Using the Spatial Quality and Anomalies Diagnosis (SQUAD) Tool to Identify and Correct Data Anomalies in Master Facility Lists

Global Operational Guide

September 2018

Using the Spatial Quality and Anomalies Diagnosis (SQUAD) Tool to Identify and Correct Data Anomalies in Master Facility Lists

Global Operational Guide

John Spencer, MA, MEASURE Evaluation
Becky Wilkes, MS, MEASURE Evaluation

September 2018

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, North Carolina 27516
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. MS-18-146

ISBN: 978-1-64232-086-2

ACKNOWLEDGMENTS

The authors would like to thank the United States Agency for International Development (USAID) and the United States President's Emergency Plan for AIDS Relief (PEPFAR) for their support of this work.

We thank Jenny Mwanza, of the USAID- and PEPFAR-funded MEASURE Evaluation, Palladium, for her input. Thanks also to Jason Smith of MEASURE Evaluation, University of North Carolina at Chapel Hill (UNC) for managerial guidance throughout the development process. Thanks to Veronica Escamilla, Zaharani Kalungwa, and Beatus Kibiti of MEASURE Evaluation, UNC and Nigeria, for their help on an earlier version of this document. And we thank the knowledge management team at MEASURE Evaluation, UNC for editorial, design, and production services.

CONTENTS

ABBREVIATIONS

CSV	comma-separated values
GIS	geographic information system
GPS	global positioning system
km	kilometer
MFL	master facility list
PEPFAR	United States President’s Emergency Plan for AIDS Relief
SQUAD	Spatial Quality and Anomalies Diagnosis [Tool]
UNC	University of North Carolina
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

A master facility list (MFL) is a catalog of health facilities that helps Ministry of Health officials know the name of each facility, where each one is located, and other important information. The MFL facilitates service delivery planning and helps locate health facilities near the populations who need them. Ensuring that the data in the MFL are correct—and identifying and troubleshooting errors—are critical but problematic, because the database has thousands of records and the use of a geographic information system (GIS) is required to check the location of each facility.

MEASURE Evaluation, funded by the United States Agency for International Development (USAID) and the United States President's Emergency Plan for AIDS Relief (PEPFAR), created the Spatial Quality and Anomalies Diagnosis (SQUAD) Tool to assess data quality issues in these large data sets. The tool performs a data quality check on the data set, rapidly and automatically looking for anomalies in the data that need to be investigated.

This document is a resource for developing an action plan to improve data in an MFL through the use of the SQUAD Tool. It provides detailed instructions on how to prepare the data for use with the tool and how to run the SQUAD Tool with a typical country MFL. It also provides recommendations on specific steps to follow to prioritize actions and resolve errors found in the output list provided by the tool.

The status of health facilities is always changing as new health facilities come into existence and others close or move. Because of the dynamic nature of the MFL, it is recommended that the SQUAD Tool be regularly used, in line with the local MFL governance policy.

INTRODUCTION

The master facility list (MFL) is a catalog of a country's health facilities and their corresponding global positioning system (GPS) coordinates. These kinds of geographic data coordinates are relatively easy to collect using a smartphone, tablet, or GPS receiver. When the coordinates are stored in a table or spreadsheet, they can be shown on a map. The purpose of an MFL is simple: the coordinates allow a country to know where its health facilities are located, which facilitates service delivery planning and ensures that facilities are located near the populations who need them. This enables resources to be used more wisely.

Ensuring that the data in the MFL are correct and identifying and troubleshooting any errors are critical but problematic, because of the effort required to review thousands of facility records to identify potential errors in the data set. One way to ensure that the coordinates are accurate is to plot the locations in a geographic information system (GIS) and then carefully examine the "attribute" data associated with the points. Attribute data consist of such information as the district name, facility code, facility type, funding source, and contact information. The resultant data table can be quite large, involving hundreds or thousands of points. Reviewing each point location for accuracy and correctness in such a large database can be time-consuming, tedious, and challenging, because of the complex relationship between the spatial characteristics (i.e., point locations) and the associated attributes of the data points being described.

Because of these unique data characteristics, ensuring the quality of geographic data typically requires the skills of an experienced GIS technician, which has been a barrier to a country's ability to verify the data and subsequently make use of them. A solution for staff with minimal training in GIS was therefore needed.

To that end, MEASURE Evaluation developed the Spatial Quality and Anomalies Diagnosis (SQUAD) Tool for assessing data quality issues in these large data sets. The tool performs a data quality check on large data sets, rapidly and automatically looking for anomalies in the data that need to be investigated.

Importance of Data Quality

Having high-quality data for the MFL is not just valuable but imperative. If a facility is plotted in the middle of a lake, or on the other side of the world, concern about data quality is warranted. If the data are not or cannot be trusted, they likely will not be used. And if they are not used, decisions will not be made in an informed way. Knowing where facilities are located can give program planners a better understanding of demand and need for health services. It also helps prospective clients know what services are available to them. For example, when the MFL is available online, clients can look up the nearest healthcare site. Incorrect data lead to confusion that can interfere with the adequate provision of services.

Purpose of This Document

This document is a resource for developing an action plan to improve data in an MFL through the use of the SQUAD Tool, developed by MEASURE Evaluation. Instructions on how to run the tool with a generic data set are provided. Recommendations on specific steps to follow to prioritize actions and resolve errors found in the output list provided by the SQUAD Tool are also given.

Recommendation for Using the SQUAD Tool for the MFL

The status of health facilities is always changing as new health facilities come into existence and others close or move. Because of the dynamic nature of MFLs, it is recommended that the SQUAD Tool be regularly used, in line with local MFL governance policy.

How to Use This Document

This document leads staff tasked with identifying and fixing errors in an MFL through the following steps:

Step 1. Prepare the data for use with the SQUAD Tool.

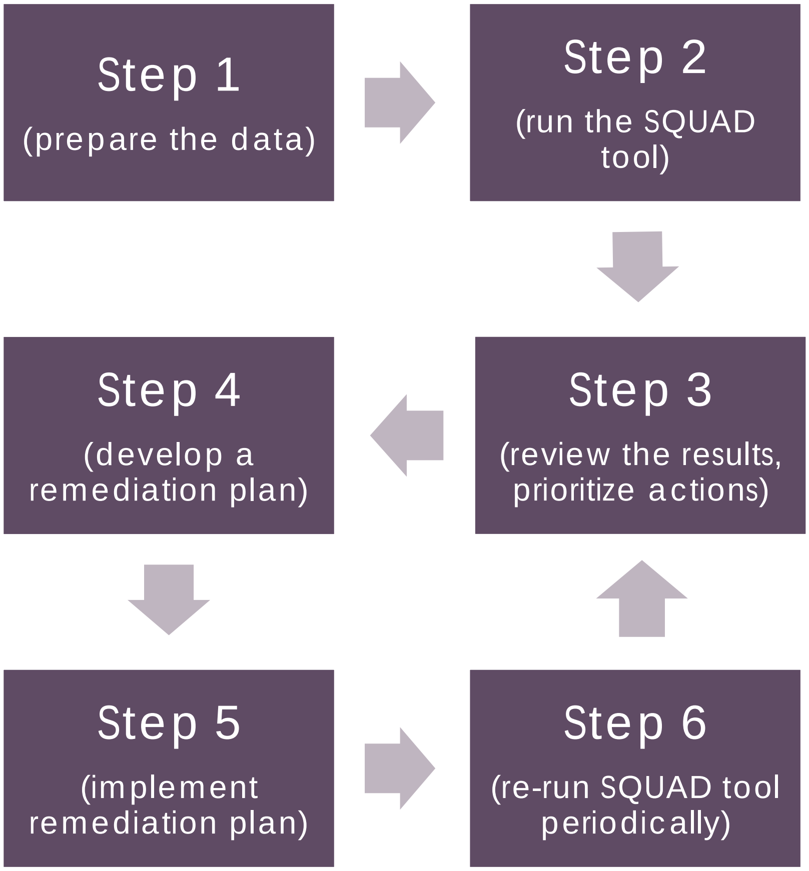
Step 2. Run the SQUAD Tool.

Step 3. Review the results of the SQUAD Tool analysis to prioritize actions.

Step 4. Develop a remediation plan.

Step 5. Implement the remediation plan.

Step 6. Periodically rerun the SQUAD Tool and repeat Steps 1 to 5.


Step 1
(prepare the data)
Step 2
(run the SQUAD tool)
Step 3
(review the results, prioritize actions)
Step 4
(develop a remediation plan)
Step 5
(implement remediation plan)
Step 6
(re-run SQUAD tool periodically)

Characteristics of MFL Data

The **spatial domain** of an MFL consists of data associated with facility locations. Examples are latitude/longitude or the name of the city, district, or region in which a facility is located. There may be other attributes associated with the data stored in an MFL, but they are not the tool's area of focus.

SQUAD TOOL OVERVIEW

The SQUAD Tool provides a framework for assessing the overall quality of a spatial database and can be run by someone with minimal GIS experience. The tool works in the spatial domain and provides a consistent, clear assessment of potential problems with a given spatial database, by looking for anomalies in the data that can indicate data quality issues. After the potential anomalies have been identified, the records in question can then be examined in detail, and any errors can be corrected. The tool offers a clear picture of the problems that may be associated with a given geographic data set, enabling the prioritization of solutions to improve the data.

The tool checks for six spatial anomalies that are based on common errors in point-location data sets. This information can be used to prioritize the type and extent of investigation needed for records that may have problems. The spatial domain anomalies the tool checks for are:

1. Missing coordinates
2. Truncated coordinates (lack of adequate precision)
3. Duplicate coordinates for distinct records
4. Duplicate key attributes (two identical names, but plotting in different locations)
5. Coordinate not located in expected district (but falling within 2 kilometers of its border)
6. Coordinate not located in expected district (but greater than 2 kilometers from its border)

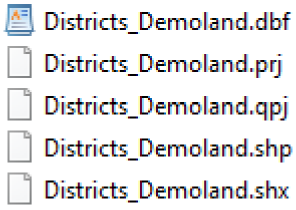
Note that the presence of an anomaly may not mean that there are errors. For example, two different sites in two different regions could actually have the same name but would be reported as an anomaly. These kinds of anomalies merit examination to ensure that they are not errors.

