ClusterApp - Tutorial

ClusterApp is an application to guide and streamline field studies based on GPS cluster data. It is developed in `shiny` (Chang et al. 2023) within R version 4.3.1 (R Core Team, 2023) and built within the `golem` framework (Fay et al. 2023).

The following document gives a tutorial on the basic usage of the app with the dataset *"wolf"* provided by the Scandinavian Wolf Research Project (SKANDULV), as well as the usage of the app when applying it on multiple individuals with the dataset *"bears"* provided by the Scandinavian Brown Bear Research Project (SBBRP). There are five chapters describing the general usage of the app. **Chapter 1: Starting the App** describes how to download and start the app. **Chapter 2: Getting to know the app** focuses on implementing a cluster analysis using the *"wolf"* dataset. **Chapter 3: Consecutive analysis** deals with important points when running subsequent analyses using an already downloaded shapefile from a previously executed cluster analysis. **Chapter 4: Analysis on multiple individuals** deals with performing cluster analysis on multiple individuals within the same GPS file. **Chapter 5: Handling GPS bursts** gives a short introduction on the possibilities of the app to work with GPS burst data, such as proximity-triggered GPS events.

1. STARTING THE APP

The app can be started by downloading the package from GitHub:

```
install.packages("devtools")
devtools::install_github("JohannaMz/ClusterApp")

library(ClusterApp)
```

If you want to follow the tutorial, you can download the GPS data. Data for the wolf examples are loaded by the command:

```
data <- ClusterApp::wolf
```

And the data for the bear examples, including multiple individuals, are loaded by the command:

```
data <- ClusterApp::bears
```

More information regarding the two datasets can be found by searching for the datasets in the *Help* panel. To be able to use the data within the app, it must be saved in a folder of your choice. It is good practice to have a separate folder per GPS data file, which can include data on one or more individuals. When running multiple cluster studies, each new GPS file should be stored or replaced here. In this folder, you will also find the output of the cluster analysis.

```
write.csv(data, "path/to/folder/data.csv")
```

After loading the package, you can start the app by running the command:

```
ClusterApp::run_app()
```

The app opens in a separate window and R runs in the background. For a better view, choose *"Open in browser"* in the upper left corner. The app will now open within your Internet browser.

The app has three main steps:

1. **Upload GPS data**
    - Upload the GPS data
    - Define the necessary column names
    - Adjust date format and coordinate system format
2. **Adjust Cluster Analysis Parameters**
    - Adjust all the parameters necessary for the cluster analysis
3. **Cluster Analysis Output**
    - Look at the results, enter data in cluster file, and download cluster -, point- and map files

2. GETTING TO KNOW THE APP

1. **Upload GPS data** (Figure 1)

General usage is shown with the *"wolf"* dataset. The data is loaded by selecting *"Upload the original GPS file here:"* and the path to your datafile will appear. Possible formats are *.csv, .shp* or *.dbf.* For the format *.csv,* you can choose the correct file separator, making sure that the columns are loaded correctly.

After uploading the data, the file path appears, and tabs *"Data"* and *"Data Summary"* give information about the uploaded data. These can be used to check if the data were uploaded correctly, and which column names fit as input for the ID (Individual ID), time stamp, and coordinate variables. The timestamp has to include both date and time in one column with the default format of "**Y**ear-**m**onth-**d**ay **H**our:**M**inute:**S**econd", but this can be adjusted based on standard R formatting.

Additionally, the coordinate system your data is originally in *(EPSG code 4326 for WGS84)* as well as in which UTM zone the data should be downloaded *(Zone 33 for Sweden),* can be adjusted. Within the analysis the data is always converted to UTM, as this coordinate system has metric units, which are more intuitive within analyses involving distance measures.

The red info sign (), always give extra information on the necessary steps within the analysis. Blue underlined text are URLs to external websites.

