CROSS-CPP

Cross-CPP MARKETPLACE USER GUIDE

CPP DATA CONSUMER GUIDE

(CROSS-SECTORIAL SERVICE PROVIDER)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780167

Introduction

Service Providers User Guide describes Service Providers functionalities within Cross-CPP Marketplace such as discovering available data, requesting data from CPP owners, managing AEON subscription channels, understanding the utilities and how to use the analytics toolbox and its requests and results, the data views filtering utility, and to understand the use of the context monitoring and extraction module.

Purpose

This guide aims to help users from Service Provider companies on how to use the platform and give knowledge about the different functionalities available.

Audience

This guide is meant for and solely for users of Cross-CPP Marketplace with Service Provider role. Other roles can find their own user guides.

Scope

The contents of this guide are meant to be taken into consideration only when using the Cross-CPP Marketplace and will only cover functionalities meant to be used by the role stated above.

Cross-CPP team does not take responsibility on bad use of the application or the data provided when not following the instructions given in this guide.

Troubleshooting

For any questions or inquiries about the use of the Cross-CPP Marketplace web application or the contents of it or this guide, , or if you find there is no content in this guide for some functionality please forward it to: marketplace-support@cross-cpp.eu.

Contact

Cross-CPP Project website: https://cross-cpp.eu

Cross-CPP Marketplace: https://datagora.eu

Marketplace support: marketplace-support@cross-cpp.eu

Context Monitoring and Extraction Module (CME): context-support@cross-cpp.eu.

Contents

Guide

There are two main functionalities in the Cross-CPP Marketplace (Figure 1):

- **Catalogue**: list of all available registered signals and their assigned measurement channels. (2.1 Catalogue)
- **Data Discovery**: discover the data available in the Cross-CPP Marketplace through a set of filters to create your Data Requests for the data you are interested in. (2.2 Data Discovery)

Cross-CPP Marketplace offers two distinct sections for Service Providers, grouped in the Service Provider Wallet, apart from the main functionalities (Figure 2):

- **Data Wallet**: is considered as the customer central point where they can take the control of their user profile options, data subscriptions and entity information. (3 Service Provider Wallet)
- **Toolbox**: offers a series of tools to get different metrics on the data received by users for your active data requests. (4 Toolbox)

At a glance, Cross-CPP Marketplace main functionalities can be access directly through the side menu

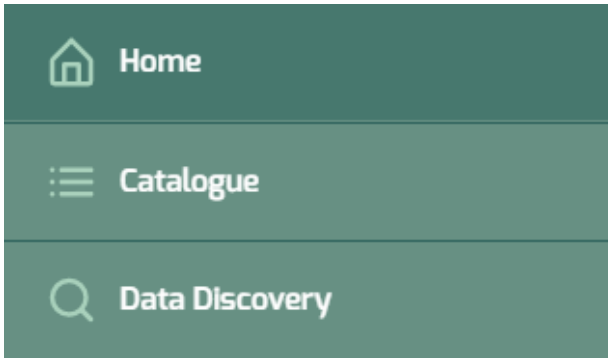


Figure 1. Main functionalities menu

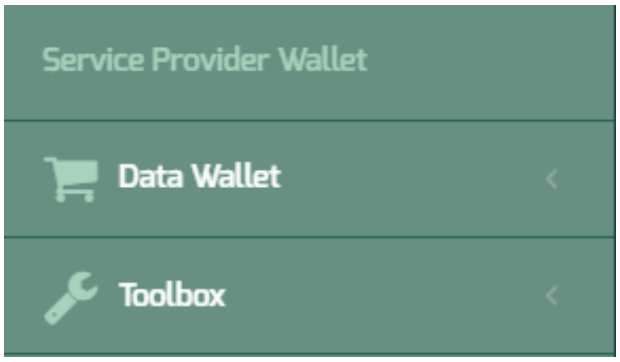


Figure 2. Service Provider Wallet menu

Other user related functions can be found in the top right corner of the toolbar which shows the F.A.Q. and a menu including Marketplace *Profile*, User *Settings* and *Log out.*

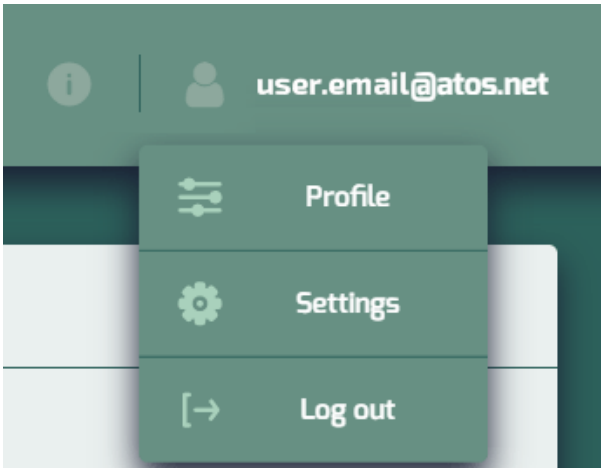


Figure 3. User menu

1. How does Cross-CPP Marketplace work?

1.1. Basic knowledge

Cross-CPP Marketplace works under the premise of letting Service Providers decide which data they want to receive and subscribe to, while giving CPP Owners control over their data at all times.

This leads to the following premises:

- Service Providers request specific data through a set of filters given, creating Data Requests
- CPP Owners decide which Data Requests they want to accept based on the filters decided by the Service Provider
- Any number of CPP Owners can accept any number of Data Requests

Data Requests are further explained in section 3.1 What is a data request?

Cross-CPP Marketplace offers two ways of getting data:

- Pull mode: traditional API request using the Data Requests identification
- Push mode: subscription to an AEON channel to get data as soon as it is available in the system

Both approaches are further explained in section 3.3 Mechanisms to subscribe to Cross-CPP data requests (pull & push mode).

Cross-CPP Marketplace offers a Toolbox including a series of Analytics and Data views. These act as a service themselves and allows Service Providers to use data coming from Data Requests the way they could do in case they have those tools in their own systems, while saving the time and effort to configure them. The Toolbox is further explained in section 4 Toolbox.

1.2. Step by step

In a nutshell, these are the steps on how to work within Cross-CPP Marketplace and the possible interactions:

1. User creation
    a. Service Provider registration
    b. System Administrator validates Service Provider registration
2. Discovering data
3. Getting data
    a. Create Data Request
    b. Subscribe to provided AEON channel for that Data Request
    c. CPP owners accept Data Request
        i. If subscribed, Service Provider would receive data as soon as it is available in the system
        ii. Service Provider can always request data collected within Data Request via API request (see Service Provider Developers Guide)
4. Analyse data
    a. Create analytics from Data Request
        i. If one-time analytics results will be instantly delivered and displayed on screen
        ii. In case analytics are services that consume data delivering results over time an AEON channel will be provided to subscribe to
    b. Delete analytics

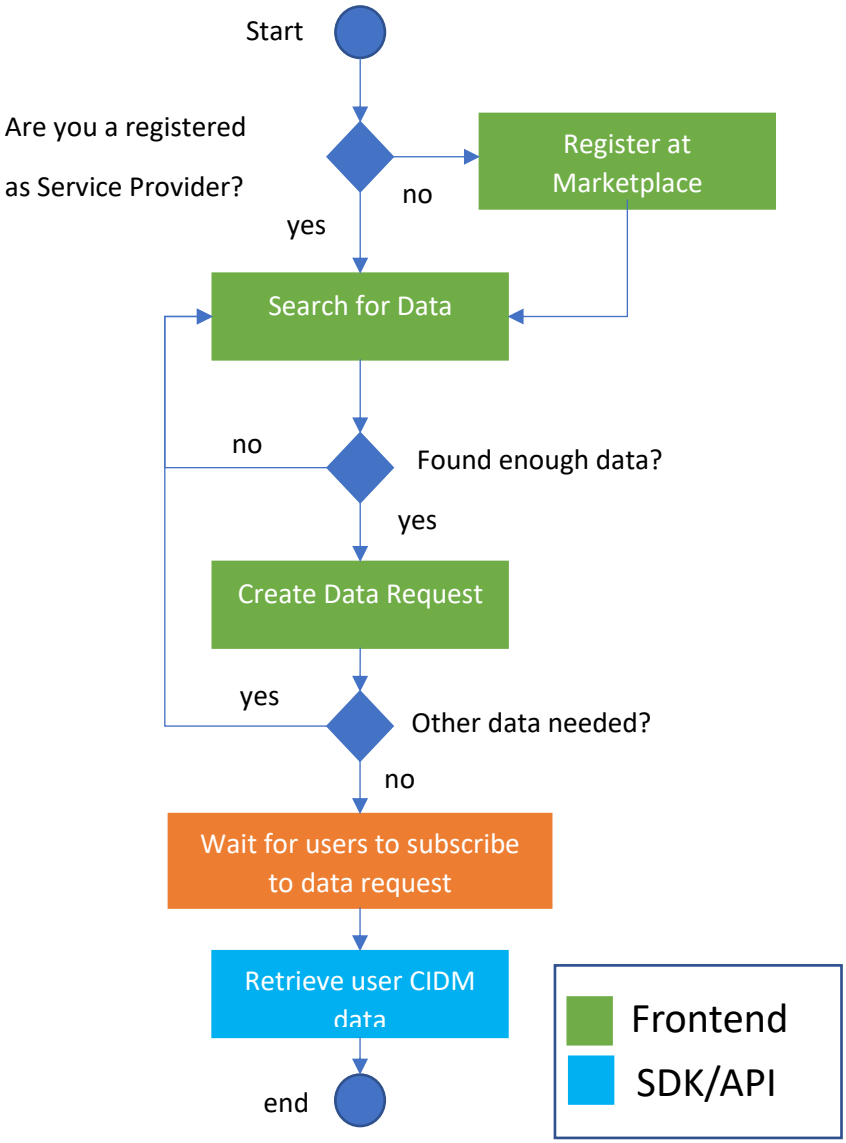


Figure 4. Cross-CPP Marketplace general workflow

1.3. Registration at Marketplace

In order to identify towards the Marketplace, Service Providers need to create an account. In the Sign up! button to create a Service Provider account request.

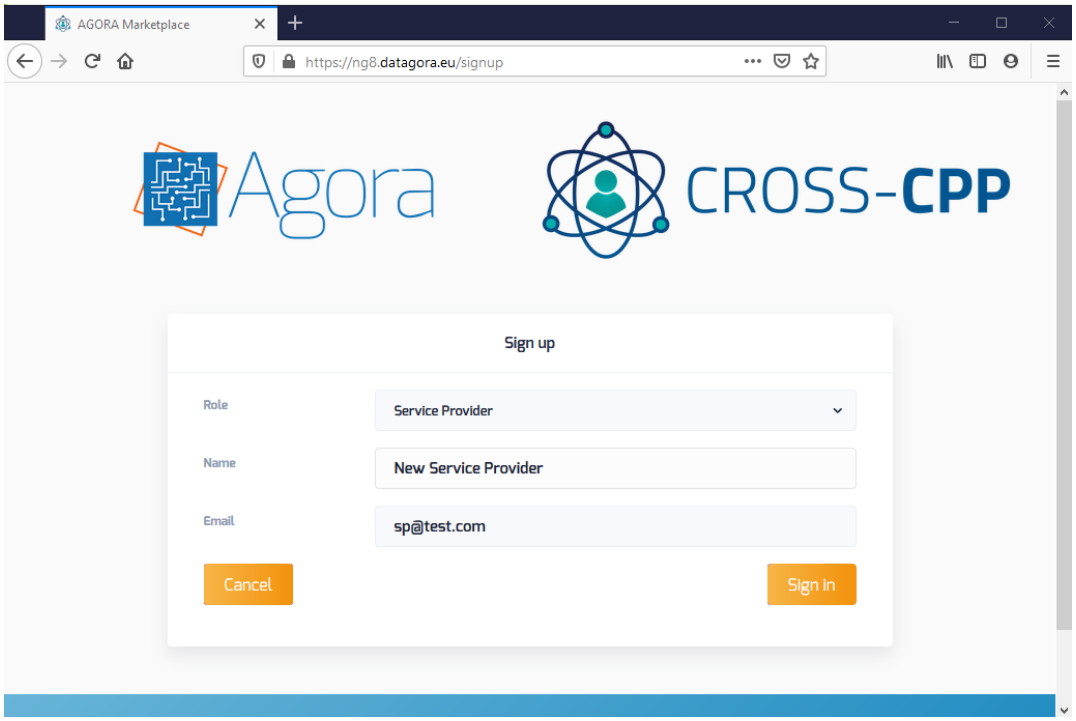


Figure 5. Cross-CPP Marketplace sign up form

After filling the sign-up form, the marketplace administrator will receive a notification to validate the service provider. Once the Service Provider user is validated, the user receives the credentials for Sign in into the Marketplace.

2. Main functionalities

2.1. Catalogue

The Catalogue (see Figure 6) shows a list of all available signals under a harmonized data model called Common Industrial Data Model (from now on CIDM – see F.A.Q in section 8) which enables to retrieve data from Data Providers (CIDM has a dedicated guide to be found in cross-cpp.eu). The data in the Cloud Storage is stored following the CIDM which harmonizes the proprietary OEM formats into brand independent data formats (made then available in the catalogue of the marketplace).

In the Catalogue, a Service Provider can search and filter by different attributes and see the channels that use each signal. Below, the filters which appears in Signals Catalogue are explained (see Figure 7):

- **Signal**: the name of the signal describing what is measured.
- **Signal type**: type of the signal data. Following the CIDM this can be "numeric", "enumeration", "information" or "general-purpose".
- **Unit**: units of the data
- **Related Channels**: data channels that use the data received from this signal. In case there is more than one that use the same signal, a list is available to select which one is of interest.
- **Channel ID**: identification of the selected channel that receives data from the signal
- **Channel Type**: selected channel type of data receiving. Following the CIDM this can be "time-series", "histogram", "geo-histogram" or "general-purpose".
- **CPP Type**: from which type of CPP this signal receives data from. At the moment this can be from vehicles or buildings.

