic. As before select both columns and create a scatterplot using insert and recommended charts. From our histogram

analysis as well as the scatter plot above we know that there is an error in the tree height data (red box), we also know that some of the dbh data (blue box) may have been entered incorrectly. Carefully review the scatterplots you create for your study sites, there may be additional errors which require further analysis. Copy and paste the scatterplot into your error recording sheet.



In the graphic below I have removed trees above 100m and DBH with values over 150cm. The graphic shows that the general relationship between trees and dbh in this data set is reasonably good and displays characteristics that we expect in a forested environment.



Diameter at breast height vs. Tree height – z-score analysis

The proposed z-score analysis is used here to identify which data points, by diameter class (10-30 therefore 10cm to 250cm), are far from the others with respect to height/diameter ratio and diameter/height ratios. The scatterplot based visual interpretation above highlighted data points in the data set that differed significantly from the entire data set. The z-score analysis will now compare these two variables within each of the diameter classes present within the data set. Should diameter / tree height values differ significantly from those within the diameter class they will return residual values at either end of a normal distribution which may warrant further analysis? We will use the z-score analysis to identify these trees and highlight them for additional analysis.

A separate excel sheet is provided to facilitate this analysis (Liberia_zscore_and_residual_analysis_for_error_identification.xlsx). Before we start the analysis we will need to prepare our diameter and height data to match the format in this sheet. Below is a screen capture of the z-score analysis data format. The spreadsheet has pre-calculated columns where we will insert the data we want to analyse. Columns A to E below are the only columns we will be changing in this analysis. The rest are to be left untouched.

	A	B	C	D	E	F	G	H
1	Cluster	tree#	plot	height	diameter	diamter class	height/diameter	diameter/height
2	GP_CP_187	18	WP	19.5	140	13	0.139285714	7.179487179
3	GP_CP_204	54	EP	8.502075	140	13	0.06072911	16.46656772
4	GP_CP_205	33	WP	27.09883	134	12	0.202230094	4.944862451
5	GP_CP_203	21	SP	25.09211	130	12	0.193016243	5.180911115
6	GP_CP_204	45	CSP	14.10374	128	11	0.110185435	9.075609669
7	GP_CP_203	33	SP	36.70736	126	11	0.291328251	3.432554159
8	GP_CP_186	45	EP	23.72815	126	11	0.188318641	5.310148759
9	GP_CP_204	42	CWP	24.53895	123	11	0.199503653	5.012439534
10	GP_CP_203	27	SP	37.2037	120	11	0.310030835	3.225485622
11	GP_CP_186	36	CWP	16.5	120	11	0.1375	7.272727273
12	GP CP 204	51	CSP	11.5	120	11	0.095833333	10.43478261
13	GP CP 205	36	SP	22.95156	118	10	0.194504785	5.141261683

In the graphic above the tree and diameter data includes the cluster number, tree number and plot. We include this information such that we are able to identify the plots and clusters, as well as the trees that display characteristics which are abnormal for their respective diameter classes. In your data analysis worksheet, create a new sheet and call it zscore-prep.

GP_CP_186	42	CSP	17.50334529	105
GP_CP_205	30	EP	22.9222201	104
GP_CP_205	57	CSP	22.9222201	100
GP_CP_204	48	CSP	21.49102527	100
GP_CP_205	51	CWP	20.69912448	100
GP_CP_204	45	SP	19.5	100
GP_CP_204	33	CSP	16.5	100
GCM_CP_184	15	CSP	15.76469373	100
GP_CP_186	42	FP	15.00606066	100

tree | **zscore-prep** | plot | basal area per hectare | trees per hectare

Zscore-prep should be an empty sheet. We will now copy from the tree sheet the columns we need for the analysis, these will be copied into your zscore-prep sheet. From the tree sheet, copy and paste the cluster (column A), tree number (column C), plot type (column B), tree height (column AB), and tree dbh (column T). Be sure that you copy the entire column into the zscore-prep sheet and make sure that you have copied the columns in the correct order. If you have not copied them in the correct order then the analyses will be incorrect. Make sure that your zscore-prep sheet looks similar to the one below. It is vitally

important that the columns be in the correct order - cluster_cluster_plot_no, tree_no, plot_plot_type,ree_total_height, tree_dbh.