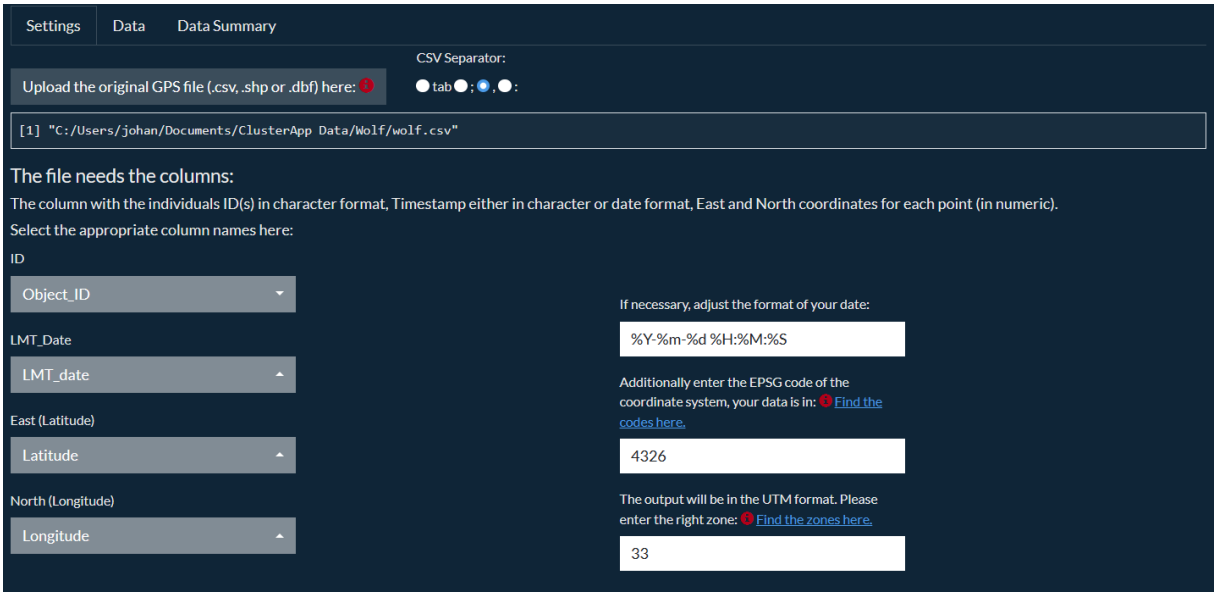

Figure 1: Filling in the settings within the tab Upload GPS data for the example data "wolf".

2. **Adjust cluster analysis parameters** (Figure 2)

In the first step, it is necessary to state a unique label of the individual(s) the input data belongs to, and which will be used throughout the entire field season. This should be a label useful for separating it from other cluster studies, that might be done at the same time. When executing consecutive cluster analyses this unique label will make sure that earlier output files with the same label are found by the app, if available, in the folder of the GPS data (further explanation under **Consecutive analysis**). For the tutorial we chose the label *"wolf_demo".*

One of the important questions regarding cluster analysis, which is directly related to the target research questions, is setting the parameters necessary to generate biologically relevant clusters. Clusters are formed by applying a buffer of *x* metres around every GPS point. If these buffers overlap (so at a maximum distance of *x* times 2), they will form clusters which will be saved as a cluster if y numbers of points are within this buffer for all points within the study period.

Therefore, it is possible to adjust the following parameters:

- Buffer size: Depending on the study species and research questions, the relevant distance criteria between GPS points can be adjusted. Changing the distance results in different cluster sizes. *We choose a distance of 100 m, which will mean locations within 200 meters of each other form a cluster.*
- Number of GPS points: Again, depending on the study species and research questions, you can set the minimum number of positions used to generate a cluster. *For this demo, set the number of points to 2.*
- Study Period: To be able to select the period during which the animal is being followed, both the starting and end date need to be set. The end date can be in the future if the study is ongoing. Often the fix rate of the GPS device is adjusted to receive GPS points more frequently, therefore it is important to set the dates correctly and thus exclude points from outside this period. *The study period of this demo data was from 28 February, 2022 to 17 April, 2022.*
- To demonstrate a case of downloading data and then using this downloaded data in a consecutive analysis, *we set the end date to 17 March, 2022*. That means only data between 28 February and 17 March, 2022 will be used for this analysis. In **Chapter 3: Consecutive analysis** we explain the column *"Some additional info, if you have done an analysis before"*.

The settings of a 100 meter buffer around GPS points, with a minimum two points per cluster, have been used in previous wolf predation studies, e.g., trying to estimate the extent of scavenging within the Scandinavian wolf population (Wikenros et al. 2023).

Further optional adjustments include the filtering of the initial GPS data to only include points every x minutes. The default is to use all GPS data points that are available. The summary of the time difference shows that for the wolf data, the time differences between positions stays constant at approximately one hour (60 minutes). Therefore, no adjustments are necessary here. More detailed explanation can be found in **Chapter 5: Handling GPS bursts.**

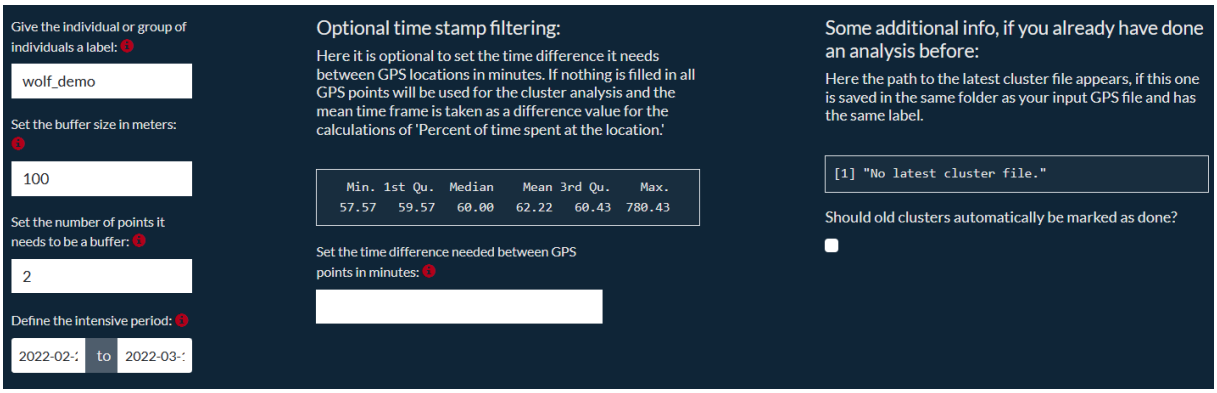

Figure 2: Setting the necessary parameters within the tab Adjust Cluster Analysis Parameters, which is separated into three columns. The first columns deal with the important parameters, the second column gives a summary of the time difference between GPS fixes and the third column related to the previous cluster analysis files.