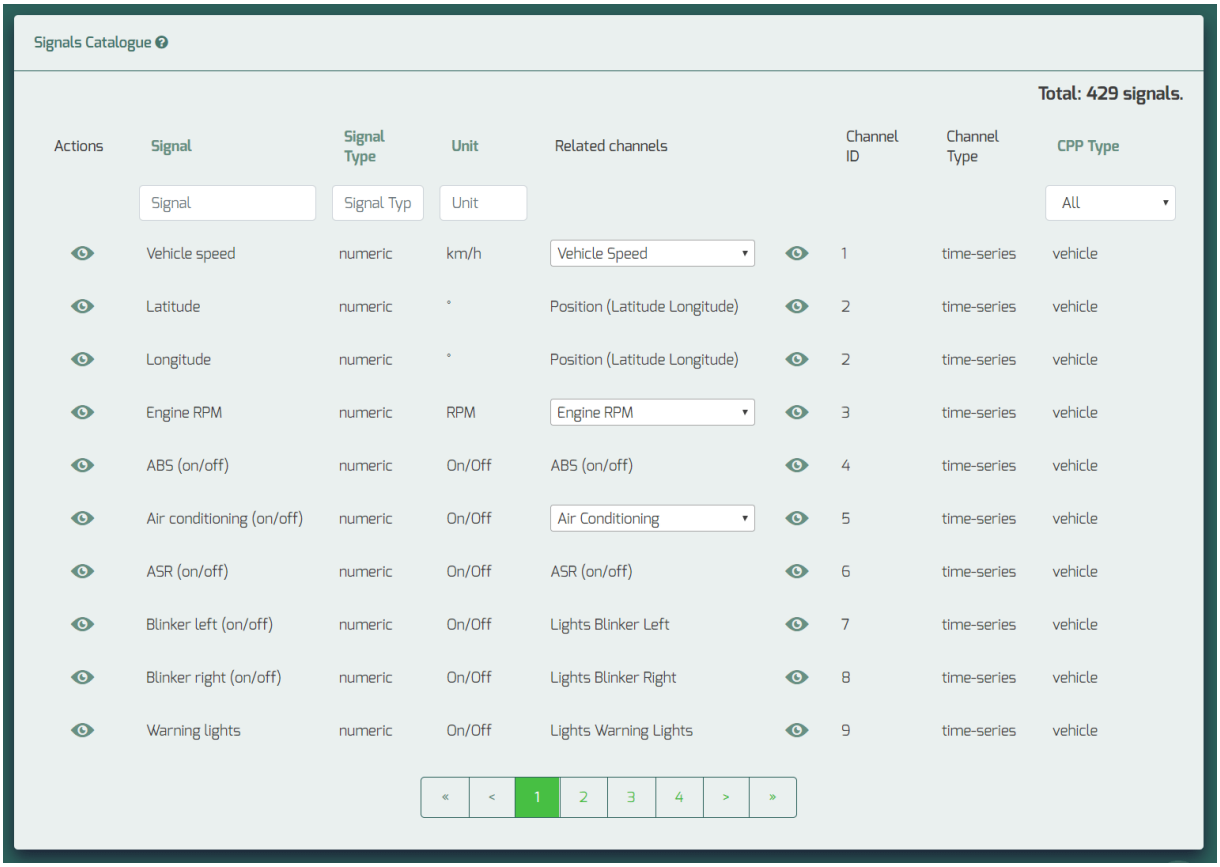


Figure 6. Signals catalogue

Also, each signal and channel have its own detailed view with all available information, including direct links to each other.

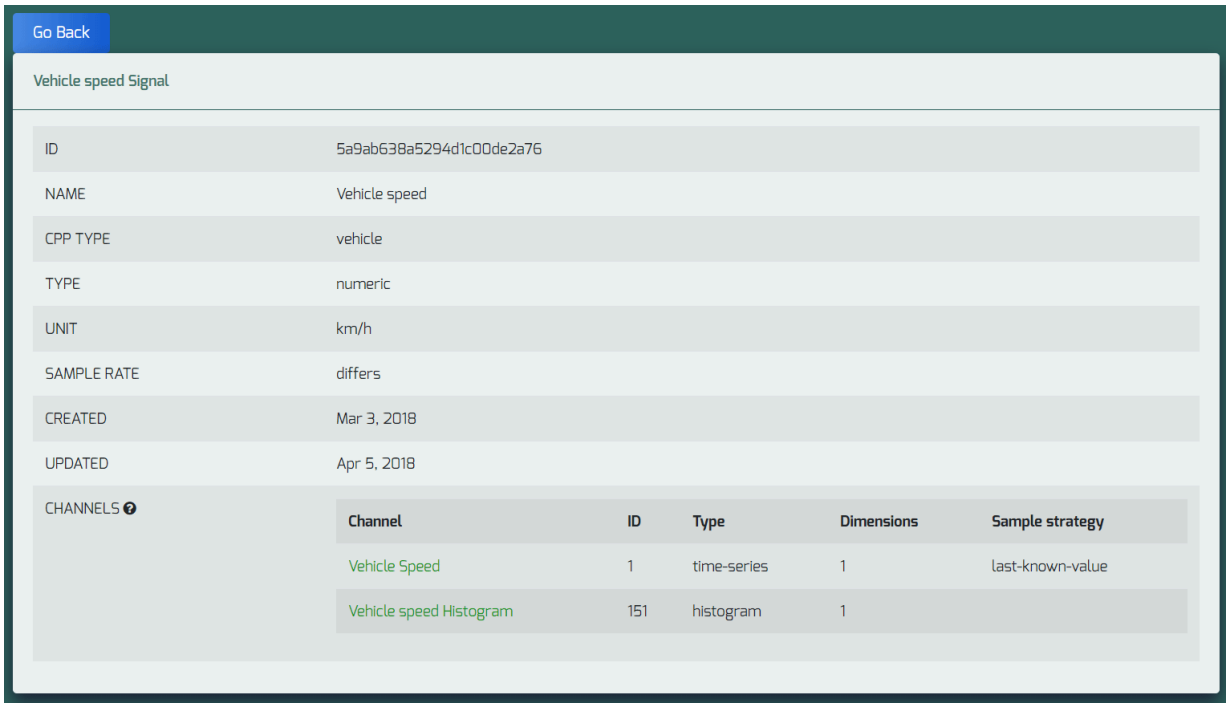


Figure 7. Signal details

2.2. Data Discovery

The Data Discovery offers Service Providers a tool to select which data do they want to receive from CPP owners. A wide variety of filters are provided in order to help Service Providers to narrow their desired results.

In a first step, the Service Provider has to select the signals for the data of interest. If none is selected, a search for all channels will be performed. Once a signal is selected the channel that uses that signal will be added to the selection list. Selected channels can be removed from the list at any time.

If one or more channels are selected suggestions can be requested. This is possible thanks to the implementation of the Context Monitoring and Extraction module (CME) (check additional information in F.A.Q. in section 0) within the Data Discovery process. In case suggestions are requested, the CME will provide a list of channels related to those already selected[1]. Channels of interest can be added directly from the suggestion list. Also, suggestions can be requested again each time a new channel is added.

Once the channels are selected, the user may apply several filters to the data packages (see Figure 8):

- **Date and duration** section filter the results by data package submission date, measurement recording date and measurement time duration.
- The **geographical selection** offers a way to narrow results to view data only from a specific geographical area. This can be set selecting a country in the list, or drawing a customized area entering the area bounds coordinates.

[1] Suggestions are provided based on a context model that is the result of analysis of the physical relation between signals (e.g. in case temperature is a selected signal, the CME would suggest the humidity and sun intensity

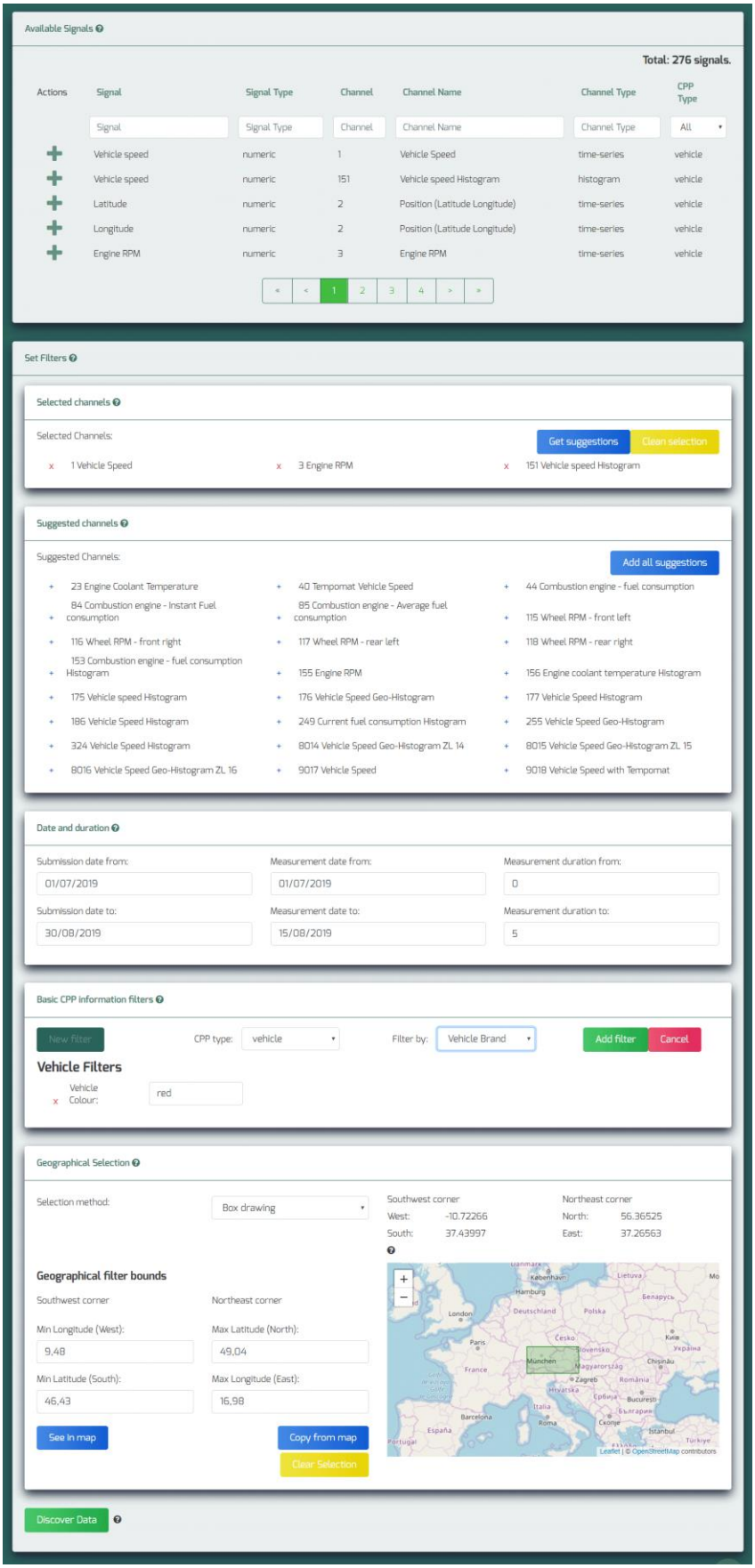


Figure 8. Data Discovery filters

Once the filters are set, Service Providers can see the data they would receive from a data request with selected configuration (See Figure 9). In this process only metadata is used in this process and no real data can be shown or retrieved through this functionality:

- A box with some **general statistics** shows the total of CPP owners and the number of selected channels that contains any data, and the number of data packages retrieved in the discovery.
- For each CPP type with data within selected channels a box is displayed showing the number of entities and channels with data, and the minimum, maximum and average duration of its measurements.
- The **heatmap** offers an overview of the geographical areas that contain data, in case it is desired to narrow the discovery to a specific area.
- The **pie charts** show how the data is distributed across the different channels. Sectors can be hidden if needed.
- The **line and bar charts** display the amount of data packages retrieved depending on its submission date and time duration.

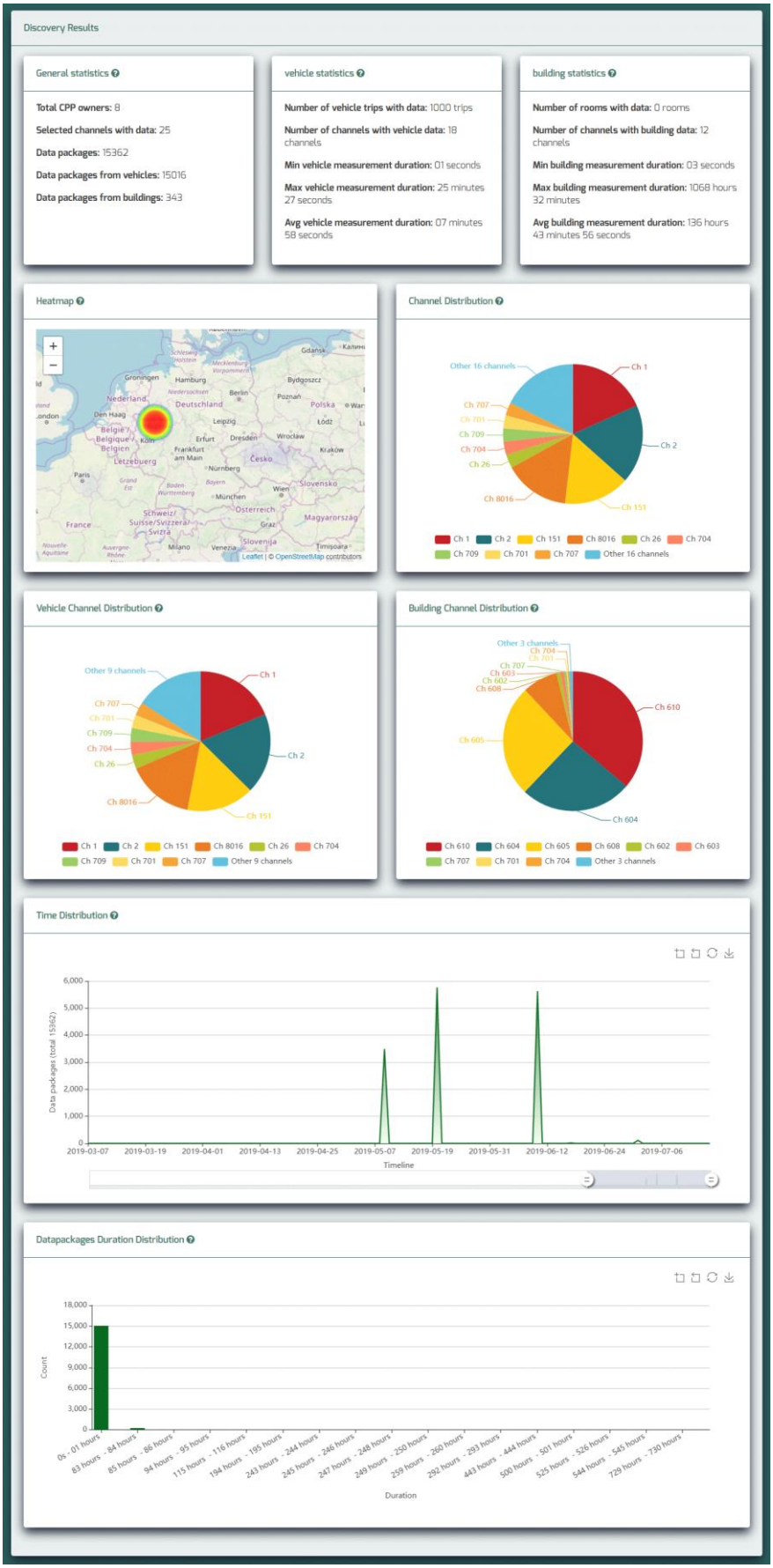


Figure 9. Data Discovery results

If the Service Provider is not satisfied with the search results the filters can be modified without having to restart the process from scratch. Service Providers can adjust the filters and browse the data until they find what they were looking for.