Although a further assessment of data quality will likely be required, the process described here is a starting point for examining a large data set in an efficient manner. Data quality assessments should be conducted on a regular basis and other errors can be caught through these assessments.

SQUAD TOOL REQUIREMENTS

Data Requirements

To run the SQUAD Tool, two data sets, in the form of *shapefiles*, are required. A shapefile is a geographically formatted data set that contains geometry and attribute data; it works with most GIS programs. (NOTE: In a desktop file system, a "shapefile" will appear as a **set** of files with the same names but different extensions, such as .shp, .shx, and .dbf; however, in a GIS program, they will appear as a single file with the .shp extension.



Districts_Demoland.dbf
Districts_Demoland.prj
Districts_Demoland.qpj
Districts_Demoland.shp
Districts_Demoland.shx

Shapefiles can be tricky to move around, as all the parts in the set are needed; it's generally best to zip a set of them together when sharing them or moving them, or to move them or rename them by using a GIS program.

The two data sets needed are:

A geographic file (shapefile) consisting of point locations. This is the MFL data file, including a list of all facilities, the district in which each facility is located, and the geographic coordinates for each facility. (NOTE: The instructions below include information on how to convert from a spreadsheet to a shapefile format in QGIS: a popular, free, and open-source GIS program that is available for different operating systems, such as Windows, Mac, and Linux.)

A geographic file (shapefile) consisting of polygons that represent administrative units. This is the administrative boundaries file for the country of interest. It must be an authoritative file—that is, accurate and up-to-date.

Software Requirements

QGIS and SQUAD Plug-In

The SQUAD Tool is a QGIS plug-in that assesses the quality of large spatial data sets. These operational guidelines use QGIS Version 3.0. Although this guide provides a step-by-step overview of running the tool, users can benefit from general familiarity with the basic operation of QGIS. The QGIS site (www.qgis.org) has links to tutorials and other online resources that can help the user learn the software.

Excel

Excel is a valuable tool for the data quality review exercise; it is useful for data cleaning prior to analysis and for the display of the data post-analysis. The results of the SQUAD Tool analysis can be exported to Excel, making it possible to produce data visualizations and dashboards that can facilitate the use of the results.

If you wish to follow along with the steps outlined below, download the sample files located here:

https://github.com/andre-ws/squad-plugin/blob/master/data/Demoland.zip

STEPS TO RUN THE SQUAD TOOL

Before you run the SQUAD Tool, you must make sure that your data are properly structured. (See the Data Requirements section above.) The data must be in a standard format before the SQUAD Tool can be used.

Step 1A. Prepare the data for use with the SQUAD Tool: Ensure that the facility site data file is properly structured.

Do you have a shapefile containing the site data?

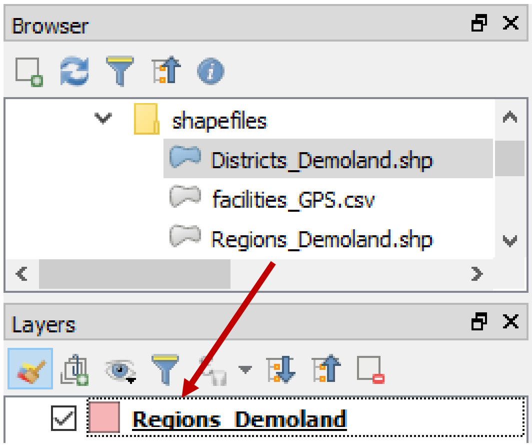
If you do, then proceed to Step 1B. If you do not, then follow these steps.

Do you have a CSV file?

CSV stands for "comma-separated values." If your data are in another format, such as an Excel file, you need to open that file and save it as a .csv file. **The CSV file needs to have one column that contains latitude (Y) values (in decimal degrees*) and one column that contains longitude (X) values. (See the *Note below for more information.)**

Use a CSV of the site file to create a shapefile in QGIS

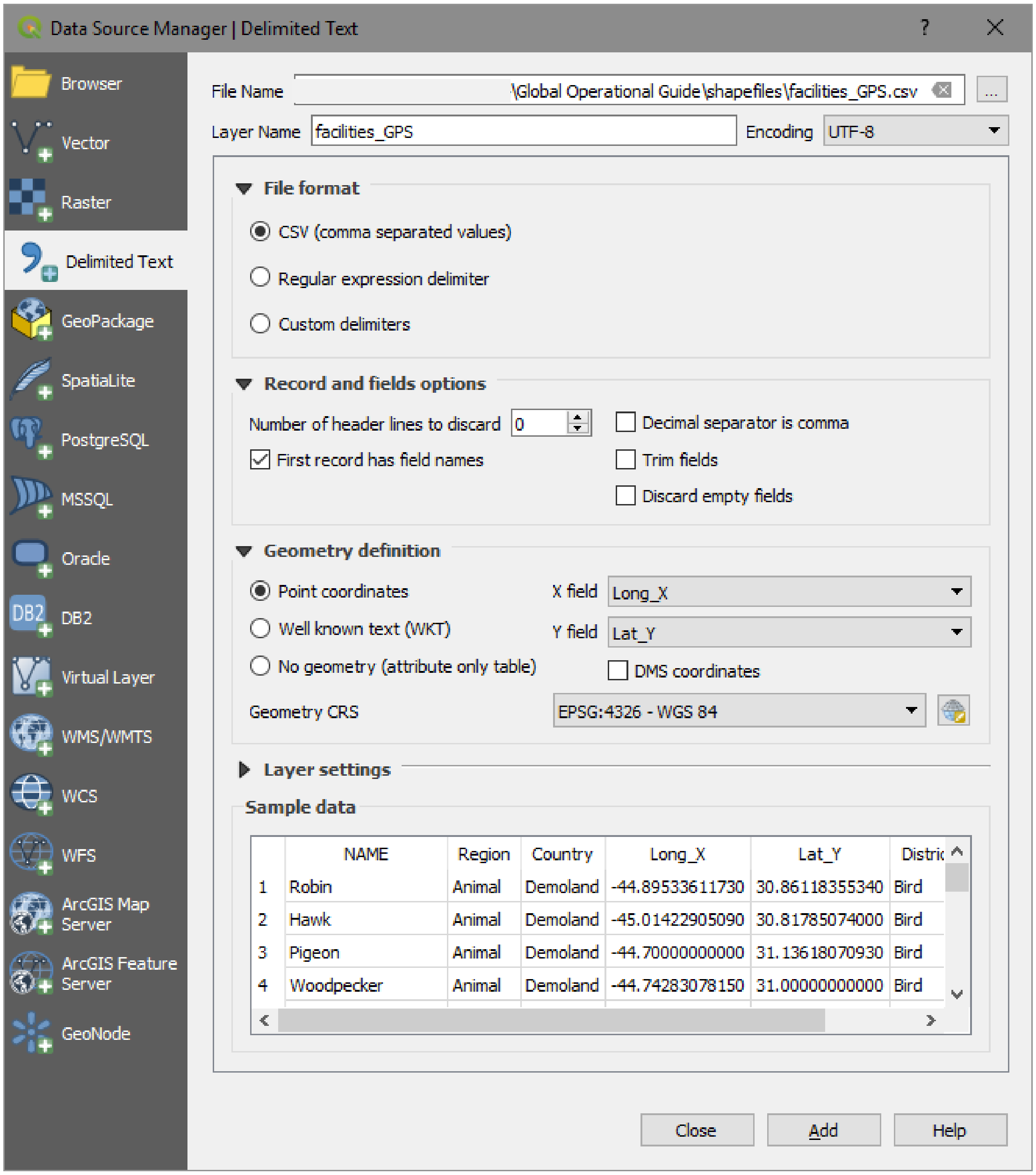
(A) First, open QGIS and load in a map of your country by dragging the name of a shapefile (one containing administrative units) in the Browser panel down into the Layers panel. (For this illustration, we have used *Regions_Demoland.shp*.) You should see the map load in the main map window. This will give you a base map, enabling you to see whether your site coordinates load from the spreadsheet into the correct location.



(B) Now choose **Layer>Add Layer>Add Delimited Text Layer…** A dialogue box will open, as shown below.

(C) You then need to use the longitude/latitude (X/Y) values in your spreadsheet to plot the points. Under **File Name**, choose the *facilities_GPS.csv* file. Check that the first record has field names and that the X field is the "longitude" column and the Y field is the "latitude" column. You can check the sample data window at the bottom to make sure that the data are being read correctly. The dialogue box should look like this:

***NOTE:** Some GPS receiver units can be set to collect data in other formats, such as "degrees-minutes-seconds" or even "UTM (meters)." If this is the case for your data, they need to be converted to "decimal degrees" by a qualified GIS technician. If the latitude and longitude were stored together in the same column or field (i.e., "23.78236 E, 3.98467 N"), they need to be placed in separate fields by a data technician. Last, if the coordinates are south or west of the equator, make sure that they are preceded by a minus (-) sign.