3. Cluster analysis output

After setting all the data file options and the cluster analysis parameters, the cluster analysis can be performed under the tab "*Clusters Analysis Output".* A message will notify you if the analysis ran successfully and the output will appear (Figure 3).

Alternatively, a cluster shapefile can be uploaded from one of the folders. This is an option if the output of an earlier cluster analysis should be inspected without the need for a new analysis. When this is uploaded, some properties regarding the track and GPS points in the interactive map are not possible, as these are exclusively related to a performed cluster analysis.

Figure 3: Successfully performed cluster analysis. Alternatively, a latest cluster file can be loaded here.

Two tables are created with the analysis which will be visible in two different tabs (*"Clusters"* and *"Points Table"*).

The clusters table appears first as the default setting and contains all the data concerning the clusters established by the analysis (Figure 4). The table includes columns regarding the ID of the individual (selected as "ID" column), the unique "Cluster ID" (built as ID underscore Cluster ID) and relevant

information regarding the clusters (e.g., number of points in cluster, first and last date of visit at the cluster, centre points of the clusters). The identifier (i.e., "percent of time spent at the cluster") is calculated as the number of points the individual could have been at the cluster between the first and last time it visited (taking the mean time difference GPS points are taken) divided by the actual number of points within this cluster (further explanation in **Chapter 5: Handling GPS bursts**).

The data in the following four columns: "State", "Event", "Date done" and "Fieldworker" can be manually adjusted by the user by double clicking on it. "State" refers to the state the cluster has: all clusters that developed during this cluster analysis will be named "New", while clusters that expanded since the latest cluster analysis are stated as "Points added". In addition, the user can choose to manually adjust the state of a cluster. When clusters have been visited, the user may want to add what was found at the location (Event), when the cluster was visited (Date done) and the name of the fieldworker who visited the cluster. These adjustments can change the aesthetics of the interactive map (see below) and streamline data management.

The top of the data table also allows for specific filtering on all columns. This can either be useful for overview purposes within the map or downloading/looking at only specific clusters, e.g., specific IDs, within a specific time frame, etc.

Clusters | Points Table

The clusters table shows all clusters that were formed by the number of GPS points that were buffered and grew together. The data can be filtered in the top row and only this data is downloaded when choosing the formats .csv and .gpx. The columns State, Event, Date done and Fieldworker can be edited. These edits will be saved when downloaded.

Show 10 entries Search:

Individual ID	Cluster ID	Number of points	Percent of time spent at the location	First date visited	Last date visited	State	Event	Date done	Fieldworker	Centerpoint East	Centerpoint North
All	All	All	All	All	All	All	All	All	All	All	All
M_22_04	M_22_04_1	8	1.14	2022-02-28T02:00:10Z	2022-02-28T09:00:38Z	New				533442.07	6883907.63
M_22_04	M_22_04_2	12	1.09	2022-02-28T10:00:09Z	2022-02-28T21:00:12Z	New				533832.99	6884600.77
M_22_04	M_22_04_3	5	1.25	2022-03-01T03:00:38Z	2022-03-01T07:00:14Z	New				523217.41	6889002.87
M_22_04	M_22_04_4	13	1.08	2022-03-01T08:00:38Z	2022-03-01T20:00:10Z	New				523668.3	6888984.43

Figure 4: Cluster table output with a selected row, which is being edited in the column State.

The cluster table can be downloaded in the output formats: *.shp, .csv* or as a *.gpx* file and will always be saved in the folder where the GPS file was initially uploaded from (Figure 5). **Downloading the cluster file as shapefile is mandatory to save data changes and if the file should be used for any consecutive analyses.** This option will always download the entire cluster table in UTM and the specified zone in the folder of the original GPS data. Downloading *.csv* and *.gpx* files are optional for further purposes such as easy adjustments and data management in excel or loading the file onto a GPS device. The *.gpx* is downloaded as a track in WGS 84 for direct import into any further applications. If filtering options were applied in the table, only the filtered data frame will be downloaded as the *.csv* and/or *.gpx* file.

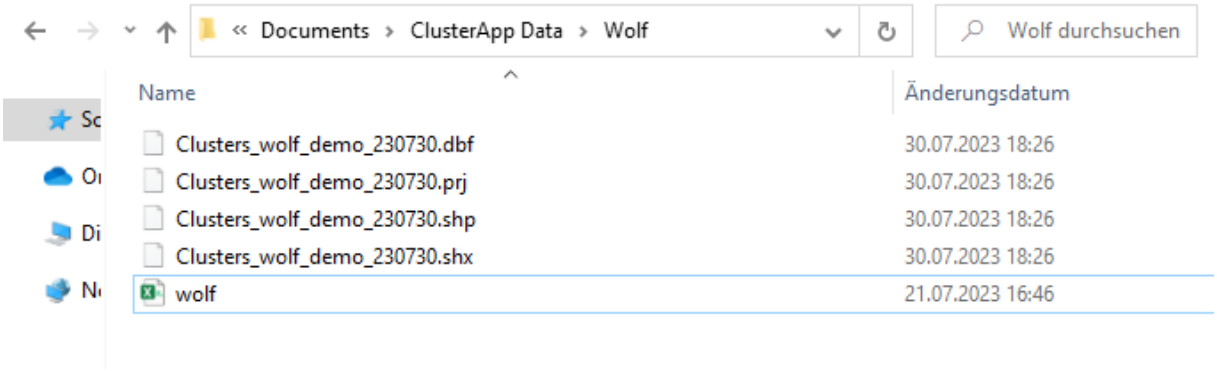

Figure 5: Folder with the initial GPS file named wolf and the downloaded shapefile for the clusters that were identified in this cluster analysis. The name is built according to the label given and the date that it is downloaded. This shapefile can now serve as a previous cluster analysis file.

