cirad
AGRICULTURAL RESEARCH
FOR DEVELOPMENT

Clean up report

SIT suppression trial - Sudan

Facundo Muñoz

17 August, 2023

Contents

Table 1: Data sets included in the initial transfer.

File	Name	Description
1	all data density.csv	Number of catches of culex, wild male _Anopheles_, and female _Anopheles_, and status code per adult trap and day. Further variables include attributes of the trap (location, sector) or the sampling unit (land classification)
2	breeding.csv	Number of larvae of _Anopheles_ and number of dips by date and breeding site. The breeding site is characterised by GPS coordinates, the sector and the area.
3	swarm data.xlsx	Number of catches of sterile and wild males during swarms, by day of occurrence.

1 Introduction

The present document describes the data processing steps that were performed upon the consolidated data sets from Sudan's *Anopheles arabiensis* suppression trial experiment in the Merowe area, Northern State.

The data include 3 observation processes of *Anopheles arabiensis*:

1. Adults captured in **adult traps** installed in fixed locations and surveyed periodically during the experiment.

   Three **traps** were installed in each of 98 **clusters** which were sampled from a 100 x 100 m grid proportionally to strata defined by 4 **land-classes**, within each of 9 **sectors** in which the 3 **areas** in which the study region was divided.

2. Larvae found in **breeding sites** within the perimeter of a sector. Sampling is carried out only when suitable breeding sites (e.g. ponds) are found.

3. Adults captured during **swarms**, which occur in specific places for short periods of time. Experts used their field knowledge to anticipate times and sites of swarming events, which may or may not take place.

2 Production of complementary data sets

Initially, the project data was synthesised into 3 data sets, corresponding to the types of observations collected (Table 1).

Table 2: Input data sets integrated and developed during the mission.

File	Name	Description
4	cleaned density data.csv	A corrected version of the initial file 'all data density.csv'.
5	all traps.shp	Point locations of traps, their identification code and sector.
6	River_bank.shp	Polygons of sampled clusters of class river bank.
7	round_dates.csv	Starting dates of survey rounds in each area.
8	Mosaic_field.shp	Polygons of sampled clusters of class mosaic field.
9	Sectors.shp	Polygons of the sectors, with identification code.
10	Release.xlsx	Release dates and number of individuals released in the sit sector.
11	Mosaic_trres.shp	Polygons of sampled clusters of class mosaic trees.
12	Settelments.shp	Polygons of sampled clusters of class settlements.

Data about the experimental set up, such as the number of traps and their locations, number and dates of releases, the relationship between traps, sampling units, sectors, and areas, the coordinate reference systems of the coordinates, meaning of status codes, etc. were not explicitly available nor documented. Instead, it was either implicit in reports, presentations, or dispersed across geographical files in different computers or folders.

Having this information explicitly structured as data tables enable verification of the internal consistency of the data. Which is what we do in the next section. Moreover, it makes it easier for new researchers that join the team to understand the structure of the data. This is even beneficial for the original researchers that come back to the project after some time off.

In consequence, we collected the data files listed in Table 2. Some of these files, such as the Shapefiles (`.shp`) existed already, but were stored in different places. We merely gathered and integrated them as first-class members of the project's data. In contrast, `Release.xlsx` was created from a table in a report, and `round_dates.csv` was developed from the directory structure where survey files were organised. Finally, file number 4 is a replacement version for file number 1.

cirad

3 Data verification steps

The experimental data were collected on the field using GPS-enabled tablets with pre-specified forms, during survey *rounds* spread over one or several non-consecutive days, which were generally conducted once or twice a month.

During such rounds, operators collected insects from all traps, stored them in a labelled box and recorded a new entry in the device. Whenever there was a breeding site near the trap, the water was sampled in search of larvae. The numbers of dips and larvae counts were recorded together with the adult trap data.

The labelled adult samples were then transported to the laboratory and counted by species, sex, and sterilisation status.

Overall, this collection procedure is robust and reliable. However, some common issues required a more thorough verification. For instance, upon inspection of the data, we have identified:

1. Records with incorrect trap or sector number.

   The device did not perform any validation of the numeric fields for the trap code and the sector number, and the traps were not labelled. For instance, it is easy to type 185 instead of 85, and not notice.

   This could be alleviated by labelling traps in the field, validating the data entry with the sector, or with the coordinates, or using a QR or a bar-code in the labels that can be read by the device. Moreover, since it is the GPS itself that guides the operator to the location of a given trap, perhaps the trap code could be recorded automatically without user intervention, preventing these mistakes altogether.