File Name \Global Operational Guide\shapefiles\facilities_GPS.csv

Layer Name facilities_GPS

Encoding UTF-8

File format

● CSV (comma separated values)

○ Regular expression delimiter

○ Custom delimiters

Record and fields options

Number of header lines to discard 0

☑ First record has field names

□ Decimal separator is comma

□ Trim fields

□ Discard empty fields

Geometry definition

● Point coordinates

○ Well known text (WKT)

○ No geometry (attribute only table)

X field Long_X

Y field Lat_Y

□ DMS coordinates

Geometry CRS EPSG:4326 - WGS 84

Layer settings

Sample data

	NAME	Region	Country	Long_X	Lat_Y	Distric
1	Robin	Animal	Demoland	-44.89533611730	30.86118355340	Bird
2	Hawk	Animal	Demoland	-45.01422905090	30.81785074000	Bird
3	Pigeon	Animal	Demoland	-44.70000000000	31.13618070930	Bird
4	Woodpecker	Animal	Demoland	-44.74283078150	31.00000000000	Bird

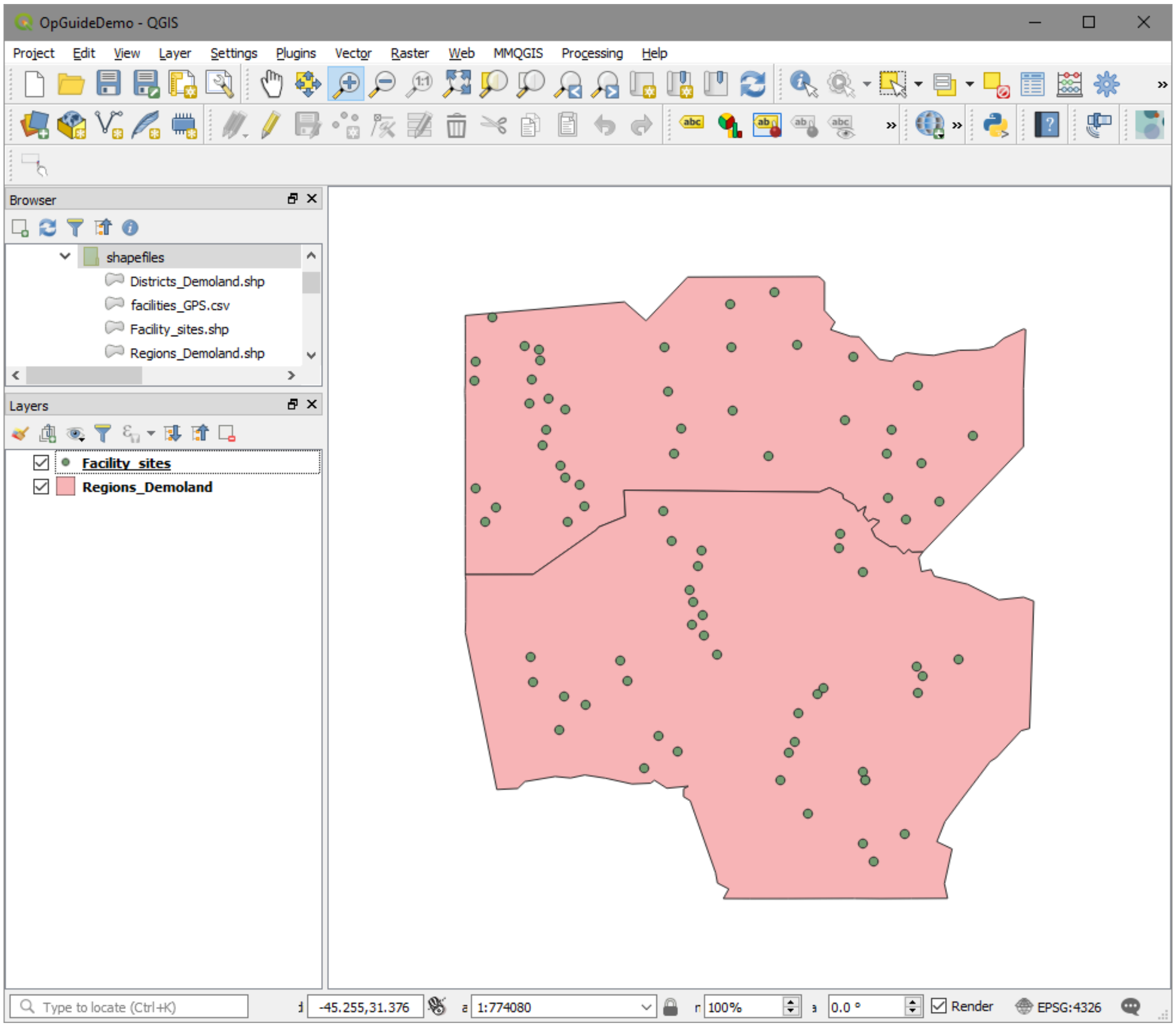
(D) Click the "Add" button at the bottom of the window and then click "Close." Your points should plot on the map.

(E) Next you need to save this temporary file to a Shapefile that you can use for your analysis. To do this, right-click on the name of the points in the Layers panel and select **Export>Save Features As…**

(F) Under **Format**, choose ESRI Shapefile, and then **browse** to your desired location and give the file a name, such as "*Facility_sites*." Keep "**Add saved file to map**" checked, and keep all the fields selected, plus all the other defaults. Then choose **OK**. After a moment, the new shapefile will appear on the map and its name will be in the Layers Panel.

(G) OPTIONAL: If you would like to change the appearance of your map, right-click on any of the files in the Layers panel, choose "Properties," and click on the "Symbology" tab to set colors and symbols. Then right-click on any of the files in the Layers panel and choose "Open Attribute Table" to view all the data fields for each shapefile.

Here is a sample finished screen shot:


OpGuideDemo - QGIS
Project Edit View Layer Settings Plugins Vector Raster Web MMQGIS Processing Help
Browser
Layers
Type to locate (Ctrl+K)
Render

Browser:
- shapefiles
  - Districts_Demoland.shp
  - facilities_GPS.csv
  - Facility_sites.shp
  - Regions_Demoland.shp

Layers:
- ☑ ● Facility_sites
- ☑ 🟥 Regions_Demoland

-45.255,31.376 | 1:774080 | 100% | 0.0 ° | ☑ Render | EPSG:4326

Step 1B. Prepare the data for use with the SQUAD Tool: Ensure that there is a unique ID for each district.

Once you have the two required shapefiles listed in the Data Requirements section on a previous page, the first task in preparing the data is to ensure that the administrative boundaries file to be used as the input has a field with a unique code for the district that can be joined to the MFL's district field. In other words, *both the MFL file and the district file should have a field for district that uses the same unique ID*.

If the administrative boundaries file does not contain unique IDs for each area, it is possible that you will need to use Excel or QGIS to create such a file.

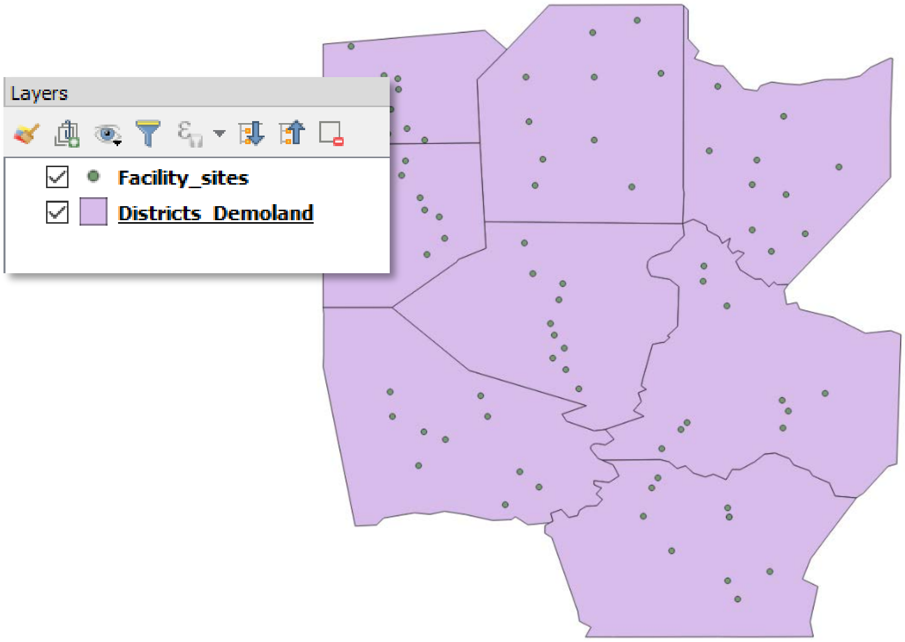
Once the two files needed for input—the district file and the MFL facility list file—both contain matching unique codes, the data are now ready to run the SQUAD Tool.

Step 2. Run the SQUAD Tool to discover anomalies in the spatial domain.

This section assumes you have formatted your data for use with the tool, as described and shown in the previous section, and you are ready to begin working with the SQUAD Tool. The two files we will use as input with the tool are called:

- Districts_Demoland.shp
- Facility_sites.shp

1. If you have not already done so, **Launch** QGIS.
2. Click on **Project>New**. Browse to the **Districts_Demoland** shapefile in the Browser Panel and drag it down to the Layers Panel. Do the same with the **Facility_sites** shapefile.
3. Make sure that the **Facility_sites** layer appears above the **Districts_Demoland** layer in the Layers Panel and that both are visible on the screen in the main map window.