Once Service Providers agree with the data they would receive, they can configure a set of additional options (Figure 10):

- **Analytics**: by activating this option the Service Provider declares the intentionality of using the data to be received for analytics, so the owner can give the consent to do so (see more details in section 4 Toolbox).
- **Context**: by activating this option the Service Provider declares that he wants data to be received to be analysed by its context and to receive data fitting selected context filters. A list of filters will be provided to configure the context filter for the Data Request, such as day of the week, meteorological condition, and so on (see more details in section 3.4.1 Context Sensitive filtering). Service Providers will then receive only the data fitting those filters produced by owners that allow the context analysis.

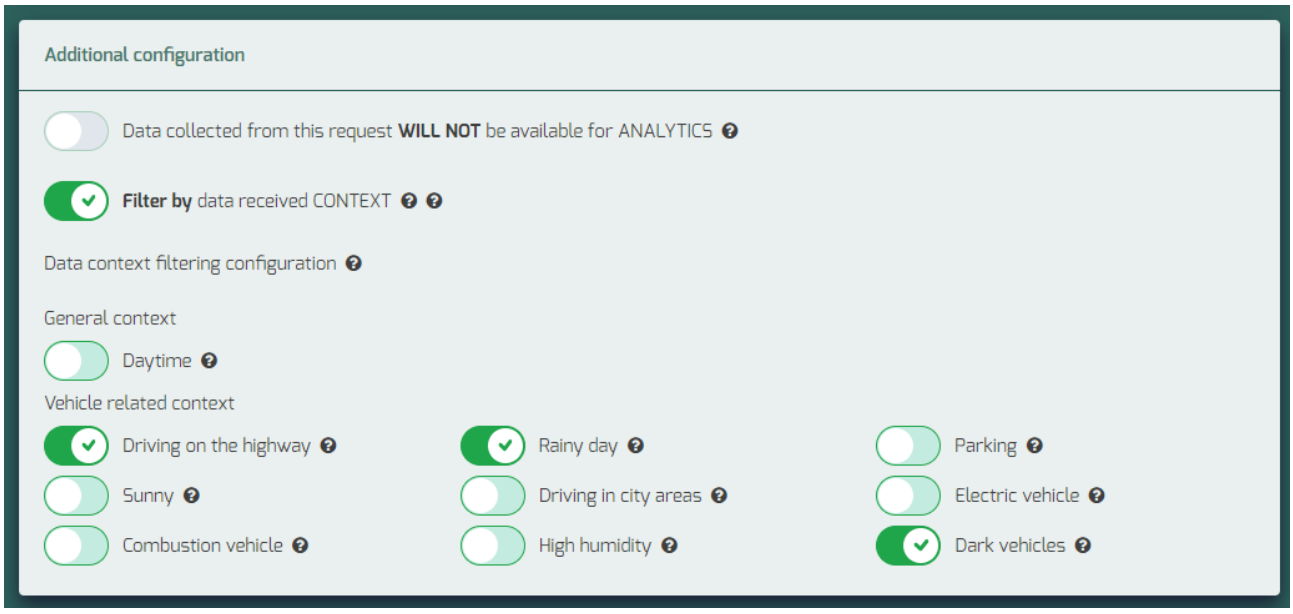


Figure 10. Data Discovery – Data Request additional configuration

Finally, they can set a description name for the data Request and indicate additional options for the data they will receive.

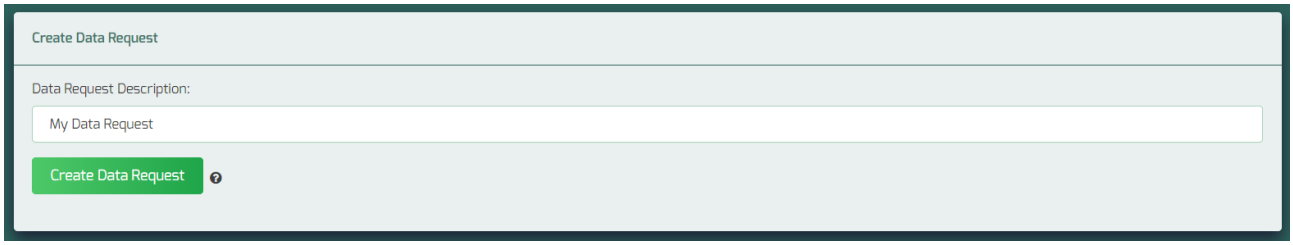


Figure 11. Data Discovery - Create Data Request

3. Service Provider Wallet

The Service Provider wallet (from now on Data Wallet) (see Figure 12) group different views for data review in which the user can:

- Data Requests:
  - List published data requests
  - Review each data request details
  - Find each data request AEON channel configuration
  - See the data packages received from each data request
- Data Transactions:
  - Get a resume of each data request package transactions

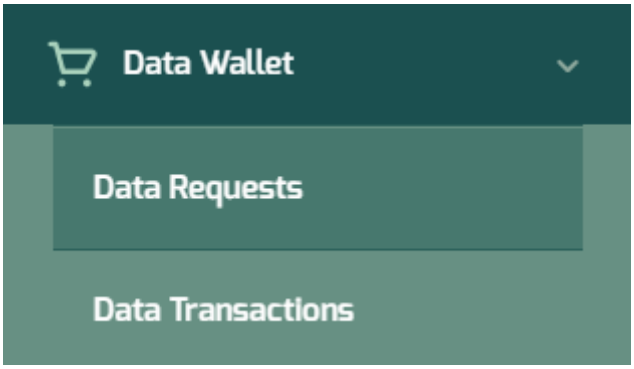


Figure 12. Data Wallet menu

3.1. What is a data request?

Data Request stands for a configuration set by a group of filters in order to receive specific data within the desired scope of our services designing. This scope is set during the Data Discovery (see section 2.2 Data Discovery above) and saved when creating the Data Request.

Once the Data Request is published by the Service Provider, the CPP owners can then review and accept it, and thus consent to deliver their data through this specific data request and to this specific Service Provider.

Service Providers will receive all data from CPP owners that accepted that request and that is exclusively for the scope of the Data Request published.

In the Data Requests view a table with published Data Request is shown. Selecting one will lead to a detailed page in which the AEON subscription channel ID and URL can be consulted, along with the full configuration and a table of known sent data packages.

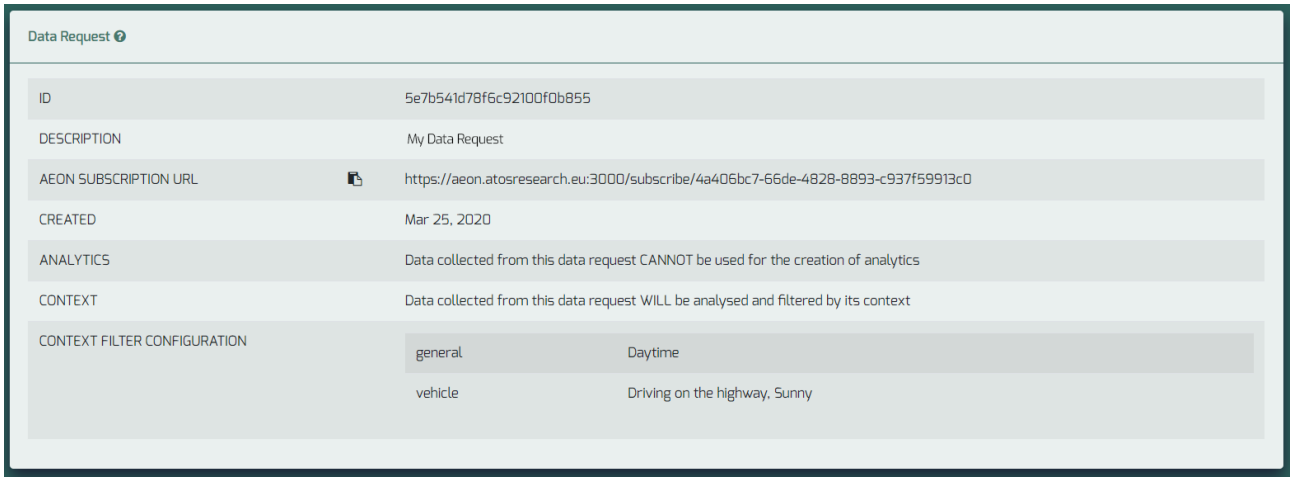


Figure 13. Data Request details

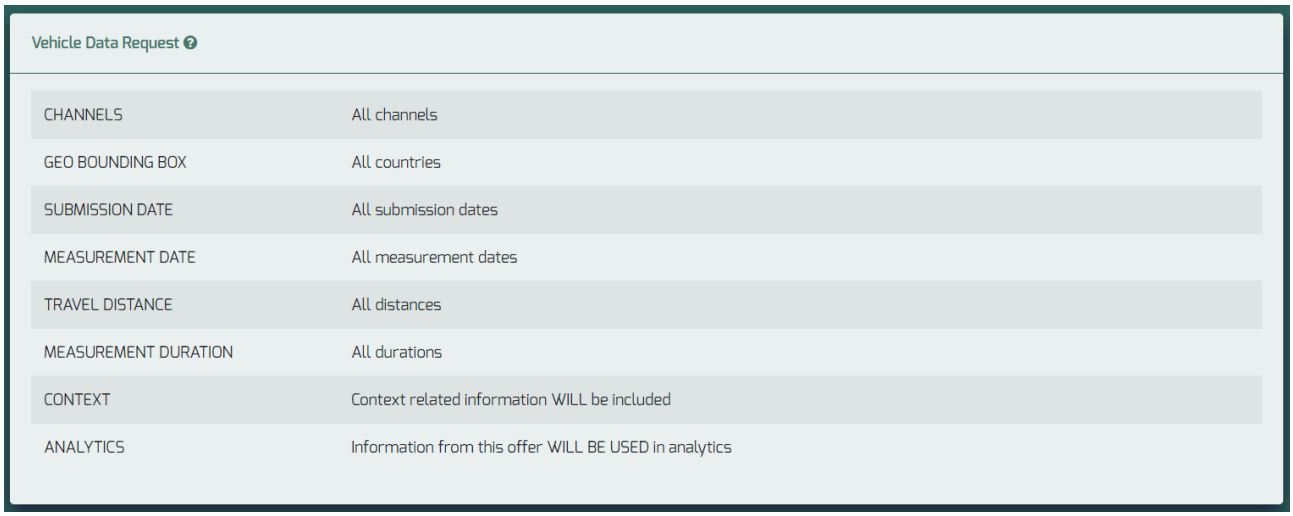


Figure 14. Data Request filter details

3.2. How to publish a data request?

A data request is configured through the Data Discovery process. For more information about the Data Discovery please go to section 2.2 Data Discovery above.

3.3. Mechanisms to subscribe to Cross-CPP data requests (pull & push mode)

There are two possible methods to retrieve data received through the published Data Requests.

3.3.1.1. Pull mode

Data can be manually pulled by the Service Provider from the Marketplace.

This mode relies on the Service Provider to get the data from an API endpoint. More information about this can be found in the Service Providers Developers Guide (to be found under cross-cpp.eu).

3.3.1.2. Push mode

Data is automatically pushed from the Marketplace to the Service Provider through a provided AEON channel.

The AEON channel subscription URL can be found in the details view provided in Data Requests view (Figure 15). The clipboard button can be clicked to copy the content of each text.

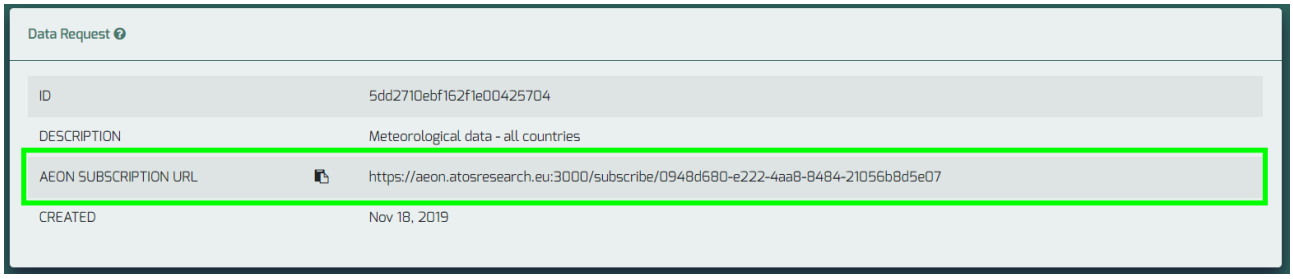


Figure 15. Data Request AEON channel details

Information about AEON configuration can be found in the Service Providers Developers guide (to be found under cross-cpp.eu).

3.4. Additional Data Request configuration

There are some additional configuration options available during the creation of the Data Request such as context sensitive filtering and declaration of purposes to which the CPP owner must give consent in order to approve sharing data.