	A	B	C	D	E	F
1	cluster_cluster_plot_no	tree_no	plot_plot_type	tree_total_height	tree_dbh	
2	GP_CP_187	18	WP	19.5	140	
3	GP_CP_204	54	EP	8.502075382	140	
4	GP_CP_205	33	WP	27.09883264	134	
5	GP_CP_203	21	SP	25.09211162	130	
6	GP_CP_204	45	CSP	14.10373569	128	
7	GP_CP_203	33	SP	36.70735964	126	
8	GP_CP_186	45	EP	23.72814882	126	
9	GP_CP_204	42	CWP	24.53894938	123	
10	GP_CP_203	27	SP	37.20370017	120	

Once you have confirmed that your zscore-prep data sheet is correct, it is time to transfer this information to the Liberia_zscore_and_residual_analysis_for_error_identification.xlsx worksheet. Open this worksheet and navigate to the residual_analysis worksheet. Here you will see the randomly selected data used to prepare this manual. We will now copy the data from the zscore-prep sheet to the residual_analysis sheet and commence with the analysis. Select the data from zscore-prep, do not include the headings just the data. Be sure to select all of the information, in the example below the data set has 699 trees (699 rows).

	A	B	C	D	E	F	G	H
1	cluster	cluster	plot no	tree no	plot type	tree total height	tree dbh	
2	GP_CP_187	18	WP			19.5	140	
3	GP_CP_204	54	EP			8.502075382	140	
4	GP_CP_205	33	WP			27.09883264	134	
5	GP_CP_203	21	SP			25.09211162	130	
6	GP_CP_204	45	CSP			14.10373569	128	
7	GP_CP_203	33	SP			36.70735964	126	
8	GP_CP_186	45	EP			23.72814882	126	
9	GP_CP_204	42	CWP			24.53894938	123	
10	GP_CP_203	27	SP			37.20370017	120	
11	GP_CP_186	36	CWP			16.5	120	

Select the data as above and paste the columns into the residual_analysis worksheet, be sure to copy and paste starting at cell A2. Your data sheet should look like the one below. Note that the values will however differ.

	A	B	C	D	E	F	G	H	I
1	Cluster number	tree#	plot	height	diameter	diamter class	diameter range	diameter class	slope_diamy_height
2	GP_CP_187	18	WP	19.5	140	13	0	1	
3	GP_CP_204	54	EP	8.502075382	140	13	30	2	
4	GP_CP_205	33	WP	27.09883264	134	12	40	3	1.993301
5	GP_CP_203	21	SP	25.09211162	130	12	50	4	1.993301
6	GP_CP_204	45	CSP	14.10373569	128	11	60	5	0.0143900
7	GP_CP_203	33	SP	36.70735964	126	11	70	6	0.0143900
8	GP_CP_186	45	EP	23.72814882	126	11	80	7	0.0143900
9	GP_CP_204	42	CWP	24.53894938	123	11	90	8	0.0143900
10	GP_CP_203	27	SP	37.20370017	120	11	100	9	0.0143900
11	GP_CP_186	36	CWP	16.5	120	11	110	10	0.0143900
12	GP_CP_204	51	CSP	11.5	120	11	120	11	0.0143900

As soon as you copy the new data into the sheet the analysis will begin, you should see a progress message in the bottom of excel indicating that the analysis is underway and counting the progress. See below a

screen shot of this message. Depending on specs of your laptop, the analysis could take between 5 and 15 min, perhaps longer if you are using an older model laptop.

0.413578542	94.19412839	4.245157888
0.413578542	94.19412839	1.313596958
0.413578542	94.19412839	2.897798157
0.413578542	94.19412839	3.566863586
0.413578542	94.19412839	0.325733242
0.413578542	94.19412839	-3.674266758
0.413578542	94.19412839	-3.082355286