The cluster table can be plotted as an interactive map by pressing *"Plot Data"*. The map has the "WorldTopoMap" as a general background layer, but can be switched to "WorldImagery", and will show several layers regarding the cluster analysis. Per default the map shows cluster polygons coloured according to their state, as well as the latest GPS position of the individual(s). There are additional options to show, for example, the walking trajectory, events of the clusters or the clusters ID as text within the map. All aspects can be activated and deactivated depending on the user's choices (Figure 6).

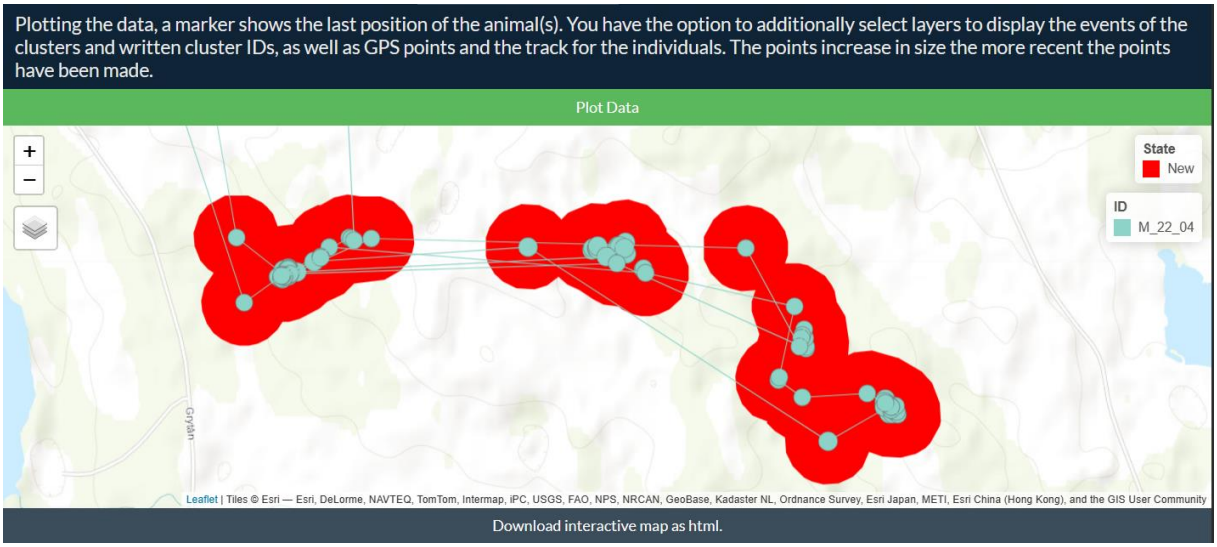

Figure 6: Interactive map output showing the clusters filled according to their state, as well as the GPS points and the track the individual walked.

Manual adjustments of the state and event column in the cluster table can be updated by plotting the data anew. These options help to illustrate visited clusters and understand the spatial distribution of the findings or 'events' around clusters. When a row is selected within the cluster table and the map is plotted anew, this cluster is highlighted (Figure 7). The map can be downloaded as an interactive file in *".html"* format.