3.4.1. Context Sensitive filtering

During the Data Discovery process, the Service Providers may apply special filters to the data packages that they want to receive. These filters are based on the context information that can be extracted by the Context Monitoring and Extraction (CME) module from the data provided by Data Owners (the CPP data) as well as CPP basic information that is not CPP runtime monitored data but basic information such as vehicle brand, model or colour.

The extracted context information is inferred from the collected data, the context models are developed for the CPP type in question, by combining several signals and applying rules to the data. An example of such extracted information is to infer if the data is collected while a vehicle is on a highway (or not) or if it's raining (or not).

These data filters will be included in the Data Request and applied to the data being collected at runtime and filtered before being sent to the Service Provider.

3.4.1.1.What is the Context Monitoring and Extraction module (CME)?

The CME is the module that by monitoring CPP data streams, basic CPP information and context models of the different CPPs being handled by the Cross-CPP Marketplace, can provide extracted context to other Cross-CPP modules for semantic adaptation, such as the here explained context sensitive filtering or for the Context Sensitive Security (see section 5.1 Context Sensitive Security).

More details can be seen in the Cross-CPP Ecosystem guide (to be found under www.cross-cpp.eu).

3.4.1.2. How to configure the context sensitive filtering for a Data Request

Service Providers may apply special filters to the data packages that they want to receive. In case the Service Provider would like to further refine the data to be received from the CPPs this can be achieved by selecting the option "Filter by data received CONTEXT" in the UI of the Marketplace within the *Additional configuration* tab. Once this is selected all possible context parameters that may be extracted once the Data Owner accepts the Data Request (see Figure 16).

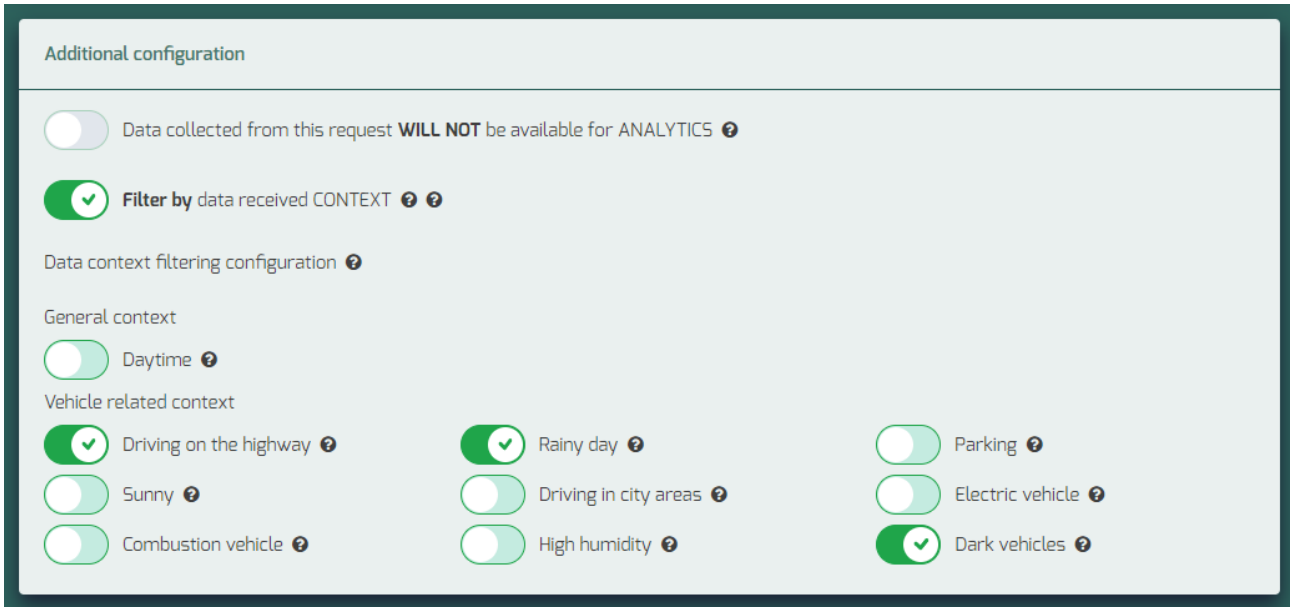


Figure 16: Context information filters

These data filters will be included in the Data Request and applied to the data being collected at runtime and filtered before being sent to the Service Provider.

Extracted context parameters can be extended by creating new context models and adding/editing the current CME rules. More information on this possibility can be seen in the Service Provider developer guide or by contact directly the CME team at context-support@cross-cpp.eu.

3.4.2. Data Views

Data Requests can be further filtered to receive specific values through dedicated AEON channels to be used as notifications or alerts, or to generate different services. For more information please see section 4.1 Data Views.

3.5. Data Transactions

Data Transactions view displays a summary of data collected from the published Data Requests, including the following information:

- total transactions
- amount of pull type transactions
- amount of push type transactions
- total size of available data
- first package received
- last package received

The details view leads to the Data Requests detail view (see Figure 13). Data Transactions view is depicted as follows including the following fields (see Figure 17):

- **Data Request**: The name of the Data Request the transactions belongs to
- **Contract transactions**: total number of transactions from users which accepted this data request.
- **Pulls**: number of times the data has been pulled
- **Pushes**: number of times the data has been sent through the AEON channel
- **Total size**: number of bytes between all transactions

Figure 17. Data Transactions list

4. Toolbox

4.1. Data Views

4.1.1. What is a Data View?

A Data View is a configured filter of values received from a Data Request measurement channels. This filtered data is sent to the Service Provider in a dedicated AEON channel in order to provide it as a notification mechanism.

The Data Views can be used also to request the latest data available for the configured value filters.

Data Views provide the primary mechanism for providing training data in the process of building Machine Learning analytics functions.

4.1.2. What is a category and how to assign it?

A category is a way to use a Data View to analyse its collected data. It can be assigned when creating the Data View or later at the details page. Once a Data View has been assigned with a category, its rows can be assigned with a value from that category to identify the type of information that row provides.

This is a mandatory step in order to use a Data View for a Machine Learning model classifying input data into the defined categories.

4.1.3. How to create a Data View

The creation page can be found through the button above the Data Views main view.

Data Views filter data coming from already existing Data Requests, meaning the Service Provider must first select the Data Request from which data shall be filtered.

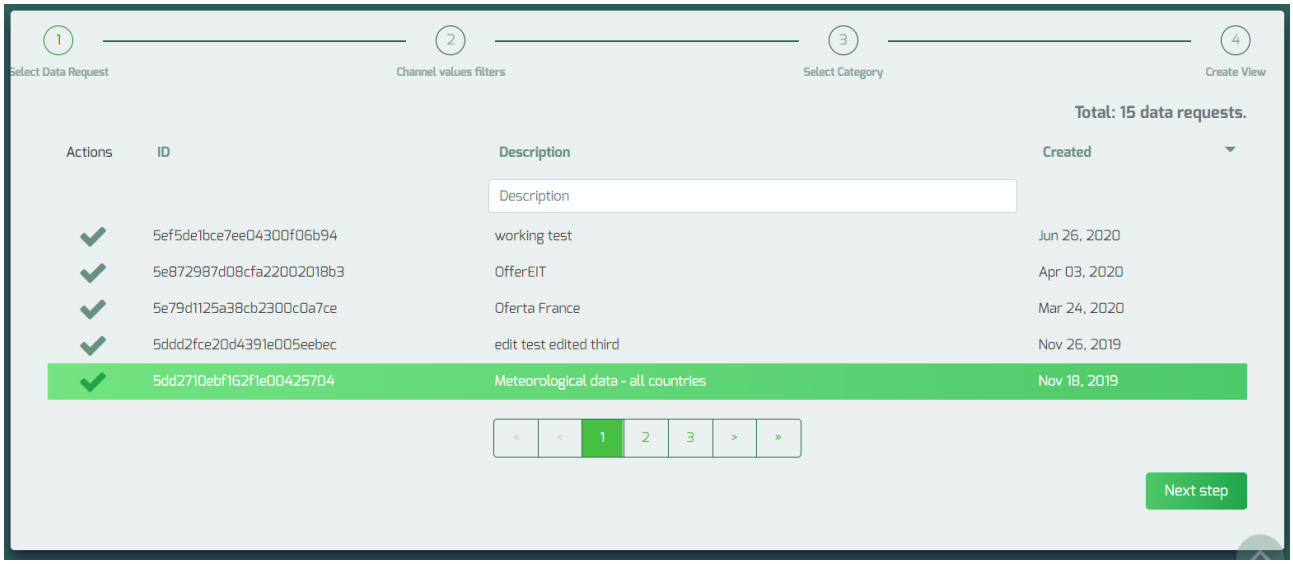


Figure 18. Create Data View – Select Data Request

Every Data Request comprises one or more measurement channels from which the request receives data. Next step is to decide from which of those measurement channels value filtering is desired.

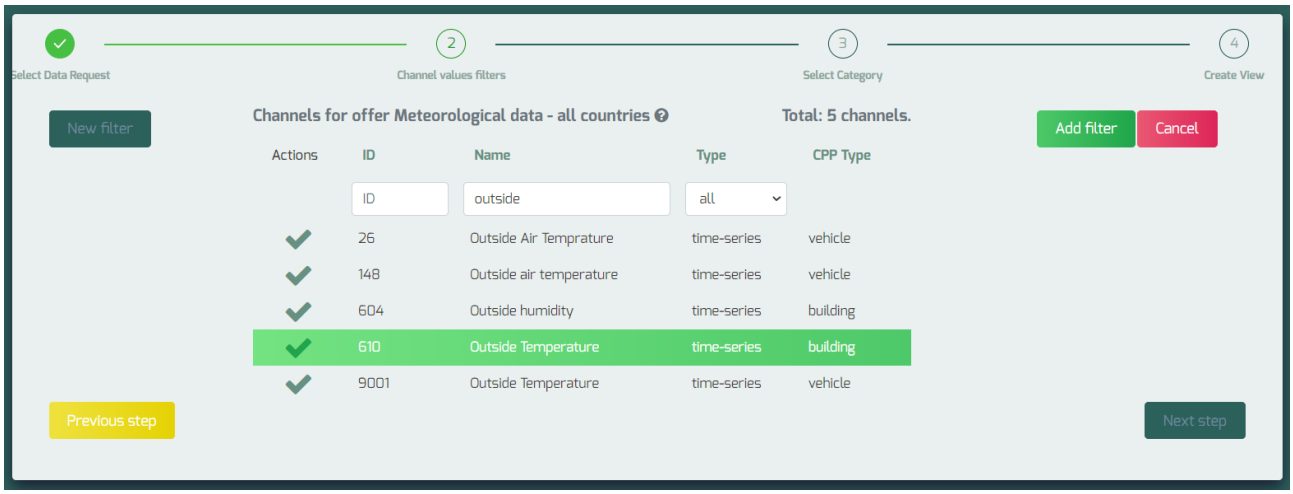


Figure 19. Create Data View - Select measurement channel to be filtered

A value filter consists of an operator and the actual value that must be configured for every selected measurement channel. Notice that a measurement channel can have more than one value filtering, such as less than a value but more than other value, but if that configuration is not possible the system will not let the user continue.

There is also the possibility that all data coming from the selected measurement channel is of interest, in which case the filter can be left empty.

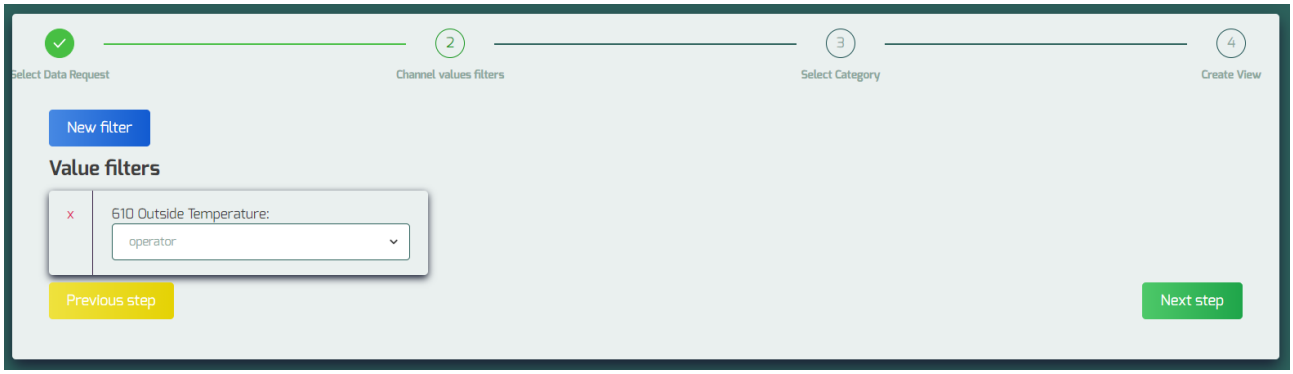


Figure 20. Create Data View – Measurement channel without value filter

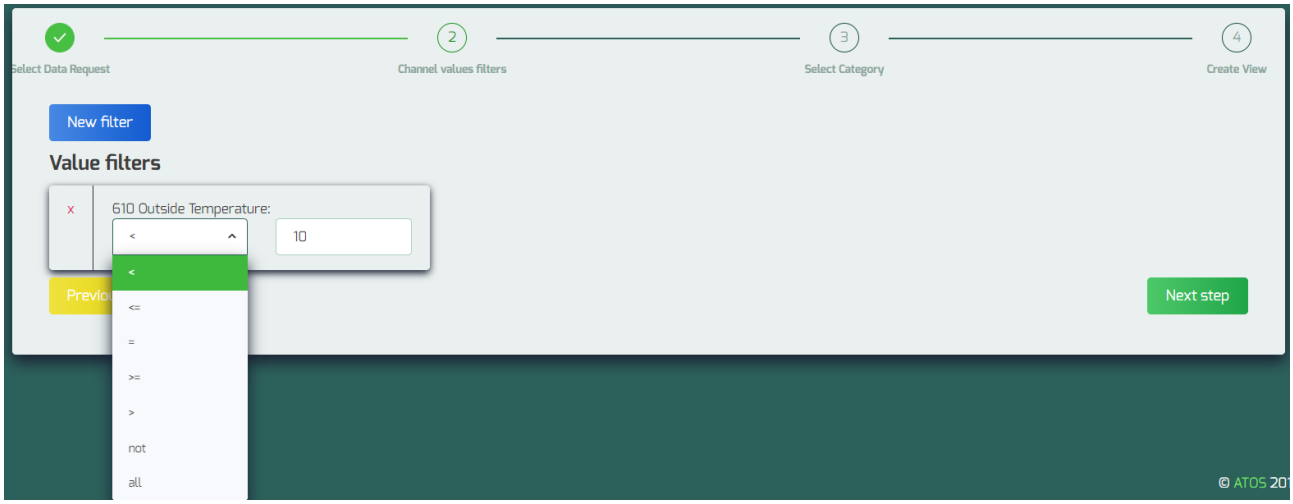


Figure 21. Create Data View - Input measurement channel value filter

Once agreed with the selected filtering, it is asked if a category is desired for the Data View. The category can be assigned now or later at the details page. The user can select an already created category or create a new one at this moment.