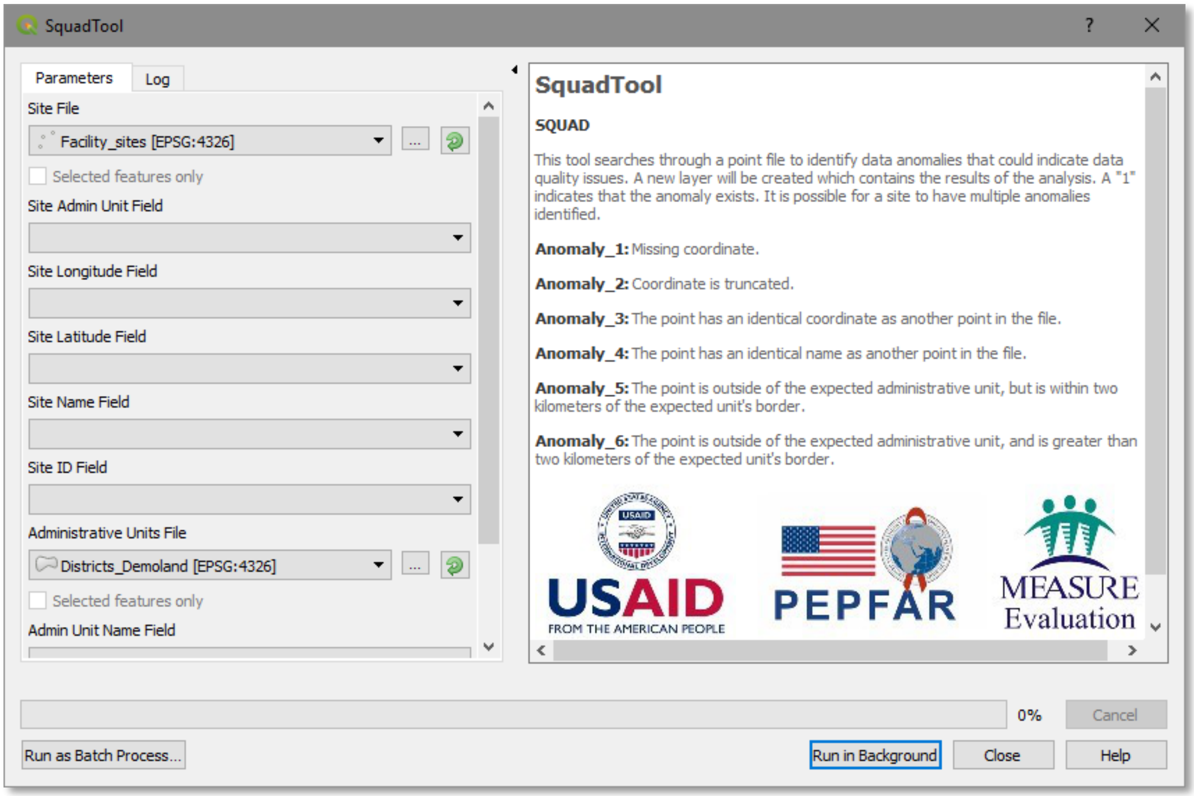
4. Go to **Plugins>Manage and Install Plugins…** Click on the **All** tab, and in the search box **type** SQUAD.

    a. For Version 2.18, click on SQUAD Tool.

    b. For Version 3.0, click on SQUAD Tool v3.

    c. **Click** the "**Install Plugin**" button, then **Close** the dialogue box.

5. Look for the SQUAD Tool icon in the toolbar and click on it to activate the tool. (Alternatively, go to **Plugins>SQUAD>SQUAD plugin**.) This will start the tool. You should see a dialogue box like this:



Site File

Facility_sites [EPSG:4326]

□ Selected features only

Site Admin Unit Field

Site Longitude Field

Site Latitude Field

Site Name Field

Site ID Field

Administrative Units File

Districts_Demoland [EPSG:4326]

□ Selected features only

Admin Unit Name Field

SquadTool

SQUAD

This tool searches through a point file to identify data anomalies that could indicate data quality issues. A new layer will be created which contains the results of the analysis. A "1" indicates that the anomaly exists. It is possible for a site to have multiple anomalies identified.

Anomaly_1: Missing coordinate.

Anomaly_2: Coordinate is truncated.

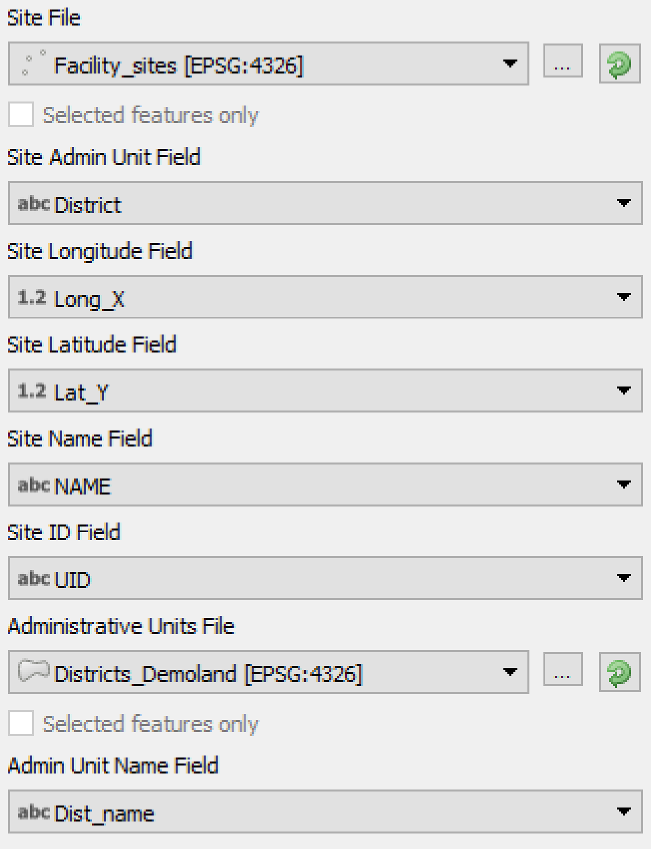
Anomaly_3: The point has an identical coordinate as another point in the file.

Anomaly_4: The point has an identical name as another point in the file.

Anomaly_5: The point is outside of the expected administrative unit, but is within two kilometers of the expected unit's border.

Anomaly_6: The point is outside of the expected administrative unit, and is greater than two kilometers of the expected unit's border.

6. Fill in the parameters needed to check for anomalies in your data set. For the data sets for Demoland, choose the following parameters:


Site File

Facility_sites [EPSG:4326]

□ Selected features only

Site Admin Unit Field

abc District

Site Longitude Field

1.2 Long_X

Site Latitude Field

1.2 Lat_Y

Site Name Field

abc NAME

Site ID Field

abc UID

Administrative Units File

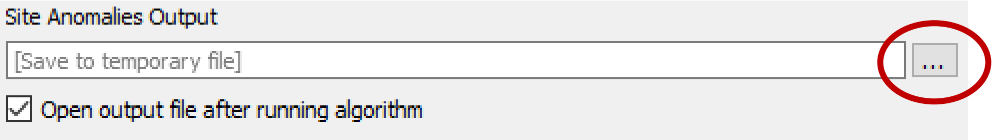
Districts_Demoland [EPSG:4326]

□ Selected features only

Admin Unit Name Field

abc Dist_name

7. In the bottom box, under "**Site Anomalies Output**," click the **Browse** button:

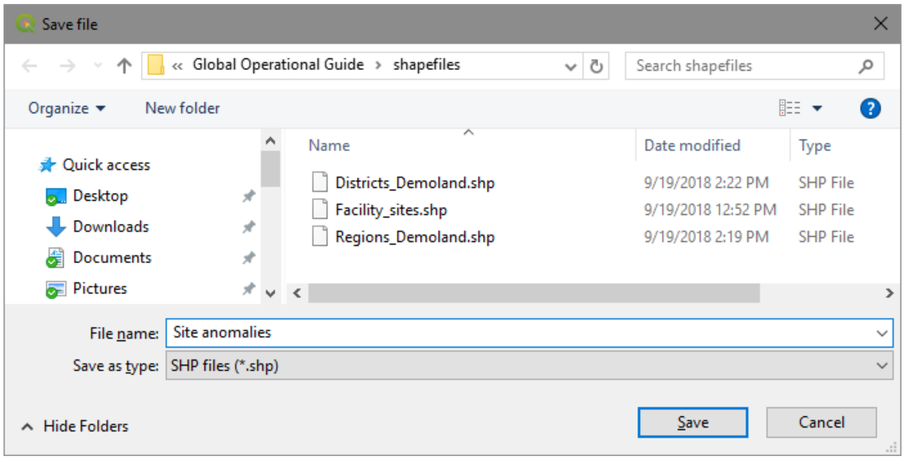

Site Anomalies Output

[Save to temporary file]

☑ Open output file after running algorithm

….and chose "Save to file."

8. In the resulting dialogue box, you should choose to save the file as either a new **Shapefile** (SHP) or as an **Excel spreadsheet** (XLSX). You may put this file in the location of your choosing. **NOTE**: If you


Save file

« Global Operational Guide › shapefiles

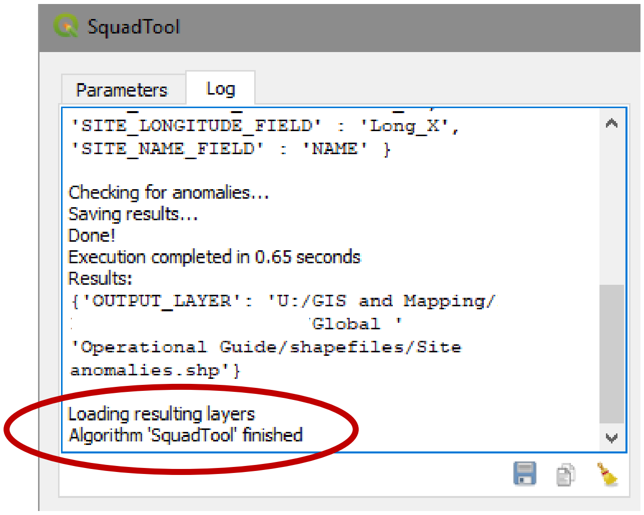
Name	Date modified	Type
Districts_Demoland.shp	9/19/2018 2:22 PM	SHP File
Facility_sites.shp	9/19/2018 12:52 PM	SHP File
Regions_Demoland.shp	9/19/2018 2:19 PM	SHP File

File name: Site anomalies

Save as type: SHP files (*.shp)

do not save to a file, a temporary output layer will be generated. You can then save this output by right-clicking on it in the Layers Panel and choosing "Save As." **If you do not do either of these things, then when you quit QGIS, the information generated by the tool will be lost.**

9. Click "**Save,**" then click on the "**Run in Background**" button in the main SQUAD Tool dialogue box. The Log tab will appear and show you a progress bar while the tool is running. After several seconds, once you get a message in the Log window that the algorithm is finished, you may click the **Close** button.