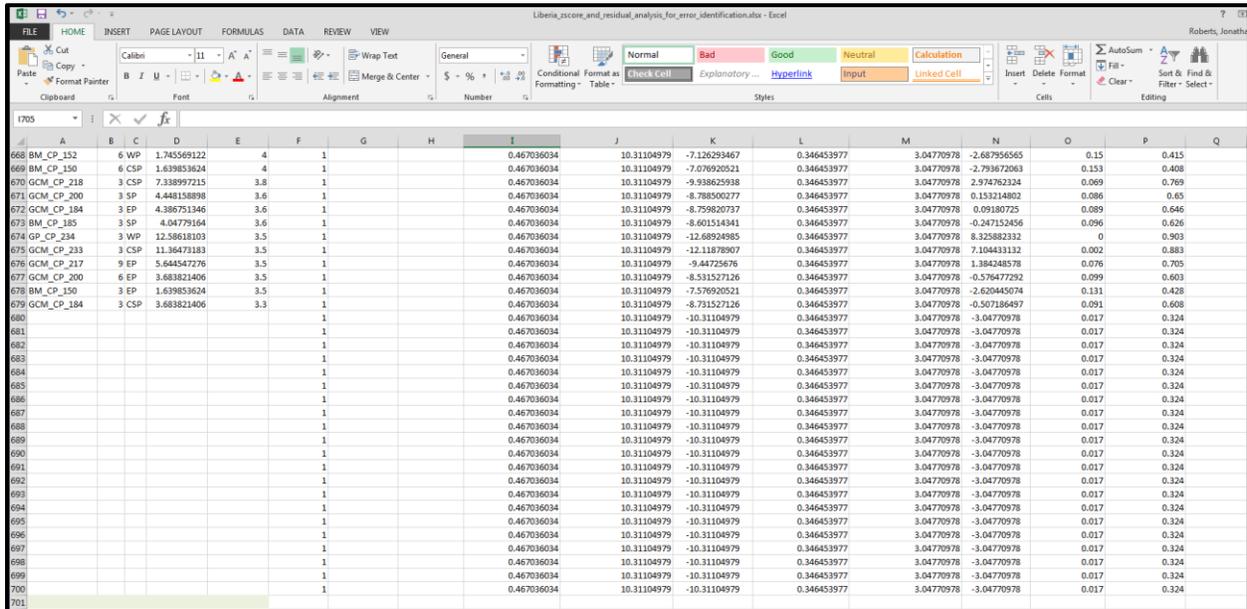
CALCULATING: (4 PROCESSOR(S)): 38%

When the analysis is complete it is important to check that the number of rows for your input data match the analysis rows on the residual analysis worksheet. To check this scroll down till the end of input data rows (the ones you just copied into the sheet). In the example below you will see that the analysis rows end at row 681 while the data runs to 700 rows (699 plus headings). All input data below row 682 are currently being excluded from the analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
674	GP_CP_234	3	WP	12.58618	3.5	1	3.596051723	0.278082763	0.612009518	2.295290213	0.491530029	1.581492258	6.070925534	-1.2755089	1	
675	GCM_CP_233	3	CSP	11.36473	3.5	1	3.247066237	0.307970311	0.612009518	2.295290213	0.491530029	1.581492258	5.360927233	-1.256610579	1	
676	GCM_CP_217	9	EP	5.644547	3.5	1	1.612727793	0.620067444	0.612009518	2.295290213	0.491530029	1.581492258	2.035925001	-1.059267133		
677	GCM_CP_200	6	EP	3.683821	3.5	1	1.052520402	0.950100348	0.612009518	2.295290213	0.491530029	1.581492258	0.8962034	-0.830582643		
678	BM_CP_150	3	EP	1.639854	3.5	1	0.468529607	2.134336839	0.612009518	2.295290213	0.491530029	1.581492258	-0.291904671	-0.101773103		
679	GCM_CP_184	3	CSP	3.683821	3.3	1	1.116309517	0.3958089	0.612009518	2.295290213	0.491530029	1.581492258	1.02598004	-0.884911897		
680	GCM_CP_218	12	SP	5.595813	3.2	1	1.748691632	0.571856113	0.612009518	2.295290213	0.491530029	1.581492258	2.312538495	-1.089751841		
681	GCM_CP_217	3	EP	5.4	3.2	1	1.6875	0.592592593	0.612009518	2.295290213	0.491530029	1.581492258	2.188046341	-1.076639871		
682	GP_CP_221	3	WP	8.038883	3	1										
683	GCM_CP_184	3	CWIP	3.413009	3	1										
684	GP_CP_205	6	WP	1.921623	3	1										
685	GP_CP_235	3	CWIP	5.139702	2.8	1										
686	GP_CP_235	6	CSP	13.15769	2.7	1										
687	GP_CP_234	3	EP	9.177281	2.7	1										
688	GCM_CP_184	3	SP	3.124598	2.6	1										
689	GCM_CP_217	6	EP	5.281121	2.5	1										
690	BM_CP_168	3	EP	3.507218	2.5	1										
691	GCM_CP_184	3	WP	2.497312	2.3	1										
692	BM_CP_185	3	CSP	2.15241	2.3	1										
693	GP_CP_235	3	EP	11.8923	2.2	1										
694	GP_CP_235	3	SP	8.042225	2.2	1										
695	GCM_CP_217	3	WP	4.3	2.2	1										
696	GCM_CP_218	3	WP	4.3	2.2	1										
697	GP_CP_235	3	CSP	9.42723	2	1										
698	BM_CP_185	6	WP	2.247984	2	1										
699	GP_CP_204	6	WP	2.192605	2	1										
700	GP_CP_223	3	EP	1.745569	2	1										

To remedy this problem we will need to extend the formulas in columns G through to O down to row 700. To extend the formulas please select a representation sample of rows from G through to O, two or three should suffice.

If on the other hand the analysis rows stretch beyond the data rows then you should see an error similar to below. You will notice that the results returned are full of errors. If this is the case delete the additional analysis rows and make sure that the both the analysis rows and the data rows match exactly. Make sure this is the case for both the residual analysis worksheet as well as the Z_score worksheet. Once all is correct and both the data and analysis columns have the same number of rows the analysis will rerun.