2. Upon detecting some mistake, the operator might abort an entry that is nevertheless recorded.

   This creates duplicate entries, some times filled with zeros, but possibly with some of the values filled in, which makes it very difficult to tell apart from the correct record automatically.

   Ideally, the device would automatically detect and prevent multiple entries.

3. The recorded GPS coordinates might be inaccurate due to weather conditions or malfunctioning.

   If the recorded trap codes were reliable, this is not necessarily a problem, since the location of the traps is known in advance. However, when the recorded coordinates are far from the location of the recorded trap code, it is difficult to tell which one is correct and which wrong.

cirad

Ideally, the device would make sure that trap code, sector and coordinates are coherent and consistent with the plan before recording a new entry.

In principle, recording redundant information can be useful in order to verify the consistency of the data. Ideally, this verification would be performed automatically by the device *in-situ*, in order to prevent mistakes. Otherwise, the verifications need to be conducted later during the data processing stage.

The result of these unverified issues is that some observations were incorrectly attributed to a different trap, possibly on a different sector, and that some fake records with values of 0 were added.

Hopefully, the prevalence of these issues is low enough so that the impact on the analysis is negligible.

Nevertheless, we describe below a few actions undertaken in order to quantify and correct these problems as much as possible.

3.1 Exclusion of entries

A different problem concerns the set of surveys used effectively in the analysis. There were a certain number of observations that in my opinion should have been excluded from the analysis. Specifically,

1. Traps 295, 296 and 297 were initially installed and used until 2015-08-30, but later abandoned because of flooding and refusal from the land owners. The traps were removed from the study. So should their observations.
2. During surveys, operators recorded whether the trap was in suitable state. Whenever traps were missing, broken, dry, etc. a special code was registered. Only traps functioning normally should be considered for analysis.

3.2 Clustering

The sampling design of the traps in the experiment consisted on choosing 3 suitable locations within each of the sample units, which were square areas of an hectare in surface that were sampled in proportion to land-use and cover strata.

This design implies that observations within a sampling unit are correlated, and this needs to be taken into account in the analysis to avoid underestimating the standard errors involved.

However, the information about the sampling unit was not included in the data set, since it has been ignored so far.

4 Trap data

The data about adult traps (sector, land classification, and coordinates) were included in the table of adult surveys, and therefore replicated as many times as collections in each trap. These values were re-assessed and re-typed at each survey, leading to a certain number of data-entry mistakes. Specifically, 76 traps were attributed two 2 or 3 different sectors, depending on the collection date (Fig. 1).