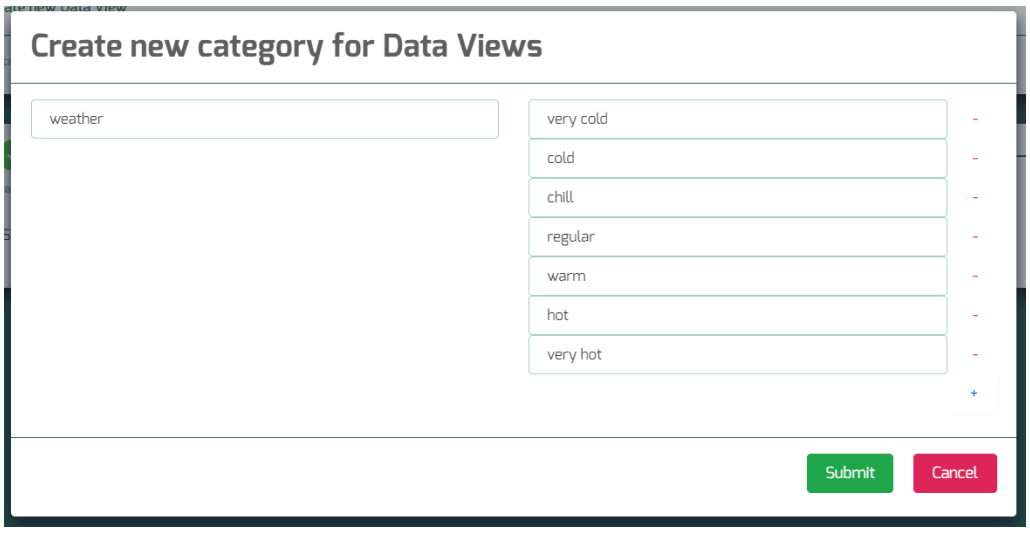


Figure 22. Create Data View – Create category

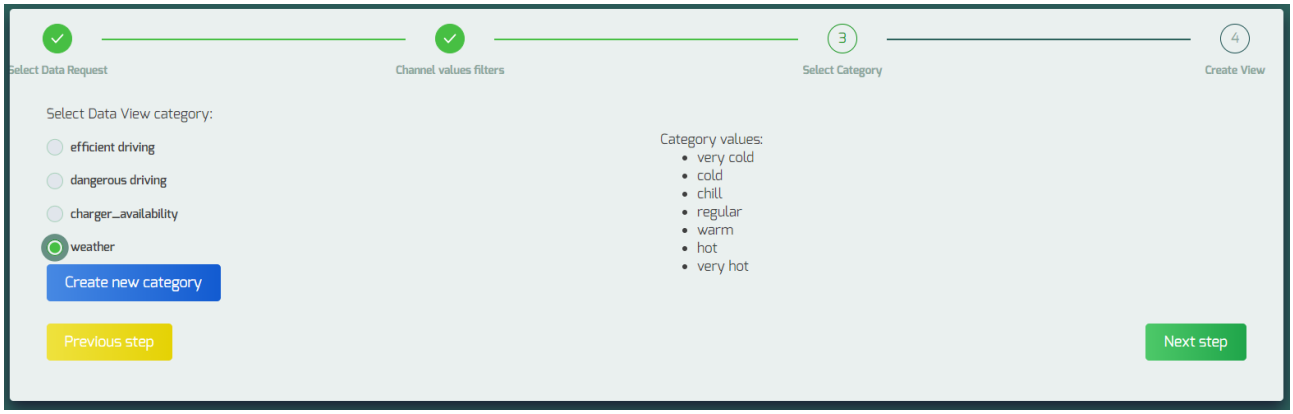


Figure 23. Create Data View – Assign category

Finally, a name and a description of the Data View are requested.

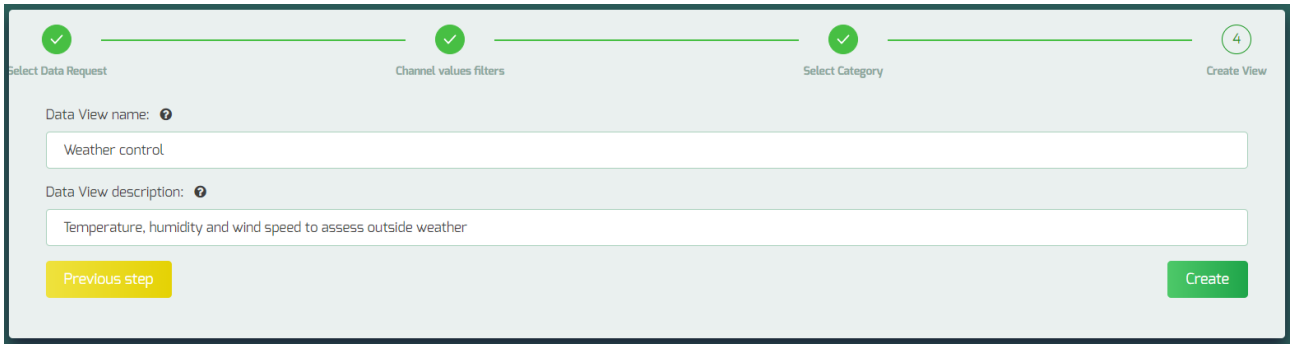


Figure 24. Create Data View - Input name and description

If successful, a view with the generated Data View name and description will be shown with a button to go to the details.

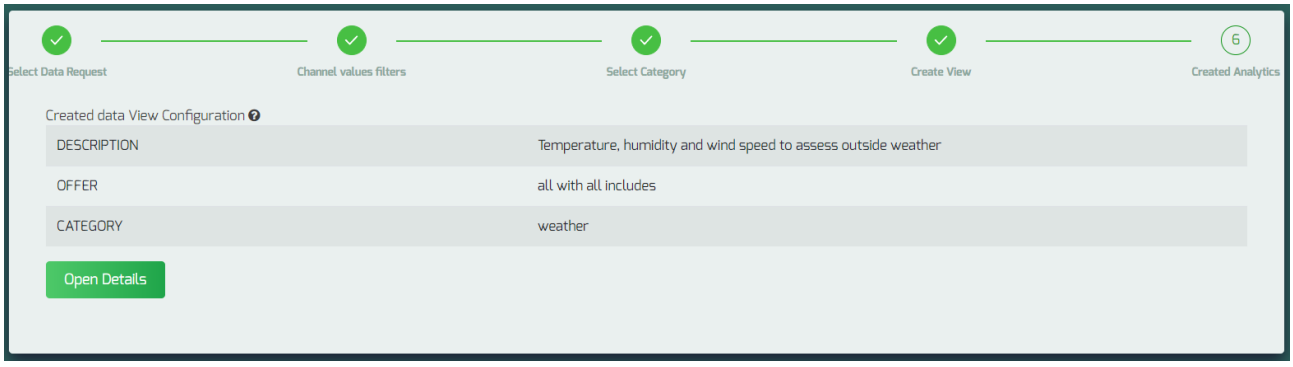


Figure 25. Create Data View - generated Data View details

4.1.4. Review Data View details

Just clicking in the Data Views menu will lead to the list view, showing a list of all Data Views created by the Service Provider.

In this list, users can filter displayed results by description, Data Request name or date.

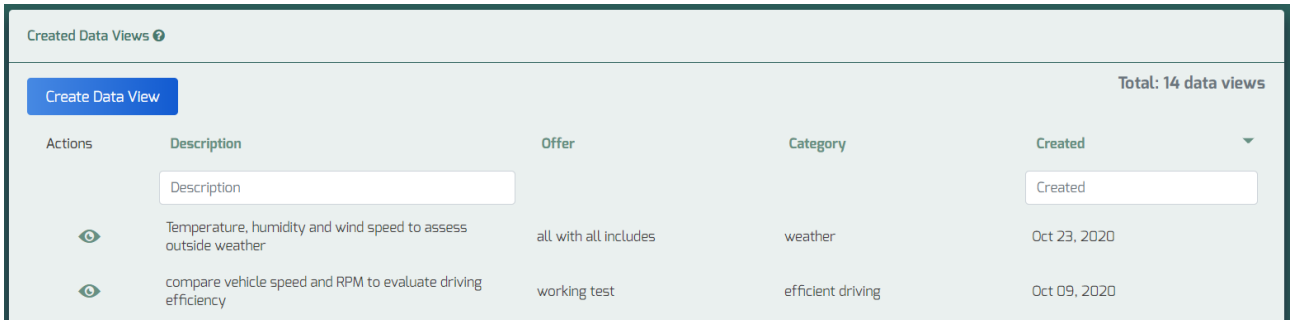


Figure 26. Created Data Views list

Clicking on any of them will open the details of that Data View, displaying the configuration, including name, description, Data Request and value filtering configuration, along with the AEON channel to subscribe to receive the data filtered.

There is also a "Delete Data View" button. By clicking and confirming this button the Service Provider will stop receiving this filtered data and the associated configuration will be deleted.

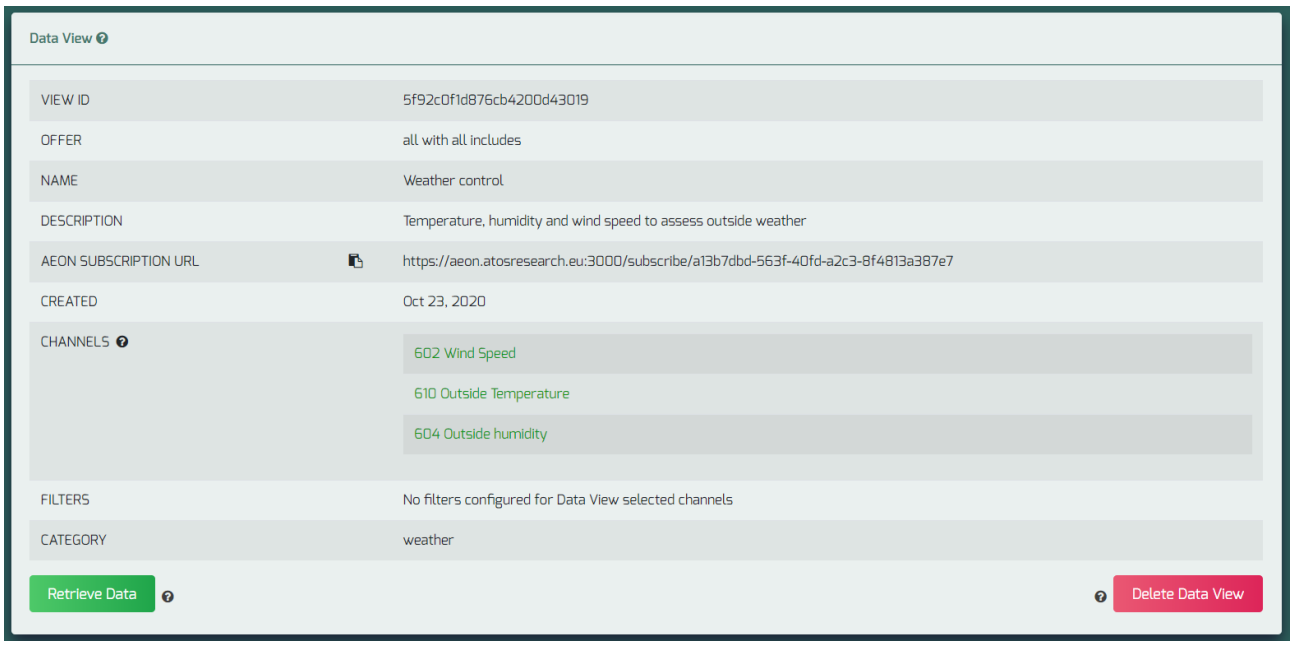


Figure 27. Data View configuration details

By clicking in the "Retrieve Data" button a table will display the latest data collected by the Data View.

Figure 28. Data View – retrieve data

4.1.5. Assign category to a Data View

To assign a category to an already created Data View just go into the Data View details page and click on the Assign Category button.