Once the analysis is complete we will identify those trees that differ significantly from the other trees within their respective diameter classes. Navigate to the Z_score sheet and scroll to the right until you find the column header 'strange value'. Those trees that do differ significantly within their diameter classes will have a value of 1 in this column. Filter the column to remove the blank cells. You should be left with a handful of trees which require additional analyses.

M	N	O	P	Q	R	S
Zscore height	Zscore diam	strange value	# of strange values	15		
3.332104173	-1.650807679	1				
-1.429476146	3.801582987	1				
3.16118282	-1.680074818	1				
-0.880142691	3.225847398	1				
-0.919434855	4.328061672	1				
-0.933979862	4.87857483	1				

Once the filter is complete, scroll to the left to record the cluster, tree number and plot of the identified trees. Copy and paste the cluster, plot and tree number into the error record sheet. This is the final data cleaning exercise that will be undertaken as part of this round of cleaning.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H
1	Cluster	tree#	plot	height	diameter	diamter class	height/diameter	diameter/height
48	GCM_CP_217	9	CWP	50.95459	88.8	7	0.57381299	1.742728061
229	GP_CP_203	18	EP	3.245296	43	3	0.075472	13.24994692
290	GP_CP_235	6	SP	25.5	30.51	2	0.835791544	1.196470588
310	GP_CP_186	15	CSP	3.831538	28	1	0.136840653	7.307769849
311	GP_CP_203	12	WP	3.107695	28	1	0.110989113	9.009892735
312	GP_CP_186	24	WP	2.839746	28	1	0.101419499	9.86003691
370	GP_CP_203	12	CSP	2.74664	21	1	0.130792382	7.645705259
650	GP_CP_236	3	SP	14.23814	4.9	1	2.90574209	0.344146166
674	GP_CP_234	3	WP	12.58618	3.5	1	3.596051723	0.278082763
675	GCM_CP_233	3	CSP	11.36473	3.5	1	3.247066237	0.307970311
686	GP_CP_235	6	CSP	13.15769	2.7	1	4.873219056	0.20520317
687	GP_CP_234	3	EP	9.177281	2.7	1	3.398992852	0.294204796
693	GP_CP_235	3	EP	11.8923	2.2	1	5.405593114	0.184993576
694	GP_CP_235	3	SP	8.042225	2.2	1	3.65555666	0.273556148
697	GP_CP_235	3	CSP	9.42723	2	1	4.713615095	0.212151391

Reporting and decisions regarding errors

All reporting associated with the data cleaning exercise will be made using the error recording sheet. A copy of this sheet is provided in Annex 1. Record all changes and or errors identified in the data cleaning process in the sheet and save it. This sheet will be used to update and or remove data based on the initial interpretation of the inventory data. The final cleaning will take place at a later time once the rest of the NFI supervision team have had time to review the results. The work undertaken here is only the beginning when it comes to the data cleaning process. In later iterations of the process we will implement additional cleaning activities relating to the work covered by Javier in the first data cleaning training exercise held in September.

Annex 1 – Error record sheets

Error recording sheet

Name of analyst:

Instructions

The following tables have been created to aid in recording the errors encountered during the data cleaning process. Data cleaning officers are encouraged to add additional rows to the tables if needs be. The information captured in these tables will be used to aid decision making with regards to corrective actions. Example errors are provided in grey text. This text should be removed as soon as the data cleaning officer is familiar with the cleaning procedure and recording of errors.

Step 1 Validation Report Analysis

Carbon related errors and corrections

Type of Error	Cluster	Plot name	Tree number	Proposed Action
Missing Coordinate	LF_CP_283	WP	6	Check team data for coordinate

Access route and photo assessment

Cluster number	Type of Error / Update	Proposed Action
GCP_CP_210	Missing photos and coordinates / none taken	Follow up with team leader

Reference photos and prominent structures

Cluster	Plot	Error Type	Proposed Action
NB_CP_099	EP	Missing reference photos	

Land use assessment

Cluster	Plot	Error Type	Proposed Action
NB_CP_099	EP	Missing land use data, land use class not specified	Flag for additional analysis and or speak to the team leader.

Harmonize non-timber forest products

Cluster	Plot	Error Type	Proposed Action
NB_CP_099	EP	Missing NTFPs	Flag for additional analysis and or speak to the team leader.

Tree species assessment

Cluster	Plot	Tree number	Error Type	Proposed Action
NB_CP_099	EP	15	Missing NTFPs	Flag for additional analysis and or speak to the team leader.

Step 2 – Quantitative Analysis of Biophysical Data

Single variables

Variable analyzed	Method	Results
Tree height	Histogram	Tree height > 100m

Two variables

Variable analyzed	Method	Results
DBH Vs Tree Height	Scatterplot	Several outliers, see excel spreadsheet

Z-score analysis

Cluster	Tree number	Plot