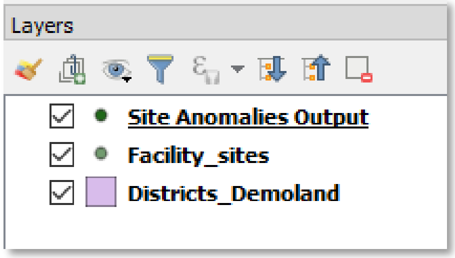


```
'SITE_LONGITUDE_FIELD' : 'Long_X',
'SITE_NAME_FIELD' : 'NAME' }

Checking for anomalies…
Saving results…
Done!
Execution completed in 0.65 seconds
Results:
{'OUTPUT_LAYER': 'U:/GIS and Mapping/
                        Global '
'Operational Guide/shapefiles/Site
anomalies.shp'}

Loading resulting layers
Algorithm 'SquadTool' finished
```

10. Your new output layer will display in the **Layers Panel** with the name "**Site Anomalies Output.**" (You may **double-click** it to see and edit its name, source file, and source file location.)

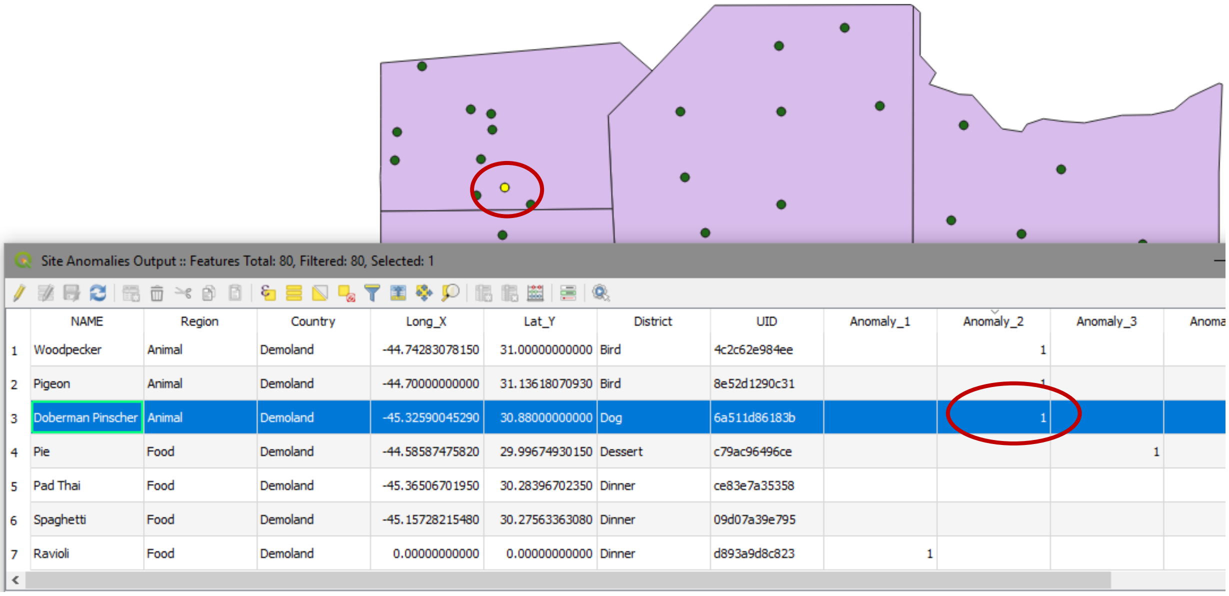


11. **Right-click** on the new layer and choose "**Open Attribute Table**" from the drop-down menu to view the contents of your newly coded file. If you saved your output as a shapefile, you can select locations in the table to highlight and view them on the map.

Steps 3 and 4. Review the results of the SQUAD Tool analysis to prioritize actions, then develop a remediation plan.

Data Quality Score

A data quality score can be helpful for prioritizing action in specific areas and monitoring progress in the remediation of errors. There are many ways to calculate a data quality score. The approach described below is simple. It provides a snapshot of the data quality and can be calculated at multiple levels (region, district, individual facility).



Site Anomalies Output :: Features Total: 80, Filtered: 80, Selected: 1

	NAME	Region	Country	Long_X	Lat_Y	District	UID	Anomaly_1	Anomaly_2	Anomaly_3	Anoma
1	Woodpecker	Animal	Demoland	-44.74283078150	31.00000000000	Bird	4c2c62e984ee		1		
2	Pigeon	Animal	Demoland	-44.70000000000	31.13618070930	Bird	8e52d1290c31		1		
3	Doberman Pinscher	Animal	Demoland	-45.32590045290	30.88000000000	Dog	6a511d86183b		1		
4	Pie	Food	Demoland	-44.58587475820	29.99674930150	Dessert	c79ac96496ce			1	
5	Pad Thai	Food	Demoland	-45.36506701950	30.28396702350	Dinner	ce83e7a35358				
6	Spaghetti	Food	Demoland	-45.15728215480	30.27563363080	Dinner	09d07a39e795				
7	Ravioli	Food	Demoland	0.00000000000	0.00000000000	Dinner	d893a9d8c823	1			

How to Calculate

If a facility has any anomaly, a data quality score of 0 is assigned. If a facility has no anomaly, then the facility receives a score of 1. To calculate an aggregate data quality score for a district or region, the sum of the data quality scores for every facility in the area of interest is divided by the number of facilities in the area. The result shows the percentage of facilities in the area that are anomaly-free.

Strengths and Limitations

This approach has some limitations. It treats any facility with an anomaly equally. Some anomalies are easier to resolve than others. For example, some anomalies may be resolved with a phone call to a site, whereas others may require visiting the site to collect a GPS coordinate. It may be preferable to differentiate between anomalies on this basis.

However, the advantage of this approach is that it is easy to calculate, and it provides a clear metric of the data quality.

Prioritizing Action Based on the Data Quality Score

1. Resolve what can be done quickly/easily (Anomaly 3 [duplicate coordinates] and 4 [duplicate facility names]).

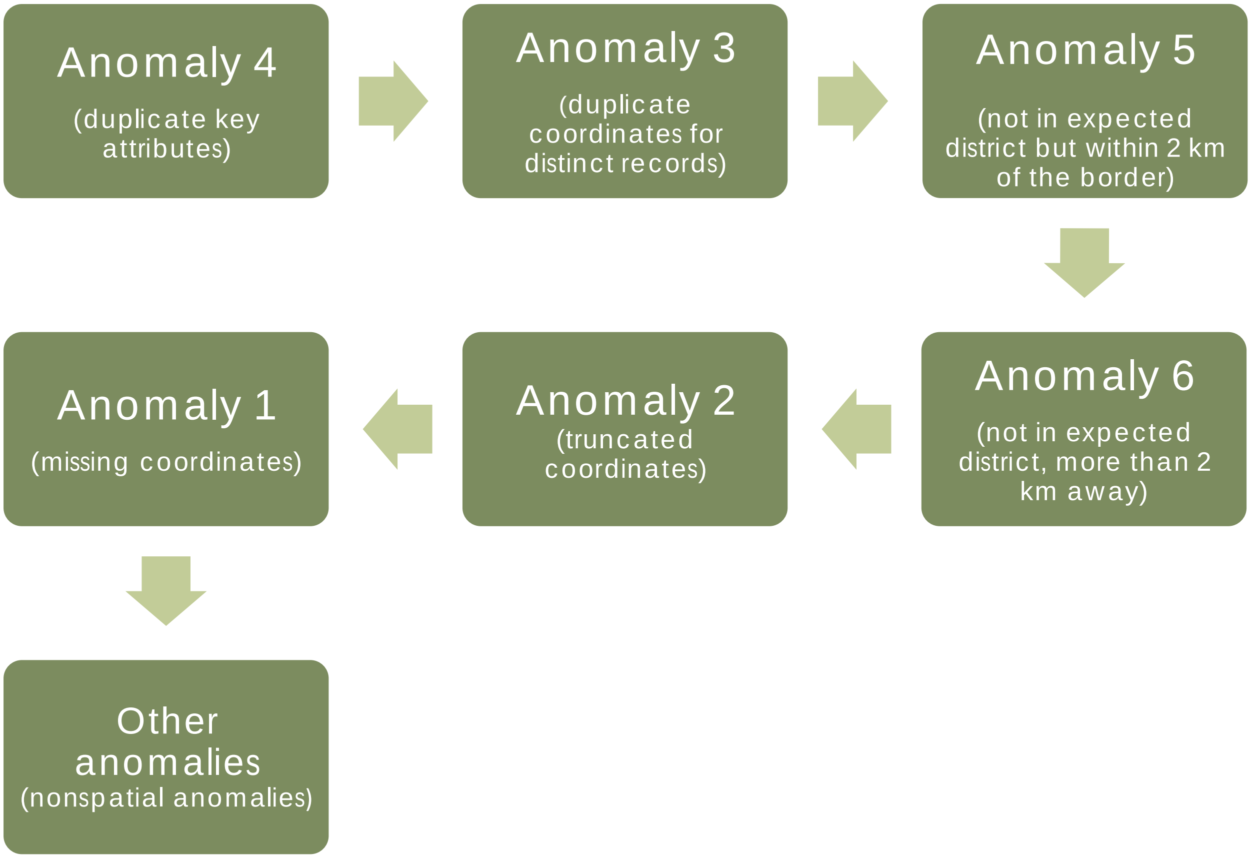
2. Each region can prioritize districts based on the data quality score and address the data quality in high-priority districts first.

Taking Corrective Action

Corrective action for the MFL data depends on the type of anomaly that has been identified. The following section provides an overview of the corrective action(s) to be taken for each type of anomaly. A flowchart shows the workflow for addressing the anomalies.

Many of the anomalies that exist in an MFL are related. For example, records with duplicate facility names could be instances where a facility's record is duplicated: once with a coordinate and another time without a coordinate. Finding duplications and deleting ones without a coordinate will address two records: one with an Anomaly 4 (duplicate facility names) and one with Anomaly 1 (missing coordinates).

Anomalies can be addressed in any order. However, the suggested workflow in the figure below introduces efficiencies in the process. In this section, a flowchart is provided for each anomaly to be addressed.