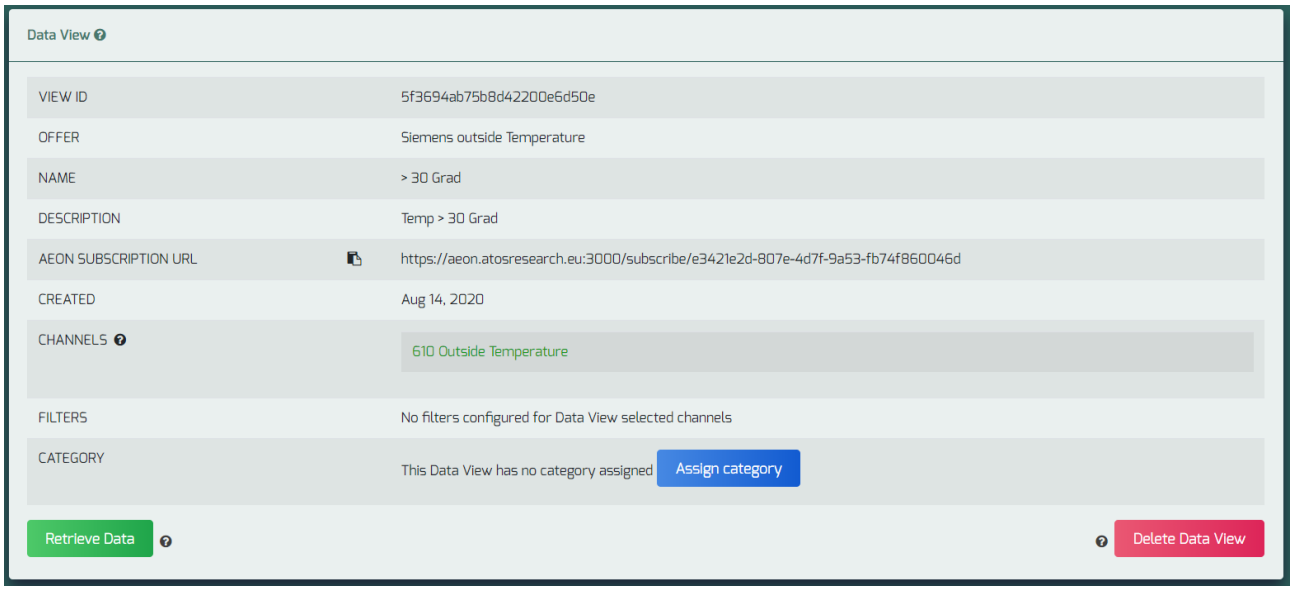


Figure 29. Data View – details view with assign button

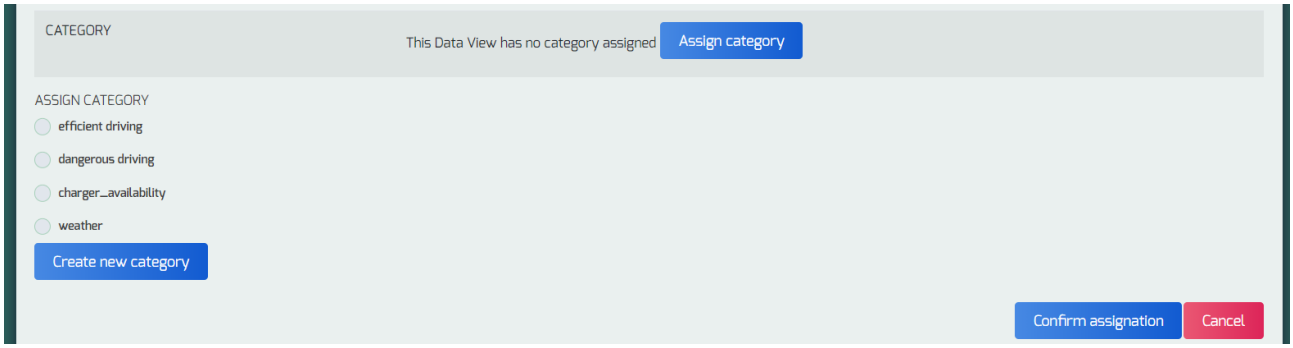


Figure 30. Data View – assign category

4.1.6. Assign category values to a Data View rows

If a Data View has a category assigned a button will appear in the Data View results table.

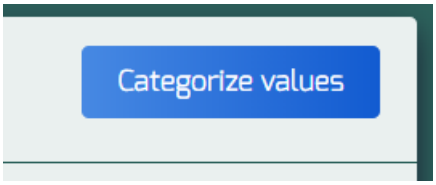


Figure 31. Data View – categorize values button

In this mode a pencil icon will appear in every row. Clicking the icon will open the value assignment popup.

Figure 32. Data View – categorize values mode

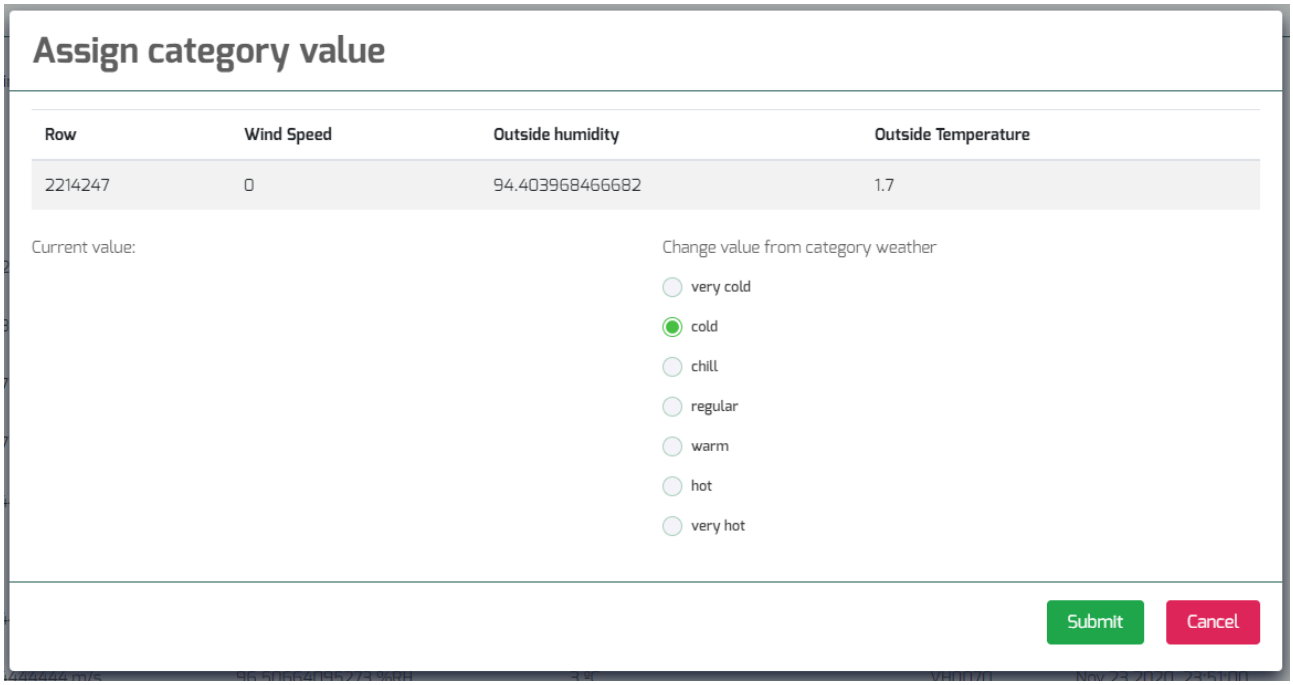


Figure 33. Data View – assign category value popup

Once all desired rows have been assigned to a value click on the "Save values" button to display a popup with a table containing the changed rows and assigned values and confirm the selection.

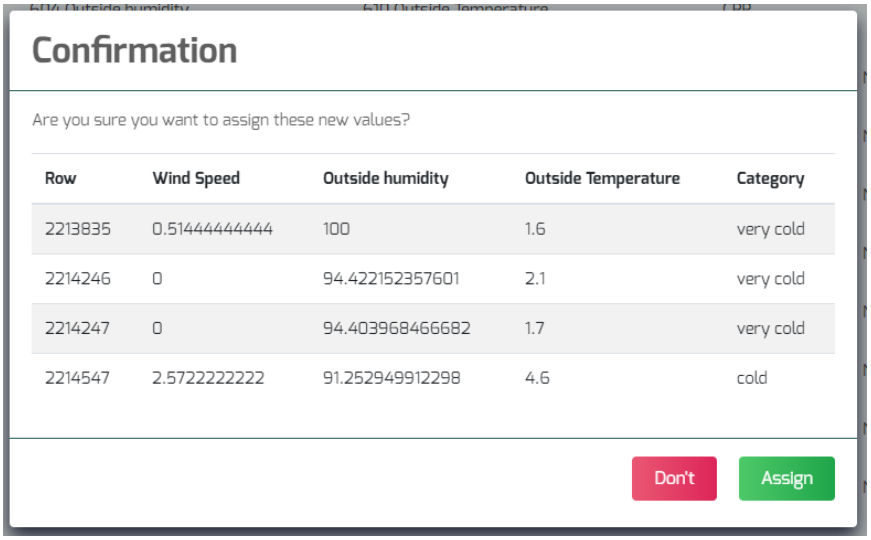


Figure 34. Data View – confirm category value assignments

4.2. Analytics

4.2.1. What is the Analytics Toolbox?

The Analytics Toolbox is a collection of tools that offers different ways of analysing the data collected from the different data requests.

Depending on the signals of the data requests some analytics may not be available.

The side menu displays the different categories of analytics. The different available tools can be found in each category.

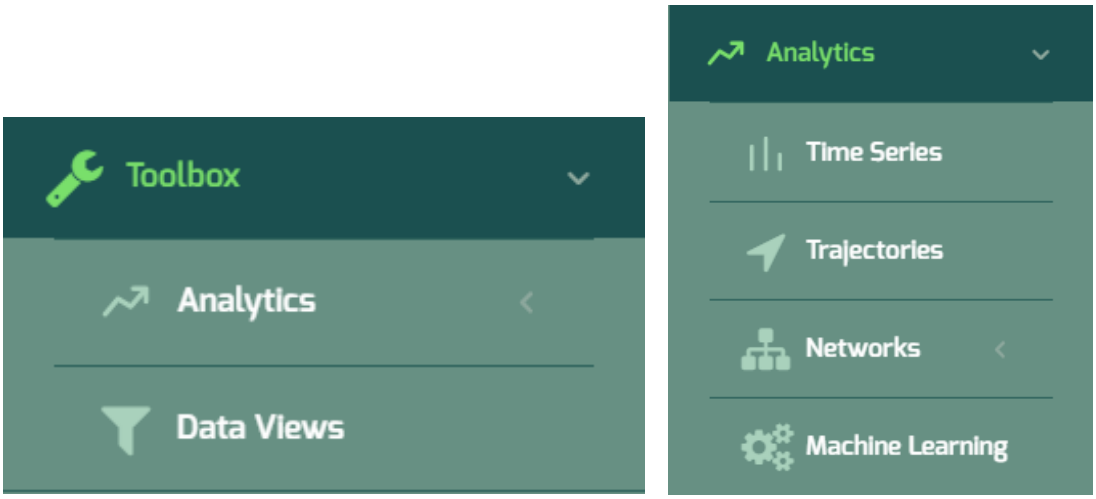


Figure 35. Toolbox menu

4.2.2. Time Series Analytics

Time series functionalities are aimed towards the management of temporal data (typically paired to timestamps that allow to trace its evolution) by enabling the inspection of its properties, as well as prediction capabilities to be able to forecast the future behaviour of a certain variable's development. In specific, the developed functions cover the following ambits:

- Drift detection
- Forecasting
- Correlation Statistics

Details about each of these subjects is presented in the following sections.

4.2.2.1. Available analytics functions services and examples

4.2.2.1.1. Drift detection

The drift present in a time series can be thought of as the changes present in the input data. In this regard, this set of analytics functions include the provision of different metrics with which to evaluate this factor, coupling it with the concept of complexity, understood as a measure of the presence of nonlinear patterns that explain the behaviour of the data.

This way, a time series is said to be highly complex when it presents underlying non-linear patterns that reveal its behaviour. On the other hand, a time series with random fluctuations in its data would have low complexity, as it would not present (as many) regular patterns in its behaviour.

With respect to the domain of this application, it provides a mean of assessing the stability with which a specific feature varies, as well as a preliminary step into further assessing its behaviour over time.

4.2.2.1.1.1. Sample Entropy

Sample Entropy is a probabilistic measure of a time series' randomness. It relies in two main parameters, an embedding dimension **e** and a tolerance value **t**, for computing the probability that, given two sets of **e** simultaneous data points that cover a distance lower than **t**, two sets of **(e+1)** simultaneous data points would also cover that distance lower than **t**.

Consequently, the parameters to be adjusted for the functioning of this process are:

- An **embedding dimension** value representing the number of points to consider in each set.
- A **tolerance** value as the estimated distance for each set of simultaneous data.
- Whether or not to **normalize** the data of the time series' values (which, depending on the nature of the data, may be necessary for obtaining a valid output from this measure).

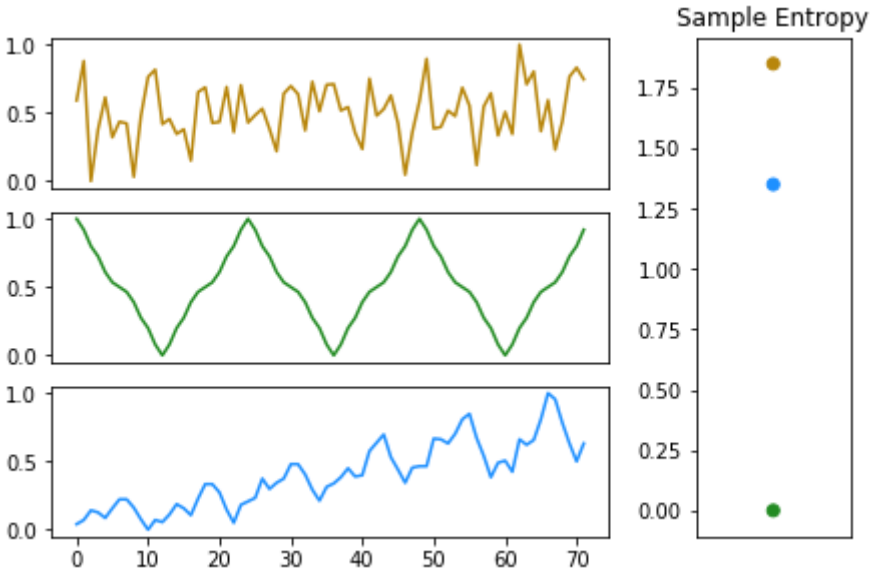


Figure 36. Expected behaviour of the Sample Entropy function given different types of input

4.2.2.1.1.2. Permutation Entropy

Permutation Entropy measures the variability of a time series behaviour. Given the ordinal pattern of sequential sets of **e** points (each spaced by **d** points from the other) in a time series, this metric is computed by assessing the behaviour of the series through all possible changes in each set's ordinal pattern.