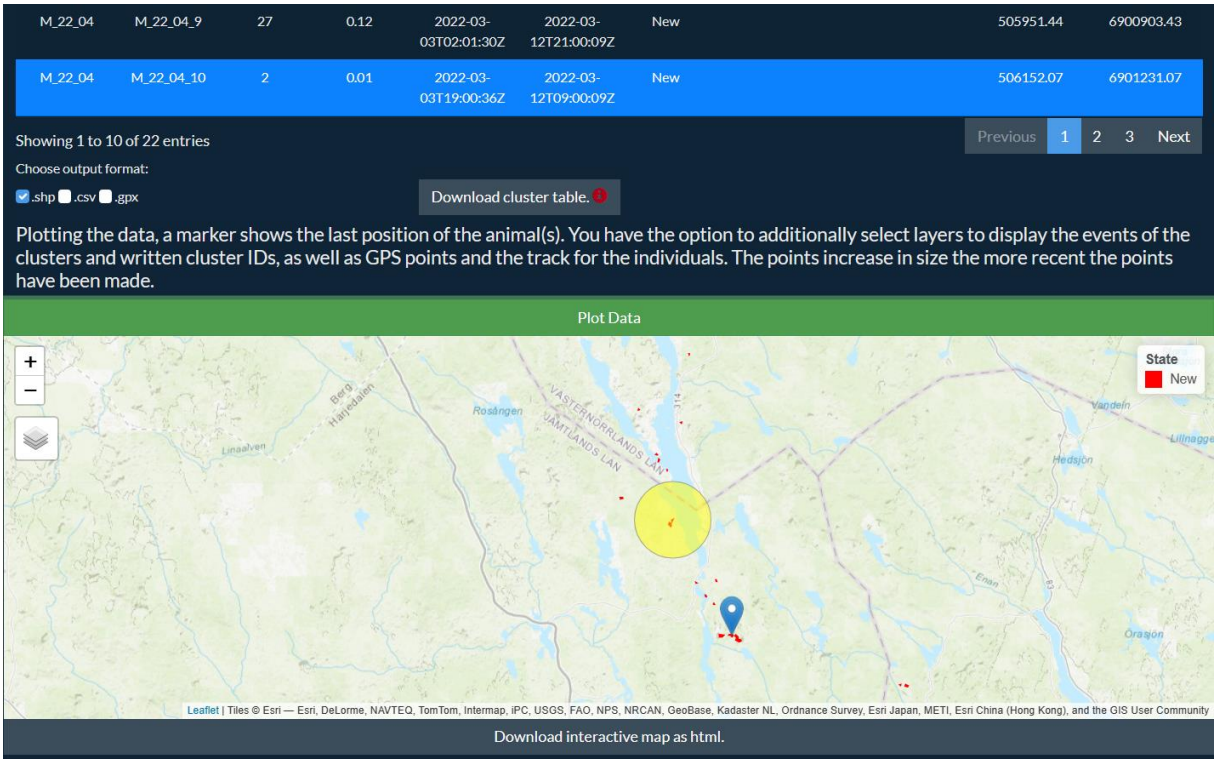

Figure 7: Output of the interactive map with the last row in the cluster table selected. This results in a light-yellow circle highlighting the cluster in the map.

The tab *"points table"* gives an overview of all GPS points that were used for the cluster analysis which can be displayed on the interactive map ("*GPS points*", Figure 8). The point ID is a combination for easier identification in the field and consists of the individual's ID, the cluster number it belongs to or SP for single points and then month, day, and hour the point was made. This way, the field personnel can get a fast overview of the distribution of the points within the visited cluster and might therefore by the experience of the species behaviour know what the event might be. The table can also be filtered to only show certain selections of the data. Again, the data can be downloaded as *.shp* or *.csv* in UTM and the specified zone or as *.gpx* file in WGS84. The shapefile will download all data, while the *.csv* and *.gpx* file will only download the selected data. This data is especially useful to load on a handheld GPS device or for any other form of orientation in the field.

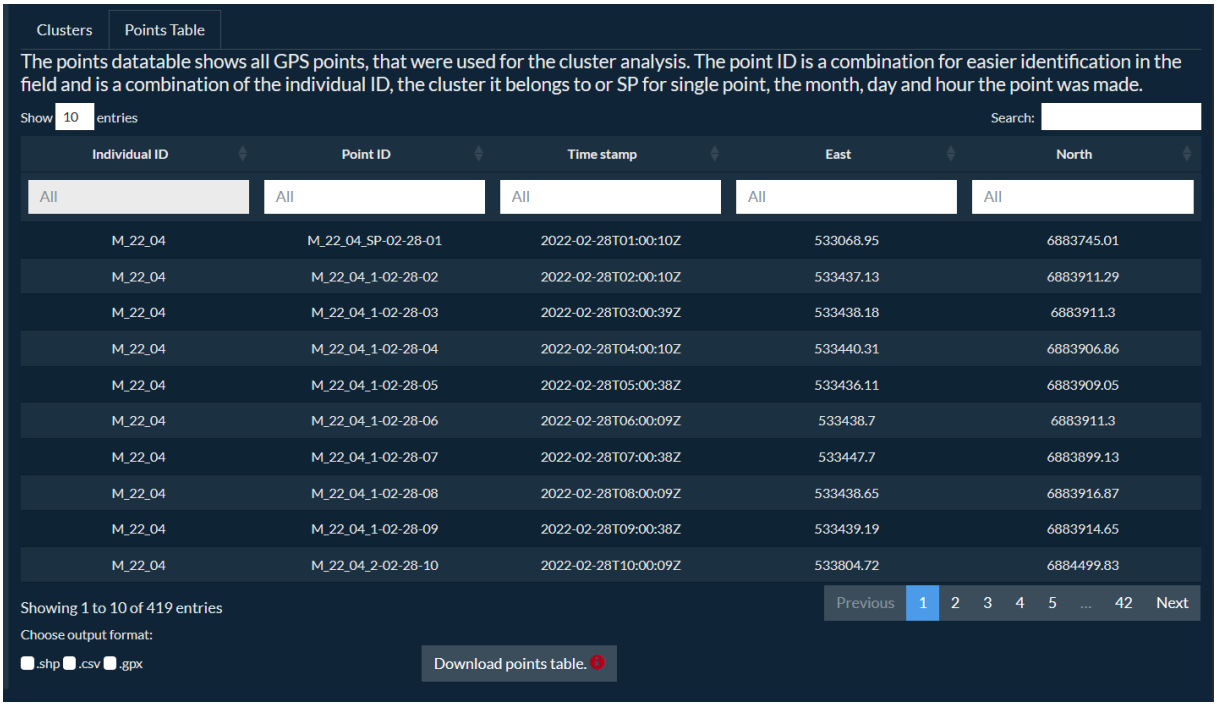

Figure 8: Output of the points table including all GPS points that were used for the analysis.