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(.fns = factor)`.
## Caused by warning:
## ! Using `across()` without supplying `.cols` was deprecated in dplyr 1.1.0.
## i Please supply `.cols` instead.
```

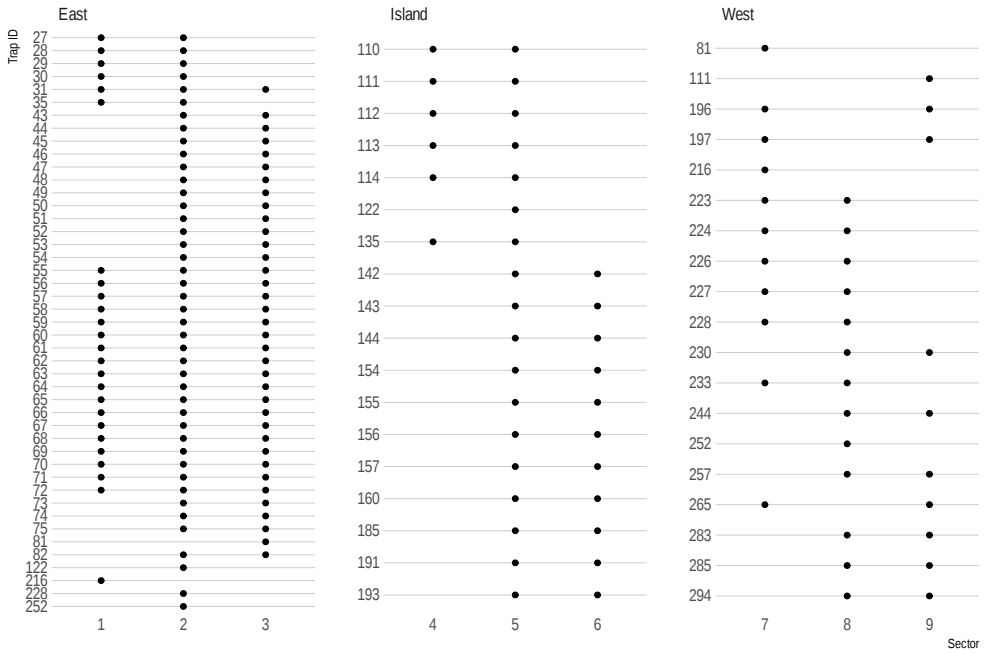

Figure 1: Sectors where traps have been recorded. Only traps with records in multiple sectors are shown.

These inconsistencies in the data could be removed simply by ignoring the sector recorded the during surveys and using the known arrangement of traps into sectors instead. However, it is possible that the trap code was mistyped while the sector was correct for some of these surveys. In these cases, this approach would only hide the problem

Similarly, coordinates were taken by GPS and recorded at each survey, resulting in some variation in recorded locations of about 50 m very common. However, some surveys were located much more far away from the actual location of the trap, with distances of up to 6 km in the most extreme cases (Fig. 2)

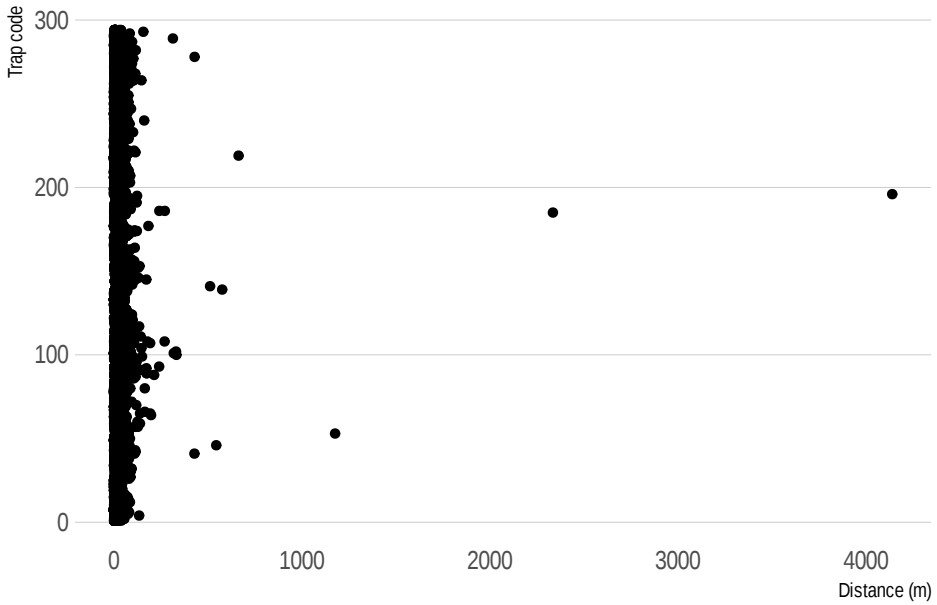

Figure 2: Distances from each survey location and the median location of all the surveys recorded at the same trap.

Upon manual inspection of some of these most extreme examples, we found that some of them corresponded to incorrect records of the trap ID, either unnoticed or noticed and aborted (but not removed), leading to a record with zeroes attributed to some other trap code.

We have manually checked these records far away from where they are supposed to be.

Furthermore, we extracted the surveys where multiple records in the same day, and manually checked these records as well.

Since some of the survey monthly "survey rounds" actually take more than one day (2 or 3), there can still be incorrect records not found by the previous procedure due to being collected in different days.

The interest in checking the absence of duplicate trap codes per survey round motivated the effort to gather the necessary information to define a variable identifying

Table 3: First 10 (of 158) surveys for which one of the verification checks fail. We selected non-matching trap and sector codes or distance from survey to recorded trap greater than 400 m. or distance from survey to recorded sector greater than 50 m.

Surv	Date	Trap	Rec. sec.	Sec.	Sec. match	D. trap (m)	D. sect. (m)
118	2014-05-22	122	2	5	FALSE	12	928
190	2014-05-22	111	9	4	FALSE	10	1899
1485	2015-05-01	55	1	3	FALSE	12	2981
1487	2015-05-01	56	1	3	FALSE	20	3002
1488	2015-05-01	59	1	3	FALSE	39	2982
1489	2015-05-01	60	1	3	FALSE	7	3055
1490	2015-05-01	61	1	3	FALSE	15	3115
1492	2015-05-01	57	1	3	FALSE	7	3068
1493	2015-05-01	58	1	3	FALSE	40	2971
1494	2015-05-01	62	1	3	FALSE	4	3112

the survey round.

5 Adult survey data

As a verification of the consistency of the records, we performed 3 checks:

1. Whether the recorded trap and sector were consistent. I.e., the trap is indeed within the sector. If this check failed, either the trap code or the sector code were wrong.
2. The distance from the survey location to the location of the recorded trap. This distance should be small. Within a range of about 10-20 m.
3. The distance from the survey location to the location of the recorded sector. This distance should be 0 in most cases, or very small.

This resulted in 163 inconsistent records that were verified and corrected manually. Table 3 displays the first 10 of these records, where the last three columns correspond which each of the verifications.

Table 4: First 10 (of 113) surveys for which one of the verification checks fail. We selected non-matching sector code and area or distance from survey to recorded sector greater than 50 m.

Surv	Date	Sector	Rec. area	Area.	Area. match	D. sect. (m)
1	2015-05-01	8	Island	West	FALSE	0
2	2017-03-30	7	Island	West	FALSE	0
3	2017-03-29	8	Island	West	FALSE	0
4	2017-04-14	7	Island	West	FALSE	0
5	2015-10-28	8	Island	West	FALSE	0
6	2015-03-14	8	Island	West	FALSE	0
7	2015-07-28	7	Island	West	FALSE	0
39	2015-05-01	1	East	East	TRUE	2983
42	2015-04-13	2	East	East	TRUE	875
114	2016-10-29	5	West	Island	FALSE	492

6 Larvae survey data

As a verification of the consistency of the records, we performed 2 checks:

1. Whether the recorded sector and areas were consistent. I.e., the sector is indeed within the area. If this check failed, either the sector code or the area were wrong.
2. The distance from the survey location to the location of the recorded sector. This distance should be 0 in most cases, or very small.

This resulted in 113 inconsistent records that were verified manually. Table 3 displays 10 of these records, where the last 2 columns correspond which each of the verifications.

The 3 records with distance to sector larger than 0 were errors in the sector code, while the rest were mistakes in the area.

The sector codes were fixed manually, whereas the recorded areas were replaced by the area corresponding to the sector.

7 Swarms survey data

In contrast to adult and larvae surveys, swarms were **only** surveyed in sector 3, where sterile males were released.

There are no more location specifics in these data. Thus, no special verifications or processing were necessary apart from date format, normalising variable names and removing calculated variables.

8 Data processing

Apart from the verification of the consistency, some processing steps have been undertaken in order to filter observations, select variables of interest, compute new variables derived from the original, or more generally format and structure the data in a form suitable and convenient for statistical analysis.

We have implemented all these steps in a data-processing pipeline in R, using the package `targets`[1].

Here we provide a brief overview of the main processes applied to the input data. General-purpose cleaning steps such as standardising variable names, declaring levels of categorical variables (factors) or applying proper format to dates are omitted. All details can be examined from the code.

8.1 Larvae surveys

- Declare the object as a set of spatial points, with the appropriate Coordinate Reference System.
- Add an identification code for the survey.
- Fix errors in sector codes.
- Replace area with correct values from the table of `sectors`.

8.2 Adult surveys

- Exclude observations from traps 295, 296 and 297, which were removed from the experiment in august 2015.
- Exclude observations from surveys where the traps was found non-suitable.
- Include columns for cluster, sector, area, land class, role and survey round
- Fix trap codes that were manually identified as incorrect and remove duplicate incorrect records
- Replace the recorded sector (which contains many errors) by the known sector of the trap

[1] https://books.ropensci.org/targets/

- Declare the object as a set of spatial points, with the appropriate Coordinate Reference System

8.3 Swarm surveys

- Added a unique identifier
- Remove calculated variables

8.4 Releases

- Fixed an error in release number
- Added a missing release
- Fixed a date written in an inconsistent format
- Added a unique identifier of releases

8.5 Survey-round dates

- From the **starting** dates of the visit-rounds in each area, compute the round for every intermediate date in the adults survey data.

8.6 Traps

- Remove residual records with missing code numbers
- Add variables for the cluster and land class, the role of the sector and the area
- Add variable for the trap type (actually constant)

8.7 Clusters

- Integrate the 4 sets of sampling units into a single geographical object with a categorical variable for the land class
- Remove irrelevant variables

8.8 Sectors

- Remove buffer sectors
- Add variables `role` (control / sit) and `area` (East / Island / West)

cirad

9 Conclusions

In order to improve the quality of the data, we conducted a series of verifications, some of which required producing new data tables with information that was implicitly spread within documents, folder structures, cartography, and the memory of the project managers.

In addition, we conducted a series of processing steps in order to prepare the data for statistical analysis.

As a result, we organised the data of the project into 8 objects: `traps`, `sampling_units`, `sectors`, `releases`, `round_dates`, `adult_surveys_clean`, `larvae_surveys` and `swarm_surveys`

We provided a repository with the processed data sets, the ones produced during the mission, and full metadata.