The parameters associated to the functioning of this process are:

- An **embedding dimension** value representing the number of points to consider in each set.
- A **delay** value referred to the number of spacing points between sets

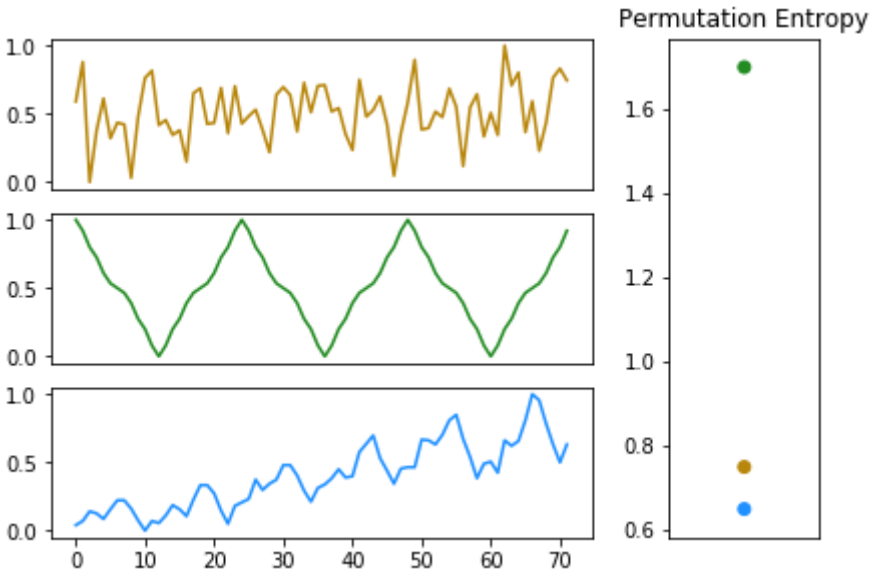


Figure 37. Expected behaviour of the Permutation Entropy function given different types of input

4.2.2.1.1.3. Irreversibility

The Irreversibility measure quantifies the variability in the behaviour of a time series when its sequence of timestamps is reversed. It is computed over sets of **e** data points given by an embedding dimension. Thus, the unique parameter on which adjustment relies the functioning of this process is:

- An **embedding dimension** value representing the number of points to consider in each set.

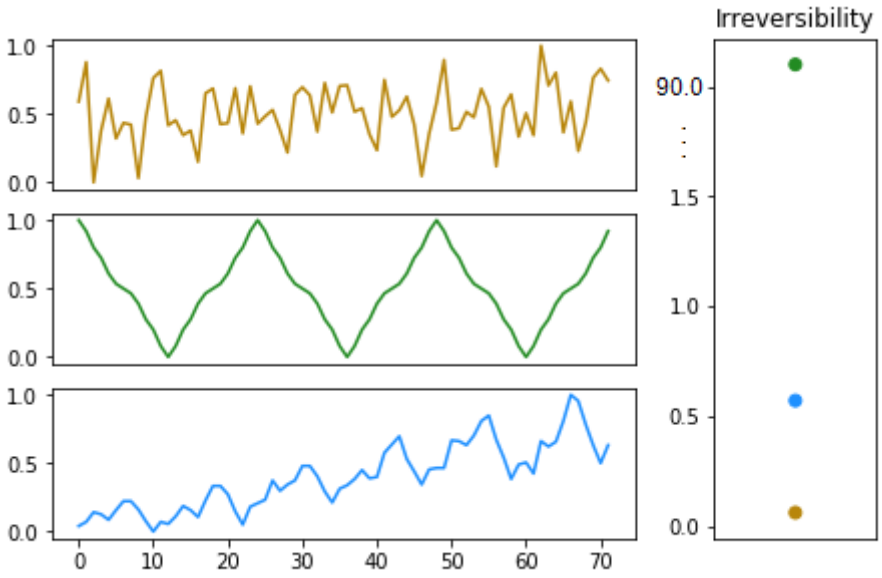


Figure 38. Expected behaviour of the Irreversibility function given different types of input

4.2.2.1.2. Value Prediction

Among the capabilities developed for the management of time series, there are functions aimed towards the prediction of their future values. These include data processing procedures with the following algorithms:

- Regression Trees
- Neural Networks
- ARIMA models

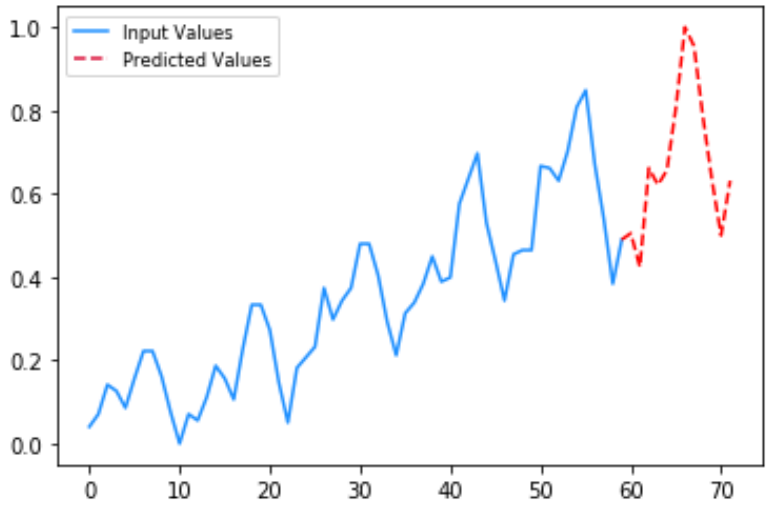


Figure 39. Prediction of a time series unknown data points

4.2.2.1.2.1. Regression Tree

The Regression Tree algorithm's mechanism emulates a flowchart where each node describes a test on the values from the assessed variable. This tests control whether or not a specific value from the series is greater or lower than a certain threshold computed by the algorithm. Coherently, each of the two branches that derive from each node represent a result of the test (lower or greater than the threshold).

New values are predicted based on the set of rules that emanate from the final tree's tests.

The parameters to be adjusted for the functioning of this process are:

- The number of data points to **forecast**.
- The index of the time series representing the point at which the **division** between training and test sets is to be performed.

4.2.2.1.2.2. Neural Network

A Multilayer Perceptron is trained and used for the prediction of new data points for a given time series.

The parameters to be adjusted for the functioning of this process are:

- The number of data points to **forecast**.
- The index of the time series representing the point at which the **division** between training and test sets is to be performed.

4.2.2.1.2.3. ARIMA

ARIMA is a time series forecasting algorithm based on the following principles:

- Autoregression (AR): uses the dependent relationship between an observation and some number of lagged observations.
  ARIMA assumes that the input data verifies that the values correspondent to a specific time depend of the values of previous data points.
- Integrated (I): differencing of observations to make the series stationary (same probability distribution for all data points).
- Moving Average (MA): uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

In coherence, the adjustable parameters associated to its functioning are

- The number of data points to **forecast**.
- A numeric value (**alpha**) for the statistical significance of the hypothesis test conducted within the model (0.05 is recommended as a widespread practice).
- The **order** of each ARIMA component. Namely, the components to adjust are:
  - **p**: related to AR's regular part. Number of lag observations included in the model.

- **d**: related to I's regular part. Number of times that the raw observations are differenced.
- **q**: related to Q's regular part. Size of the moving average window.
- **P**: related to AR's stationary part. Number of lag observations included in the model.
- **D**: related to I's stationary part. Number of times that the raw observations are differenced.
- **Q**: related to Q's stationary part. Size of the moving average window.

4.2.2.1.3. Correlation Statistics

Correlation coefficients are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

The developed features include the calculation of such coefficients based on Pearson's and Spearman's correlation.

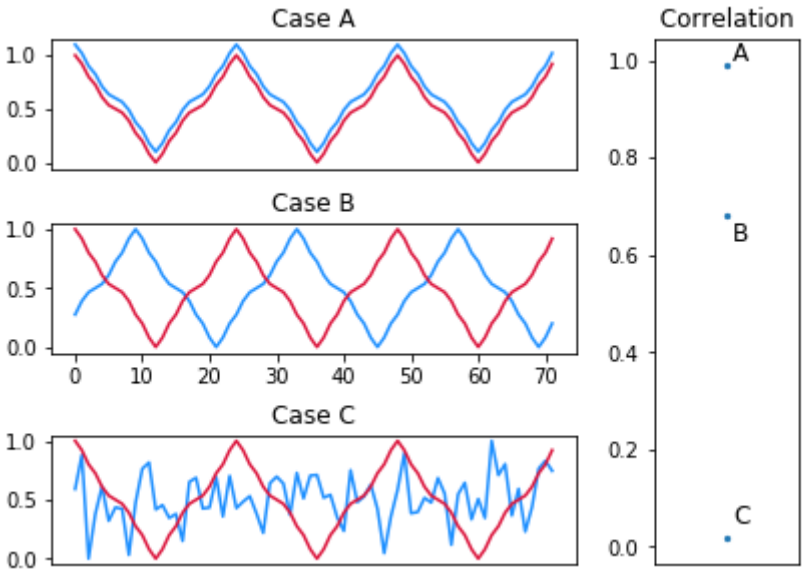


Figure 40. General expected behaviour of a correlation function

4.2.2.1.3.1. Pearson

Pearson correlation quantifies the strength of a linear relationship between paired data (the pair of time series which correlation is being computed).

It relies on assumptions of bot normality of distribution and homoscedasticity equal variance across variables) of the variables (time series) being evaluated.

4.2.2.1.3.2. Spearman

Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data (the pair of time series which correlation is being computed).

Since it is a rank correlation measure (based on ordinality of samples), it does not rely on the assumptions of normality and homoscedasticity that Pearson's correlation does.

4.2.2.2. How to create a Time Series Analytics

The creation form can be found through de button in the Time Series list view:

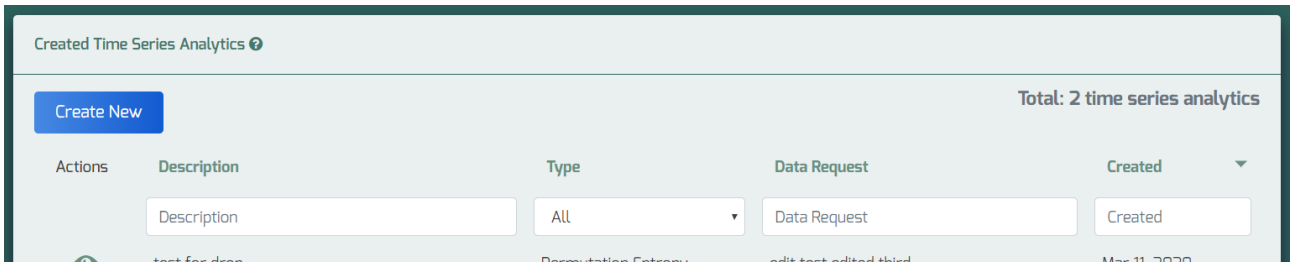


Figure 41. Create new Time Series Analytics button

Once inside a window with the form resume is shown in the top part of the view:

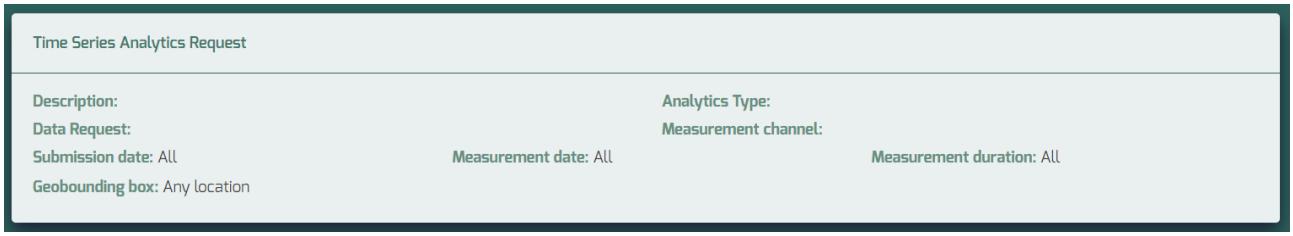


Figure 42. Time Series form resume

The first step here is to select the type of analysis to be performed. The options are displayed in the button list, with a brief explanation of the analysis selected shown at the right.

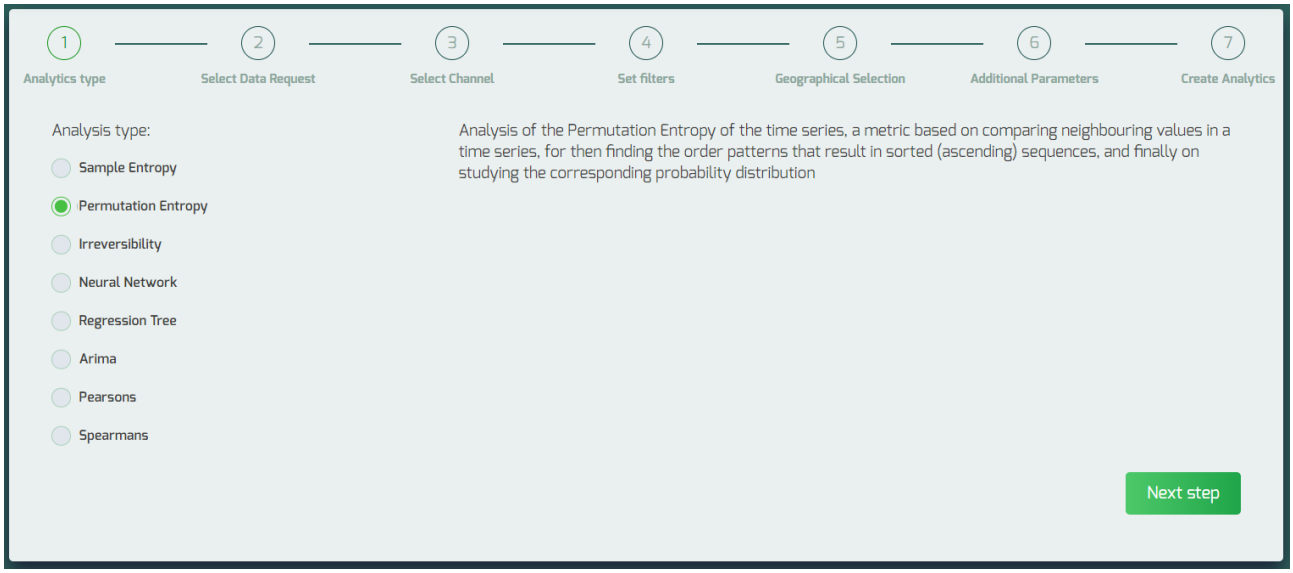


Figure 43. Create Time Series analytics - select type

In order to perform analytics on a set of data, the Service Provider needs an active data request that has been approved for analytics by the CPP owner.