3. CONSECUTIVE ANALYSIS

The first analysis has now been performed and the data downloaded.

A common procedure for executing cluster analysis, is the up-to-date visiting of clusters that were made by GPS positions. This implies that field personnel will run cluster analysis every few days with the newest sent GPS locations and visit recent clusters as soon as possible after the individual was there. Therefore, it is common, that cluster analyses are run several times within one study period.

After the shapefile of the clusters is downloaded the first time, the next analysis that is done will be a consecutive analysis. New GPS data should always be saved within the same folder, replacing the old GPS data. The procedure starts the same, by uploading the newest GPS data from the folder, setting the correct column names and the settings.

Differences will appear for the third setting column within "Adjust cluster analysis parameters", which relates to earlier cluster analysis files within the folder of the raw GPS data. If the analysis is run for the first time or the program does not detect any shapefiles with the exact string it is searching for ("Clusters underscore label underscore date *.shp*") within this folder, the message will read "*No latest cluster file"* and will run the analysis as if done for the first time. Therefore, it is important to have an individual or group label, which is unique during the entire field study and have all files relevant within the same folder (see Figure 5 showing the folder for the wolf individual with the downloaded cluster shapefile labelled "wold_demo").

When performing a consecutive cluster analysis, it is important to check if a latest cluster file is found (Figure 9), as this file will make sure the earlier established cluster IDs remain constant even with additional points that might increase, combine or create new clusters. **That is why it is important to always download the shapefile!** There can be multiple old shapefiles in the sub-folder as the program will search for the most recent file. You can optionally automatically mark all clusters from the old cluster file as "Done", otherwise this can be done manually in the "Cluster Analysis Output" tab.

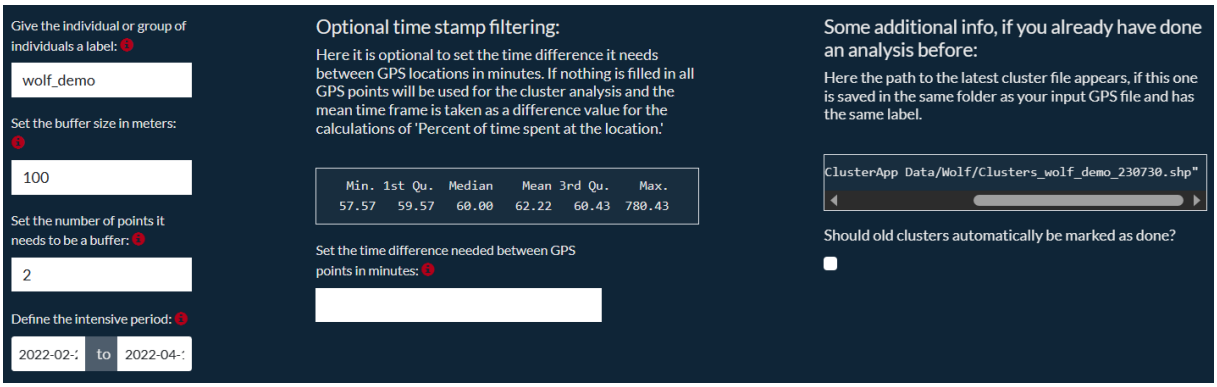

Figure 9: A previous shapefile is loaded into the third column if this can be found in the folder of the initial GPS file. The name follows a set structure, which has to include the label that was chosen. The app will always choose that shapefile with the latest download date.

Afterwards, the procedure goes on as explained for the first cluster analysis. Congratulations, you executed your first and second cluster analysis for a wolf in Sweden!

4. ANALYSIS ON MULTIPLE INDIVIDUALS

If you want to try out an analysis on multiple individuals, you can download the GPS data for the *"bears"* and save it in a folder of your choice. Again, it is good practice to now have a separate folder per group of individuals, where each new GPS file is stored or replaced. In this folder, you will also find the output of the cluster analysis.

All steps are identical to an analysis for only one individual. The *ID* column chosen in the step **Upload GPS data** should now include multiple IDs. The label set in **Adjust Cluster Analysis Parameters** is now a name for the group of individuals. We call this group of three bears *"multiplebears_demo"*. Further, we chose for a buffer size of 100 metres with 2 points needed to build a cluster within the study period May 2014 (Figure 10).

Figure 10: Setting the parameters for the group of bears.

The **Cluster Analysis Output** again shows the same output. Differences can be seen in the interactive map, which will now show the last location of the three individuals that are being followed. When

selecting to display the track and GPS points of the individuals, a new legend appears which shows the three individuals in different colours. The larger the points are the more recently they have been made, ending with the location of the latest position (Figure 11).