Anomaly 4
(duplicate key attributes)
Anomaly 3
(duplicate coordinates for distinct records)
Anomaly 5
(not in expected district but within 2 km of the border)
Anomaly 6
(not in expected district, more than 2 km away)
Anomaly 2
(truncated coordinates)
Anomaly 1
(missing coordinates)
Other anomalies
(nonspatial anomalies)

By beginning with Anomaly 4, records that are missing coordinates can be eliminated, thereby reducing the effort required to address Anomaly 1. Because Anomaly 1 is likely to be the most time-intensive anomaly to resolve, it is important to eliminate as many Anomaly 1 records as possible by addressing the other anomalies.

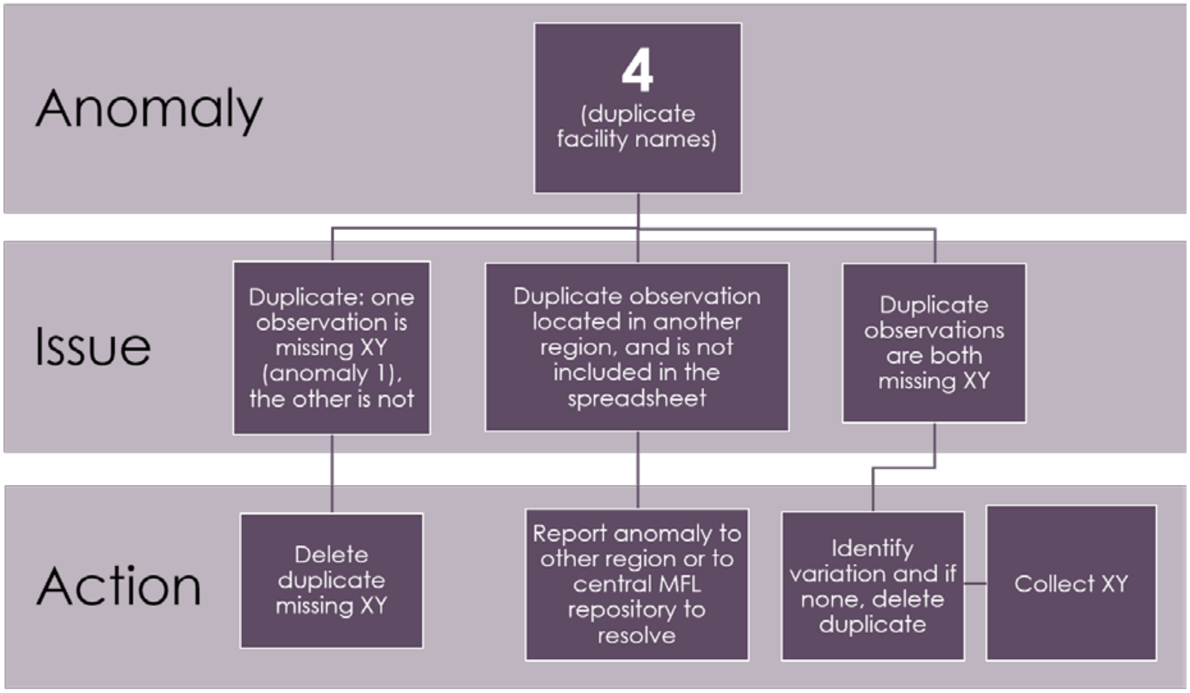
For each anomaly, there is an issue at the root of the anomaly and then an associated action.

Anomaly 4

With Anomaly 4 (duplicate facility names), it is necessary to determine why there are duplicate facility names. In some cases, complete records could be duplicated. In other instances, a duplicate could have matching attributes but otherwise be missing a latitude/longitude coordinate. Such records would also have an Anomaly 1 (missing coordinate). By removing the record with the missing coordinate, Anomaly 4 and Anomaly 1 would be resolved.

It is important to remember that if the technician remediating the anomalies is working on a subset of a larger national database, the facility with a duplicate name may not be in the subset. In other words, if the SQUAD Tool is run on a national database and finds two facilities in two different districts with the same name, a technician who has access only to one district's data would not see the facility from the other district.

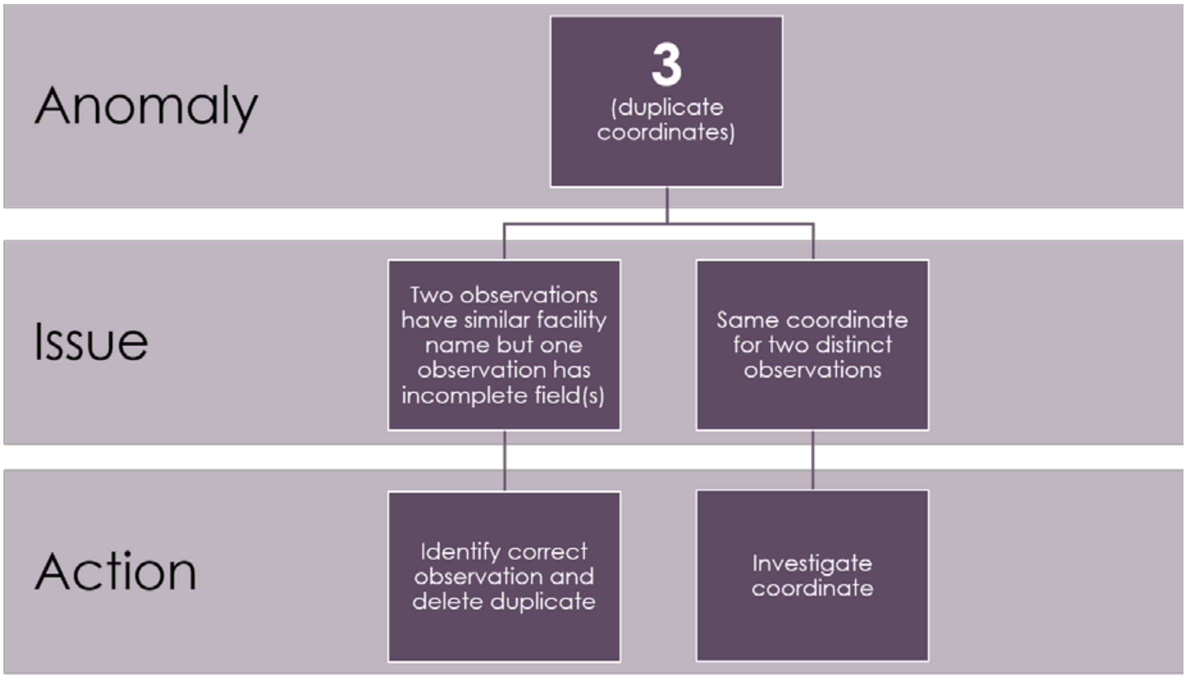
The following flowchart shows a workflow for addressing Anomaly 4 records.


Anomaly
4
(duplicate facility names)
Issue
Duplicate: one observation is missing XY (anomaly 1), the other is not
Duplicate observation located in another region, and is not included in the spreadsheet
Duplicate observations are both missing XY
Action
Delete duplicate missing XY
Report anomaly to other region or to central MFL repository to resolve
Identify variation and if none, delete duplicate
Collect XY

Anomaly 3

Anomaly 3 (duplicate coordinates) could result from two possible situations. One is where an observation has the same facility name but one of them has incomplete fields (in other words, one seems to be a duplicate of the other). In that case, the second facility should be deleted. The other is where two distinct observations have the same coordinate. This may, in fact, be correct, such as when a pharmacy is located in or at the same site as a hospital. This situation needs to be investigated, to determine if it is true.

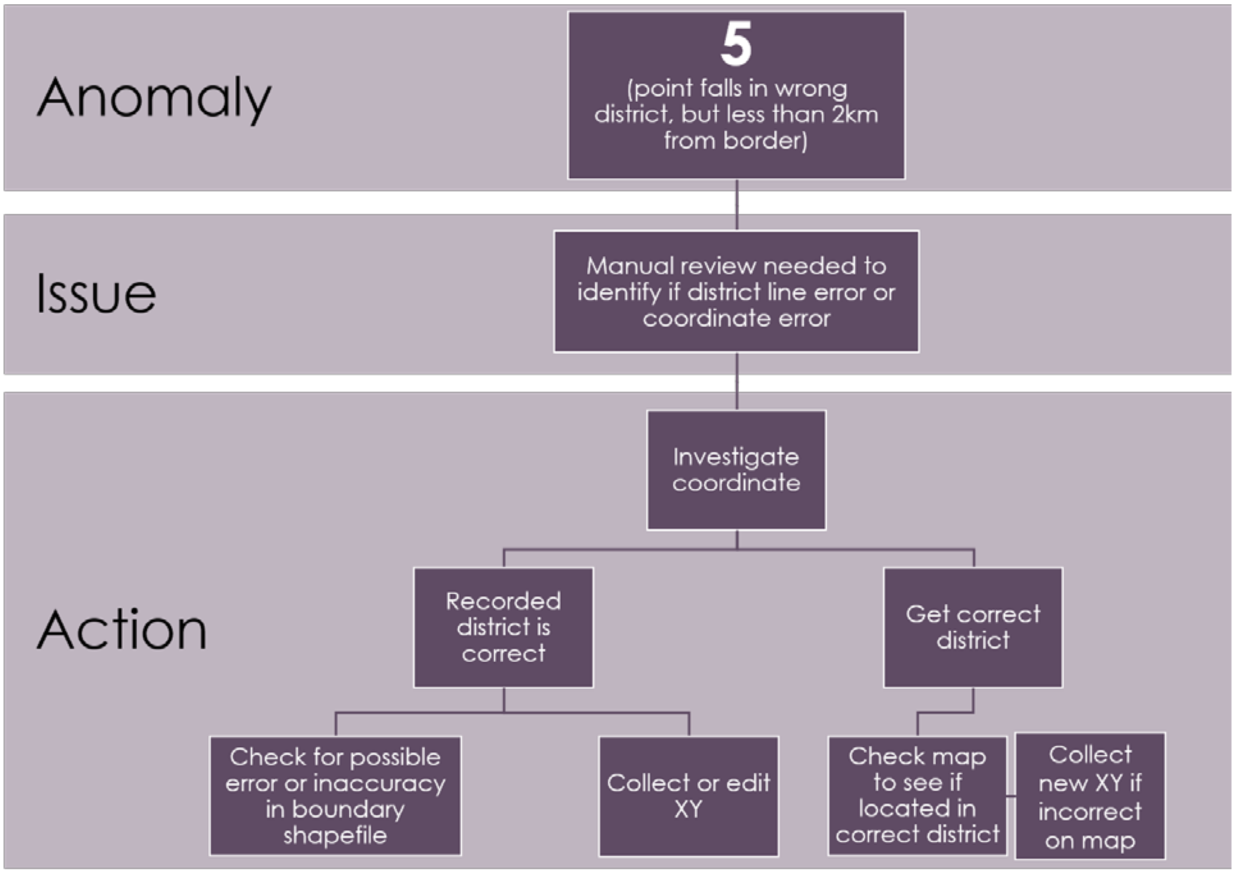
The following flowchart shows a workflow for addressing Anomaly 3 records.