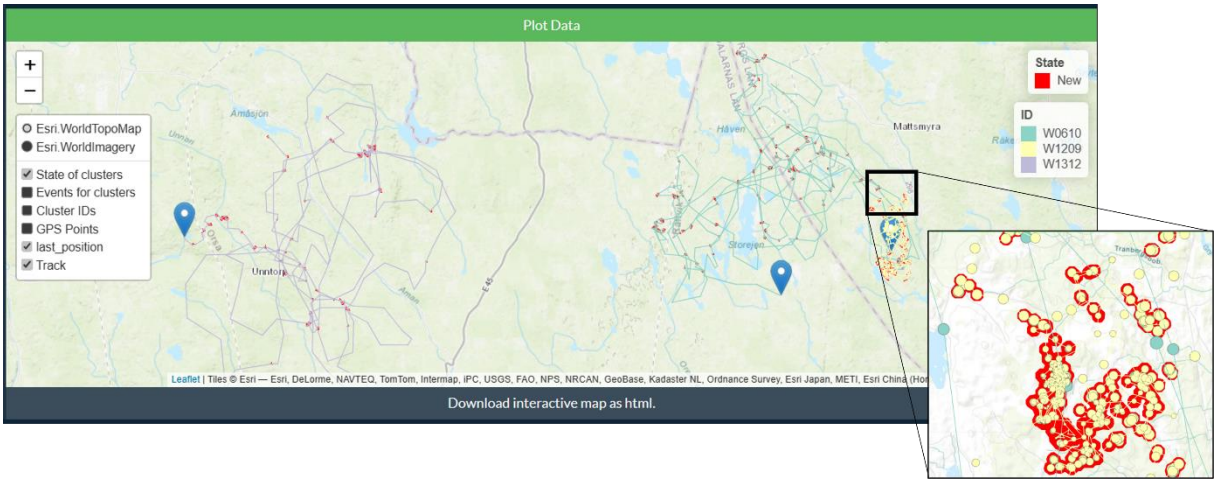

Figure 11: Cluster polygons are again all shown filled by their state. The track and GPS points of the three bears are visualized in the interactive map with different colours. Markers show the last position of each bear.

5. HANDLING GPS BURSTS

GPS collars are commonly designed to transmit GPS positions at regular intervals. Therefore, the default and the most common setting in cluster analysis is to use all GPS data points that are available within the study period. Some research projects recently explored the use of proximity sensors in collars to detect interactions between individuals of the same species or between different species. These proximity measures can provide insights into social behaviour, group dynamics or potential contacts between species. GPS data including proximity events will therefore include points at irregular time frames, which makes an additional identifier of these clusters necessary.

An example for this kind of data can be seen for a brown bear which was followed during a recent predation study in Scandinavia. This data cannot be made public within the package because the bear is still collared and active. Data will be made available once the collar has been removed.

Loading the data into app, the optional time stamp filtering already gives an indication that this data sends at irregular time steps with the mean at 26.4 minutes, however with a span of minimum 0.7 to 120 minutes. The minimum time frame between points is also visible in the density plot on the day of the 26th of May 2023, where a burst of GPS points was detected while during the rest of the period, fixes were constantly sent every half an hour (Figure 12).

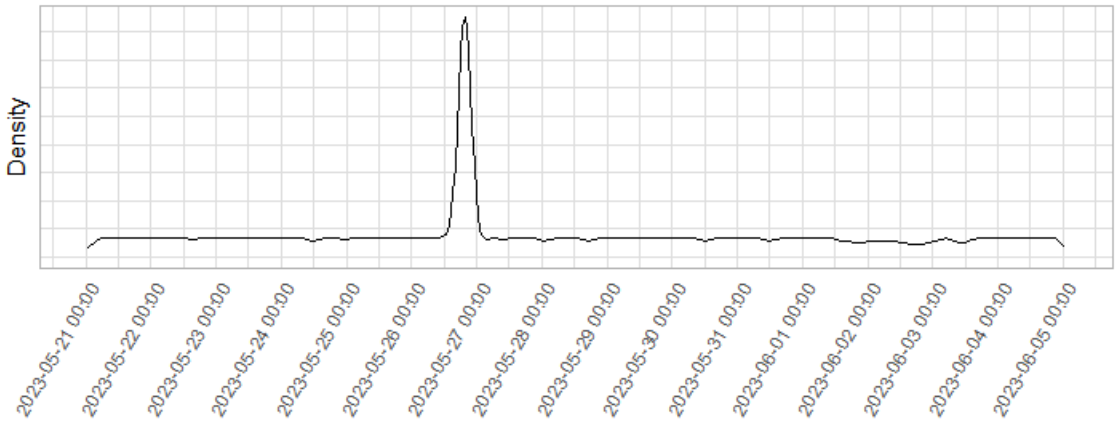

Figure 12: Density plot of send GPS positions over the time frame the individual was monitored.

In this case the user can now specify to only use a GPS fix every half an hour and discard all extra points, which will result in a study comparable as the examples above. However, if clusters of proximity data, are also interesting for the research question, all datapoints can be used for the analysis.

By doing so, the visiting identifier within the clusters table output gets more relevant: The visiting identifier (i.e., percent of time spent at the cluster) is calculated as the number of points the individual could have been at the cluster between the first and last time it visited (taking the mean time difference GPS points are taken) divided by the actual number of points within this cluster. This is only an approximation as the actual number of points that could have been sent can always differ from the calculated mean time difference. An identifier close to one implies that the individual spent most of the time between the first and last visit at this cluster, which can be related to for example bed sites used for a long timeframe. Values close to zero indicate that the individual only visited irregularly over a long timeframe, such as for example baiting sites. A visiting identifier larger than one implies that more points than the mean number of points possible in this timeframe were taken. This might relate to a burst of GPS positions in a short timeframe, such as proximity measures. Figure 13 shows the resulting cluster of the proximity event for the bear individual. The visiting identifier is at 24, with 145 positions sent between 18:26 to 20:50 on the 26th of May 2023.

While it is still important to visit clusters with an identifier value higher than one, it needs to be considered that these are not relatable to the clusters with the mean time difference between points. If the research question does not cover these specific clusters, it is advisable to set a time difference in the step cluster analysis parameters and exclude these clusters that developed from GPS bursts.

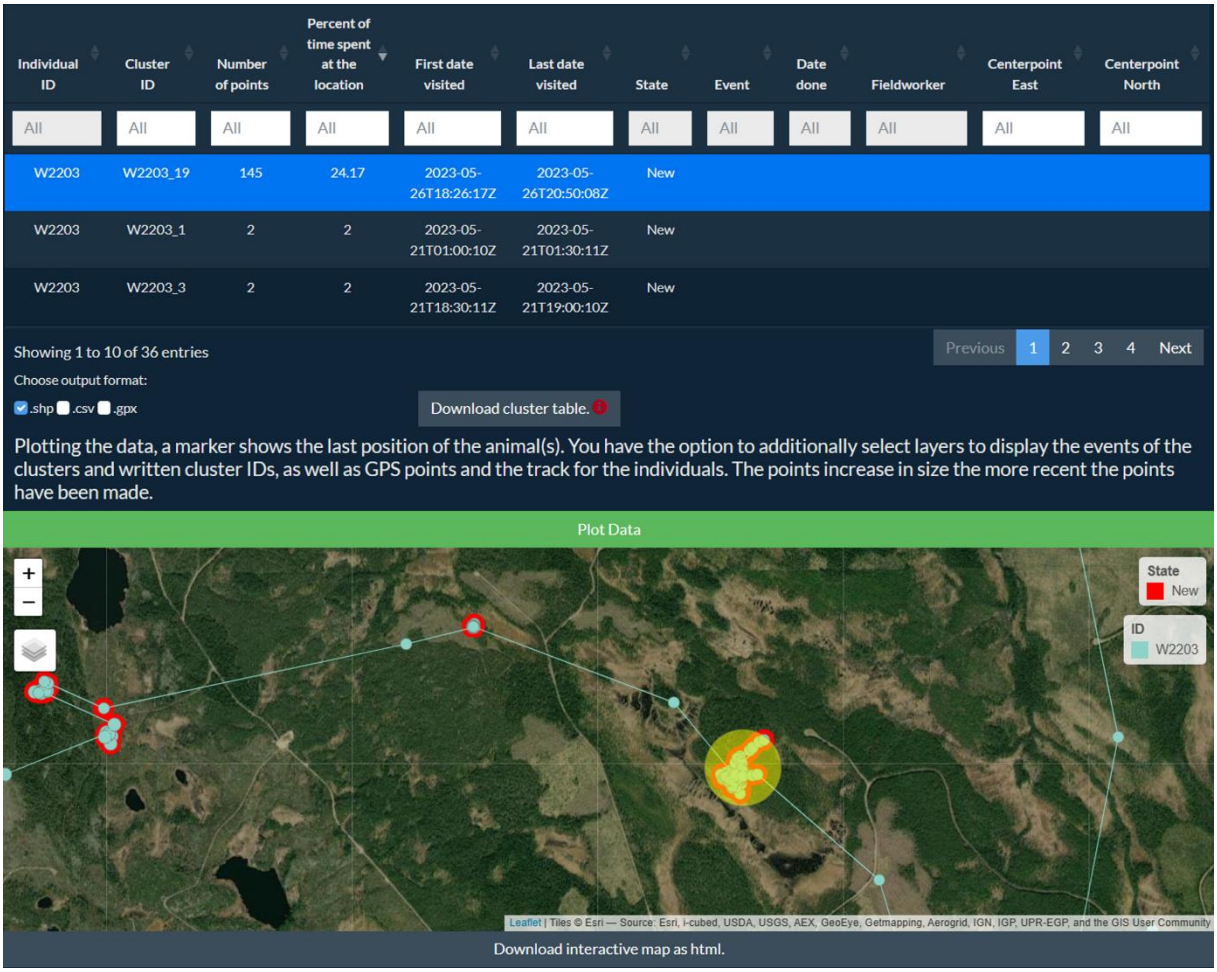

Figure 13: Output of clusters that include proximity events in the GPS data. The selected cluster developed out of a GPS burst, which can be easily recognised by the visiting identifier.

REFERENCES

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2023). _shiny: Web Application Framework for R_. R package version 1.7.4.1, https://CRAN.R-project.org/package=shiny.

Fay C, Guyader V, Rochette S, Girard C (2023). _golem: A Framework for Robust Shiny Applications_. R package version 0.4.1, https://cran.r-project.org/package=golem.

R Core Team (2923). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Wikenros, C., Di Bernardi, C., Zimmermann, B., Åkesson, M., Demski, M., Flagstad, Ø., Mattisson, J., Tallian, A., Wabakken, P. & Sand, H. (2023). Scavenging patterns of an inbred wolf population in a landscape with a pulse of human-provided carrion. Ecology and Evolution 13.