Anomaly
3
(duplicate coordinates)
Issue
Two observations have similar facility name but one observation has incomplete field(s)
Same coordinate for two distinct observations
Action
Identify correct observation and delete duplicate
Investigate coordinate

Anomaly 5

Anomaly 5 occurs when a point is falling outside of its recorded district, but less than two kilometers from its border. This anomaly requires investigation. Either the coordinate is correct and the district boundary is inaccurate, or the coordinate is wrong and needs to be recollected and/or reentered in the system.

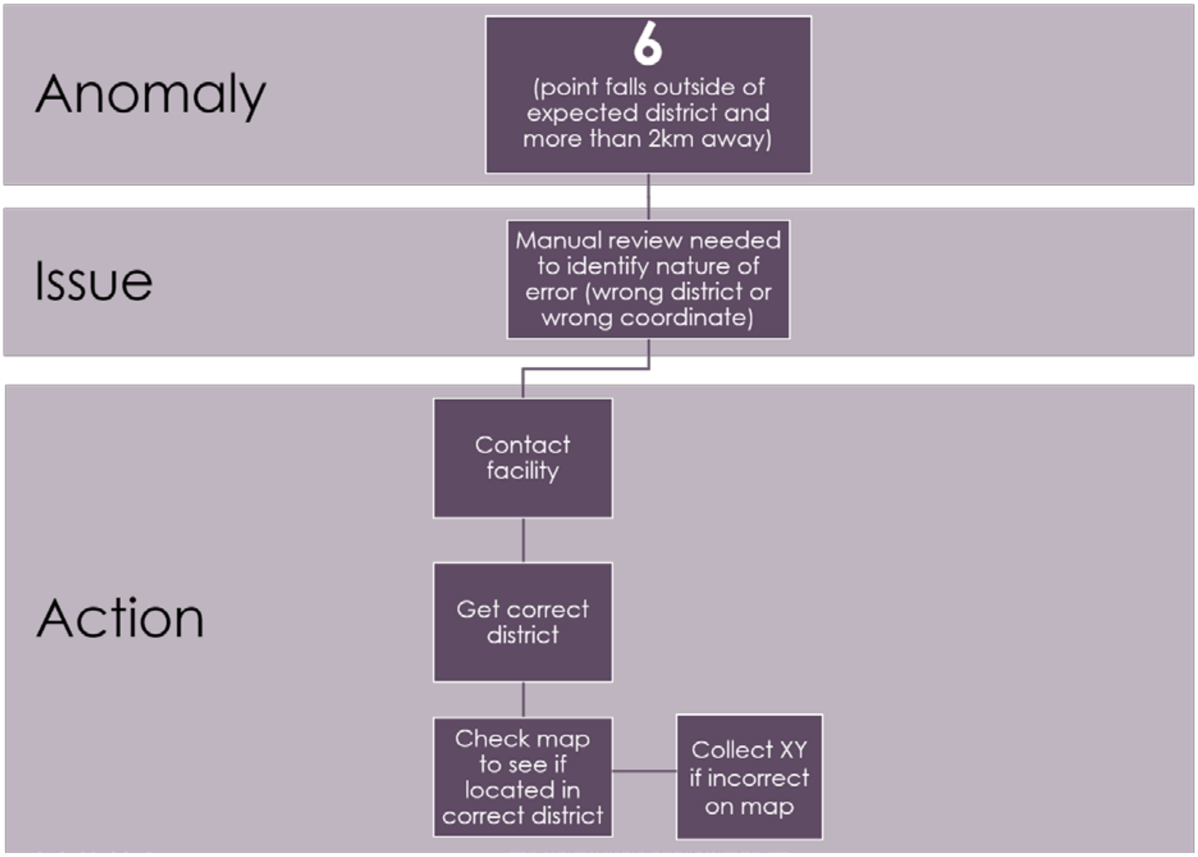
The following flowchart shows a workflow for addressing Anomaly 5 records.


Anomaly
5
(point falls in wrong district, but less than 2km from border)
Issue
Manual review needed to identify if district line error or coordinate error
Action
Investigate coordinate
Recorded district is correct
Get correct district
Check for possible error or inaccuracy in boundary shapefile
Collect or edit XY
Check map to see if located in correct district
Collect new XY if incorrect on map

Anomaly 6

Anomaly 6 occurs when a point falls far outside of its expected location. This indicates either an incorrect district field or an incorrect coordinate. The record needs to be investigated. If the district field is wrong, it will need to be corrected. If the coordinate is wrong, a new one needs to be collected and entered in the system.

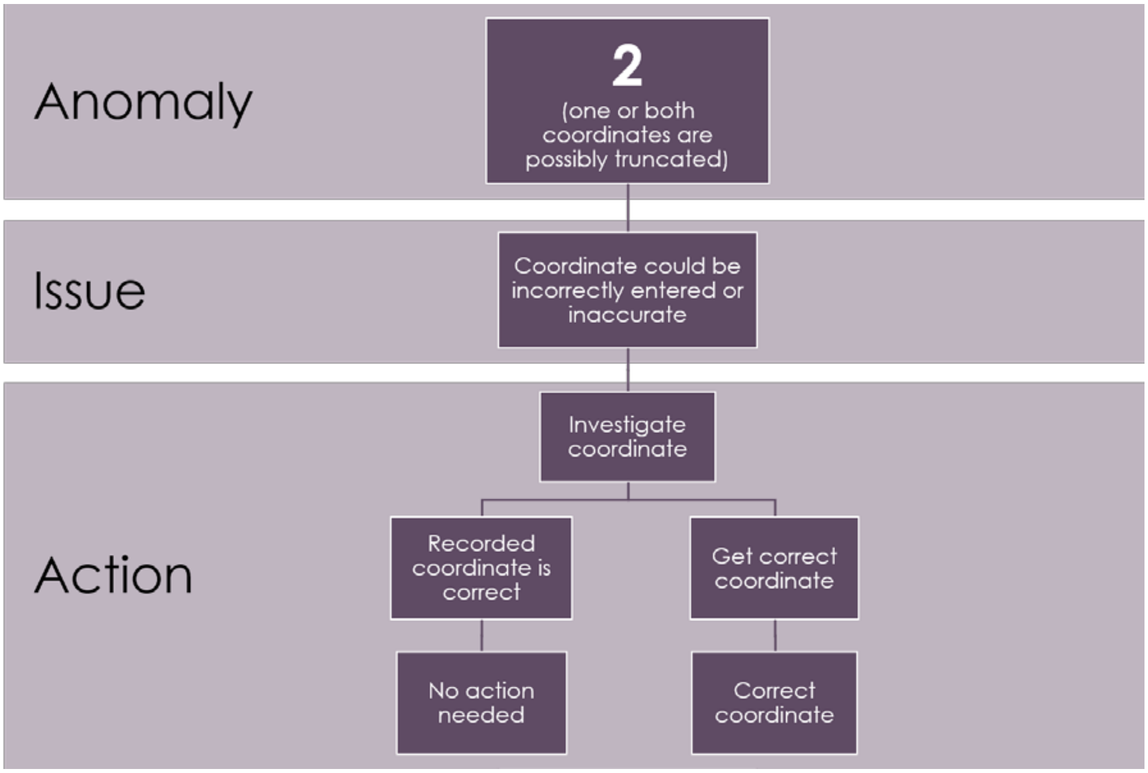
The following flowchart shows a workflow for addressing Anomaly 6 records.


Anomaly
6
(point falls outside of expected district and more than 2km away)
Issue
Manual review needed to identify nature of error (wrong district or wrong coordinate)
Action
Contact facility
Get correct district
Check map to see if located in correct district
Collect XY if incorrect on map

Anomaly 2

Anomaly 2 occurs when the X and/or Y coordinate appears to be truncated. This could indicate inaccuracy in the coordinates and requires investigation. If the coordinate is accurate, no action needs to be taken. If it is missing digits, a new coordinate needs to be taken and entered.

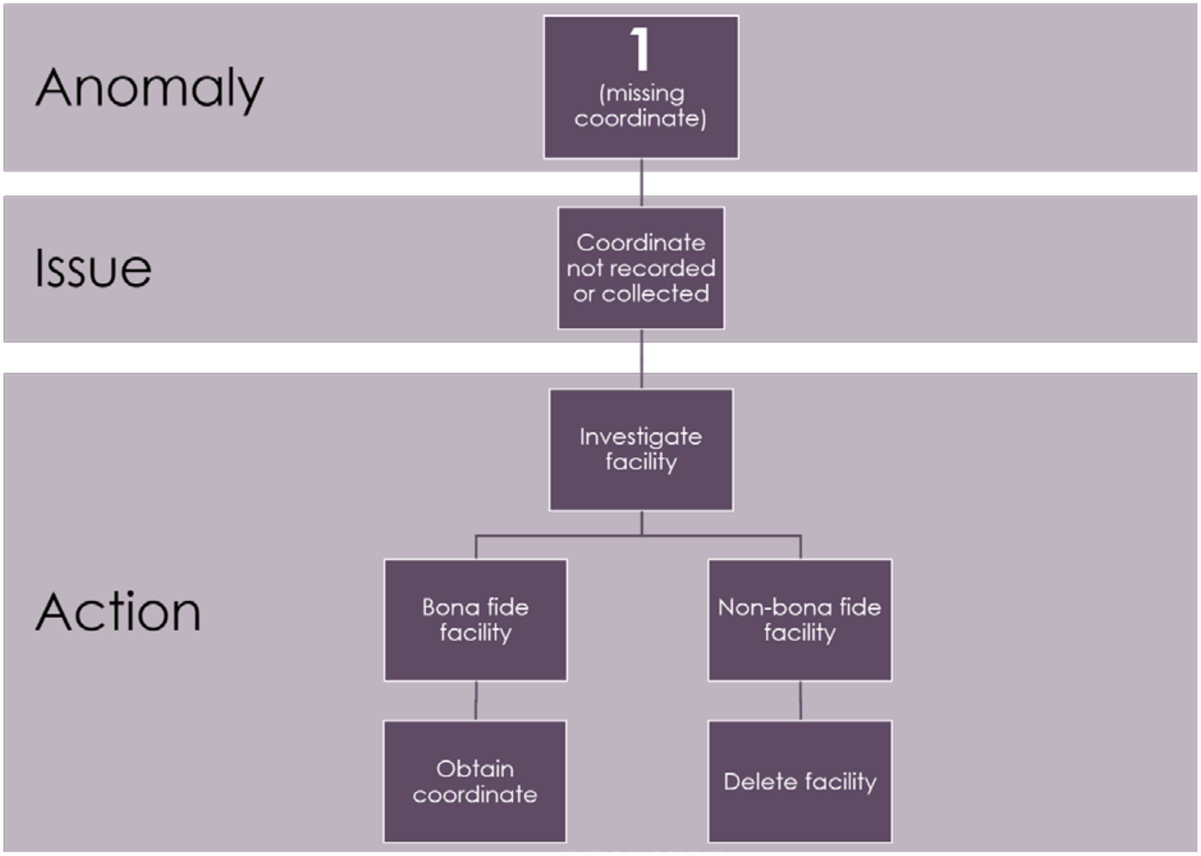
The following flowchart shows a workflow for addressing Anomaly 2 records.


Anomaly
2
(one or both coordinates are possibly truncated)
Issue
Coordinate could be incorrectly entered or inaccurate
Action
Investigate coordinate
Recorded coordinate is correct
Get correct coordinate
No action needed
Correct coordinate

Anomaly 1

Anomaly 1 indicates a missing X coordinate, Y coordinate, or both. Each anomalous record needs to be investigated. If the facility does not exist, the record needs to be deleted. If it needs a coordinate, a new one needs to be collected and entered.

The following flowchart shows a workflow for handling Anomaly 1 records.



We recommend handling each record's anomalies in the order shown above.

Other Anomalies

There could be other anomalies beyond the spatial domain, such as missing attribute data, out-of-range values, and so forth. The SQUAD tool is not designed to catch these anomalies. Resolving them will require following the procedure outlined in the MFL governance policy for data review and updating.

Step 5. Implement the remediation plan.

The next section provides a detailed review of how to handle each anomaly, including links to further resources, starting with Anomaly 1.

Anomaly 1: Missing Coordinates

Anomaly 1 occurs when there are no coordinates or when one of the coordinates is 0.0. Possible solutions are:

- Review the data collection sheet, if available, and determine whether records were omitted.
- Obtain the actual coordinates of a facility using a handheld GPS receiver or mobile phone. For more information on GPS data collection, see the following resources:

    - MEASURE Evaluation's mini-tutorials (5- to 10-minute videos) can help beginners collect GPS coordinates, download the data to a computer, and carry out other basic GIS tasks: https://www.measureevaluation.org/resources/training/online-courses-and-resources/non-certificate-courses-and-mini-tutorials/gis-step-by-step-tutorials/.
    - MEASURE Evaluation also offers a resource on collecting data using a GPS receiver: https://www.measureevaluation.org/resources/training/capacity-building-resources/geographic-information-systems-mapping-and-analysis-of-spatial-data/tools-tips/FS1383_GPScoordinates_Dec2013.pdf/view.
- Obtain the coordinates through a visual examination in Google Maps or Google Earth. For example, a cursor placed over a known nearby road intersection will display the coordinates in the lower righthand corner of the screen. More information on using Google Maps to search locations is available at http://support.google.com/maps/answer/3092445.

Anomaly 2: Truncated Coordinates

Anomaly 2 is when a coordinate is missing significant digits, resulting in a lack of adequate precision. The precision of the coordinate depends on the number of digits recorded. However, there is a paradox with precision. Enough significant digits should be included in a coordinate to meet accuracy standards for the database. But too many digits provide a precision beyond what is really needed (or even technically possible without using survey-grade GPS receivers). Typically, GPS units and GPS-enabled phones and tablets can record a GPS coordinate to six digits, which provides a location precision of roughly 10 centimeters. This level of precision is not necessary when locating a building, such as a health facility, where a precision of one meter (five digits after the decimal point) is more than sufficient.

The table below presents values and the associated precision in locations near the equator. The values change as you move further from the equator.

Coordinate	Approximate precision
23.1	10 kilometers
23.12	1 kilometer
23.123	100 meters
23.1234	10 meters
23.12345	1 meter
23.123456	10 centimeters

Possible solutions are:

- Review the data collection sheet, if available, and determine whether the records were incorrectly transferred or whether digits were left out or rounded.

- Obtain the actual coordinates of a facility using a handheld GPS receiver or mobile phone (see the resource links provided above).
- Obtain the coordinates through a visual examination in Google Maps or Google Earth. For example, a cursor placed over a known nearby road intersection will display the coordinates in the lower righthand corner of the screen (see the resource link given above).

Anomaly 3: Duplicate Coordinates for Distinct Records

Anomaly 3 is when two (or more) records have the same coordinates. This anomaly does not necessarily mean that there is an error in the data. Instead, in some instances, it may be possible for two sites to be at the same location—for example, in a medical park. For records with this anomaly, the first step is to determine whether there are, in fact, two distinct sites at that location. Here's how to proceed:

- Determine whether the anomaly is really an error. The data could be correct, such as in the case of a dispensary located at the same address as a hospital.
- If an investigation reveals that there are not two sites, then the same steps used to address Anomaly 1 and Anomaly 2 can be employed, that is, review any GPS logs, recapture the location, or use imagery.

Anomaly 4: Duplicate Facility Names

Anomaly 4 is the presence of records that contain duplicate facility names. Once again, this is not necessarily an indication of erroneous data, but it is an anomaly that merits investigation. To resolve this anomaly, it is first necessary to determine whether there are, in fact, two distinct sites with the same name. This can be done by contacting the site directly to get clarification or by reviewing documents, reports, licensing records, or other sources to determine the site's official name, which can then be corrected in the database.

- Determine whether the anomaly is really an error. The data could be correct, such as in the case of two clinics in different parts of the country having the same name.

Anomaly 5: Coordinate Not Located in Expected District, but Falling within 2 Kilometers of a Border

If a site is supposed to be in one administrative unit (such as a district) but shows up somewhere else when mapped, this is an error that needs to be resolved. When the site is near (within two kilometers of) its expected location, three potential data issues are in play:

1. The GPS coordinate could be incorrect.

2. The administrative unit field in the site list could contain incorrect information.

3. The GIS boundary file for administrative units could have errors, be inaccurate, or out of date.

To address this anomaly, the solutions are:

- Review the data collection sheet, if available, to confirm the coordinates.

- Obtain an administrative boundary file from a different source. A point falling very close to a boundary could indicate that the boundary itself is wrong or inaccurate. A good source is the Database of Global Administrative Areas: https://gadm.org or a national mapping agency.
- Validate the coordinates using imagery, such as Google Earth or Bing.
    - For information about adding imagery from within QGIS, see https://gis.stackexchange.com/questions/20191/adding-basemaps-from-google-or-bing-in-qgis. (This post talks about the OpenLayers Plugin for 2.18 and the XYZ Tile Server provider for 3.0.)
    - For information about Google Earth, see http://www.google.com/earth/.

Anomaly 6: Coordinate Not Located in Expected District and Farther than 2 Kilometers from the Border

The last type of anomaly is when a site is not located anywhere near where it should be. For example, the site is in the ocean or in another country or any location more than two kilometers from the expected administrative unit. The following are the steps to address this anomaly:

- Review the data collection sheet, if available, to confirm the coordinates.
- Check the spelling of the administrative units in the administrative boundary file.
- Check to confirm that the point file has the correct administrative area designations.
- Validate the coordinates using imagery, such as Google Earth (see above), or recollect the data using a GPS receiver or mobile device.

Step 6. Periodically rerun the SQUAD Tool and repeat Steps 1 to 5.

a. Rerun the SQUAD Tool and repeat the remediation, as needed. NOTE: The SQUAD Tool needs to be run for the country as a whole, whereas remediation can be done on a region-by-region basis. This is for the purpose of catching duplicates or other errors that can occur across region lines.

b. The MFL governance policy should have a data update policy on how frequently data are updated. After each update, the SQUAD Tool should be run on the entire country to assess the data quality.

CONCLUSION

By running the SQUAD Tool regularly and following the six steps outlined in this document, anomalies in the MFL data set can be identified and remediation plans can be made. High-quality data for the MFL are essential for the information to be trusted and useful. Regular updates and quality checks can help ensure the adequate provision of services.

MEASURE Evaluation
University of North Carolina at Chapel Hill
123 West Franklin Street, Suite 330
Chapel Hill, North Carolina 27516
Phone: +1 919-445-9350
measure@unc.edu
www.measureevaluation.org

This publication was produced with the support of the United States Agency for International Development (USAID) under the terms of MEASURE Evaluation cooperative agreement AID-OAA-L-14-00004. MEASURE Evaluation is implemented by the Carolina Population Center, University of North Carolina at Chapel Hill in partnership with ICF International; John Snow, Inc.; Management Sciences for Health; Palladium; and Tulane University. Views expressed are not necessarily those of USAID or the United States government. MS-18-146

ISBN: 978-1-64232-086-2