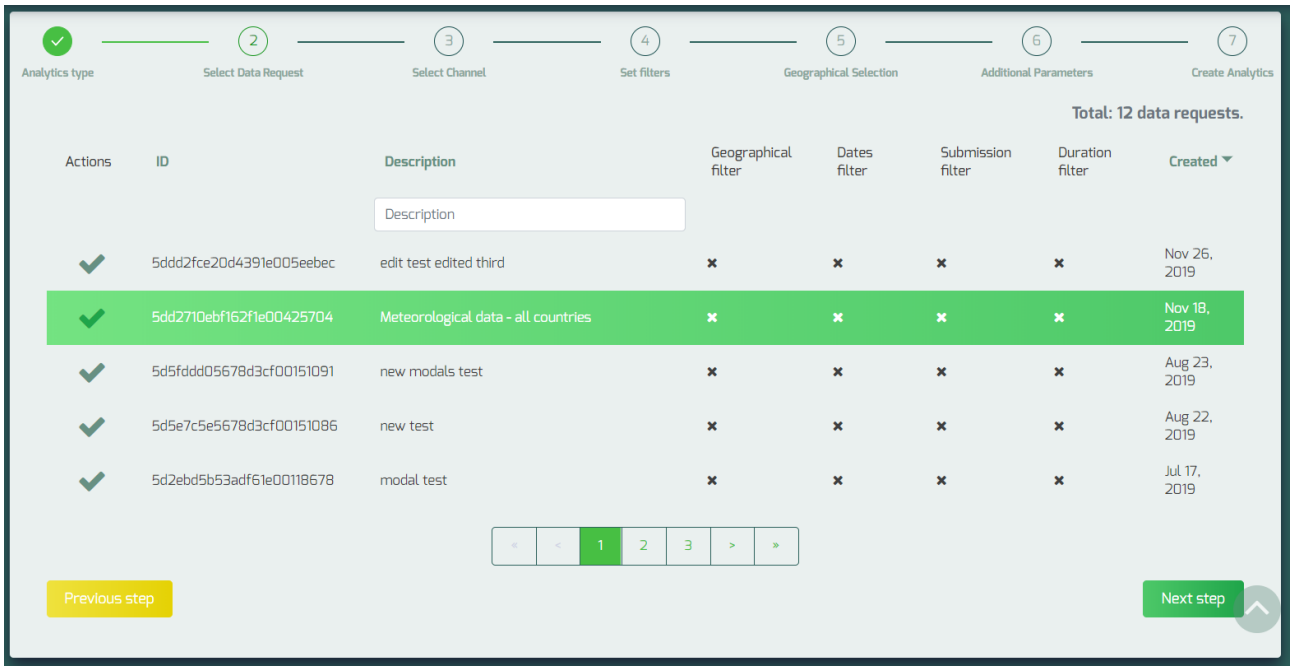


Figure 44. Create Time Series analytics - select Data Request

Every Data Request comprises one or more measurement channels from which the request receives data. Next step is to decide from which of those measurement channels data analysis is desired.

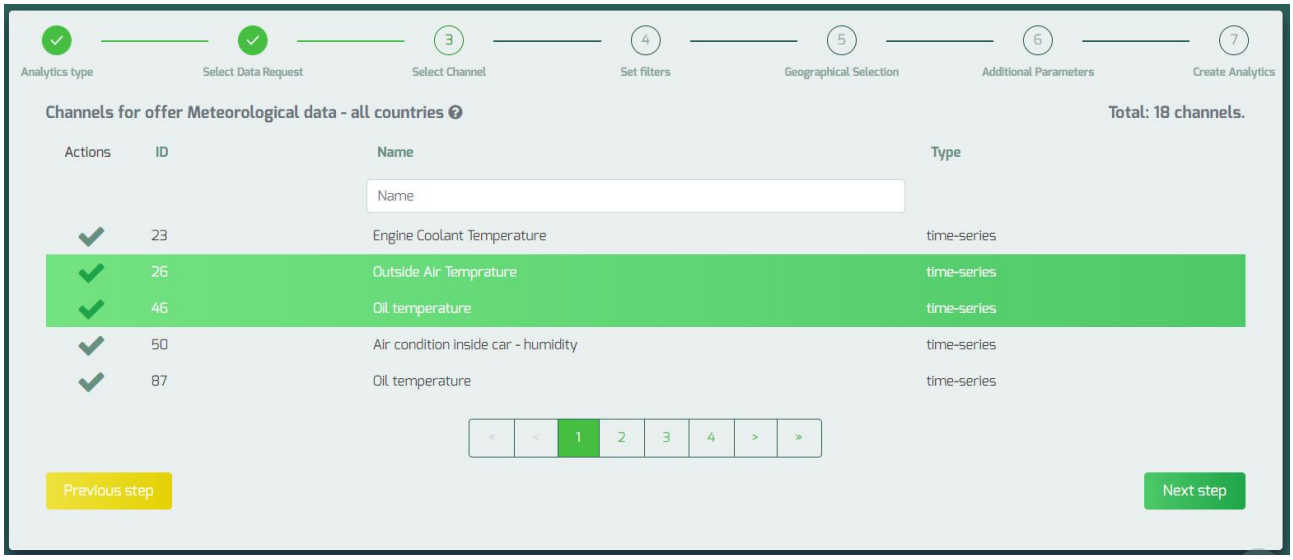


Figure 45. Create Time Series analytics - select measurement channels

From the selected Data Request initial filters (if any) data can be further filtered:

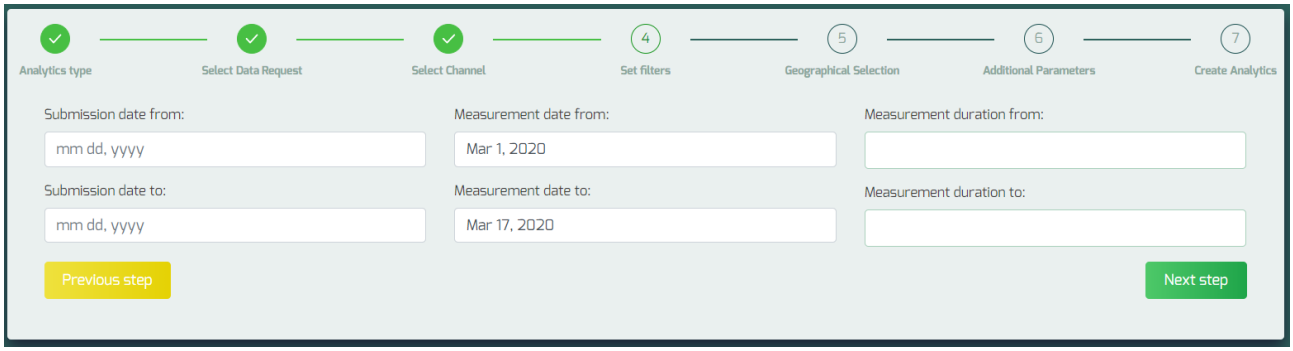


Figure 46. Create Time Series analytics - select filters

From the selected Data Request initial geographical bounding (if any) data can be further filtered:

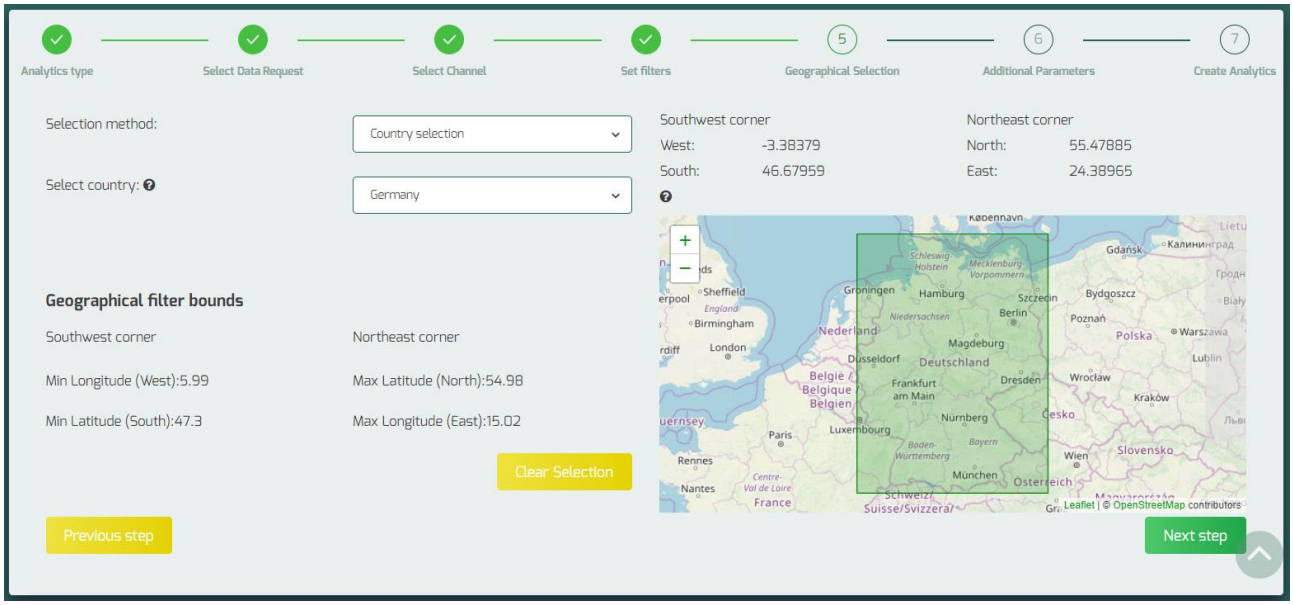


Figure 47. Create Time Series analytics - select geographical filter

Depending on the Time Series type additional parameters can be requested. A brief explanation of each parameter can be found in the question marks next to each of them:

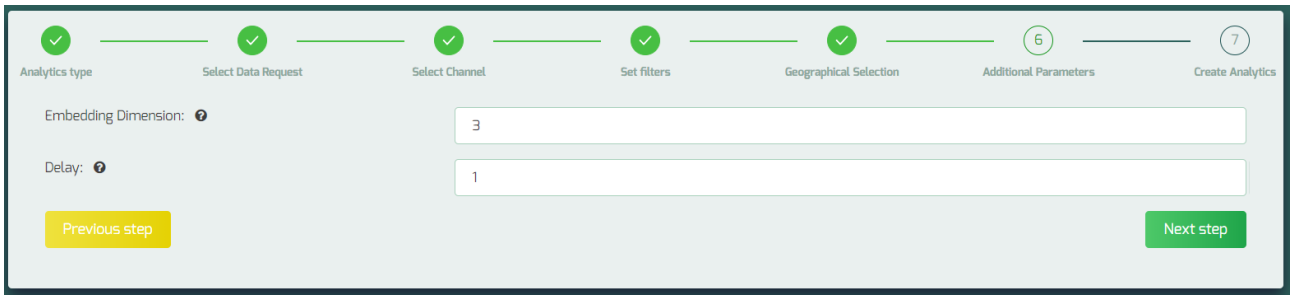


Figure 48. Create Time Series analytics - additional parameters

Finally, a description for the analytics is requested so it can be found easily by the user:

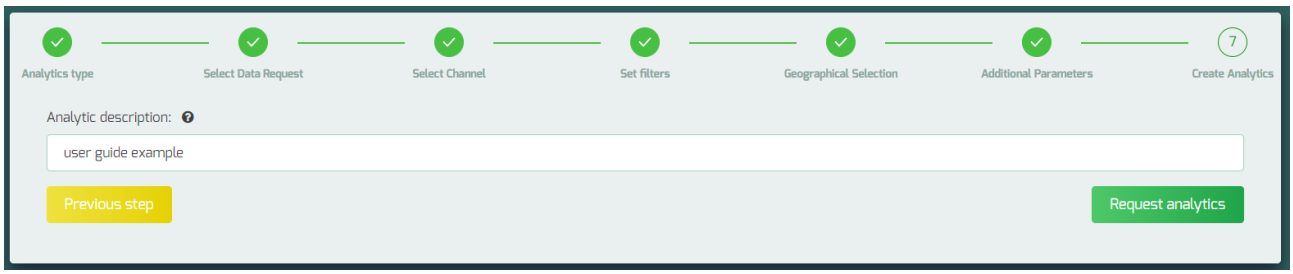


Figure 49. Create Time Series analytics - insert description

If successful, a summary of the created analytics will be displayed:

Figure 50. Create Time Series analytics - final resume

4.2.2.3. Review and stop a Time Series Analytics

Just clicking in the Time Series menu will lead to the list view, showing a list of all Time Series analytics created by the Service Provider.

In this list the user can filter the displayed results by description, analytics type, Data Request name or date.

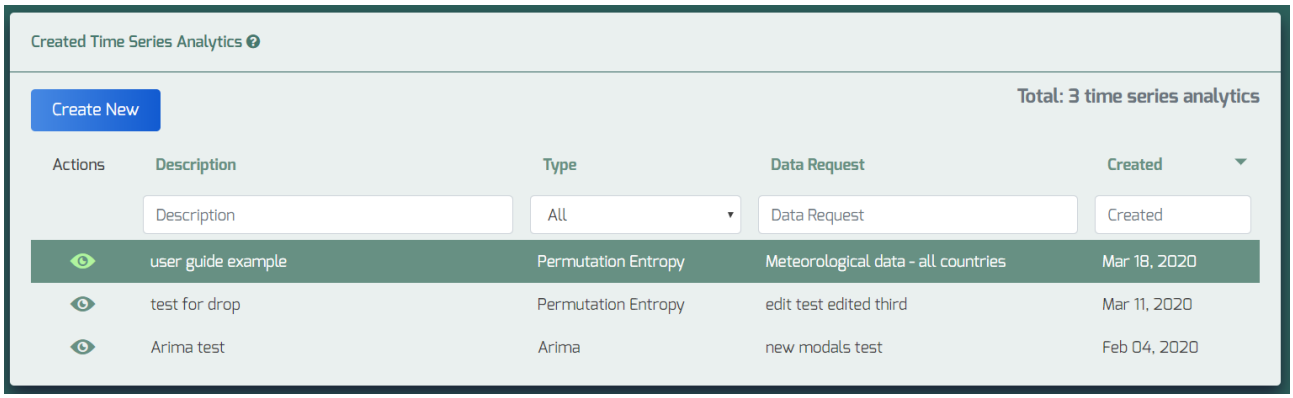


Figure 51. Created Time Series analytics list

Clicking on any of them will open the details view of that analytics, displaying the detailed configuration, including description, type and the AEON channel to subscribe to receive the analysed data.

There is also a "Delete Analytics" button. By clicking and confirming this button the Service Provider will stop receiving this analytics data and the configuration will be deleted..

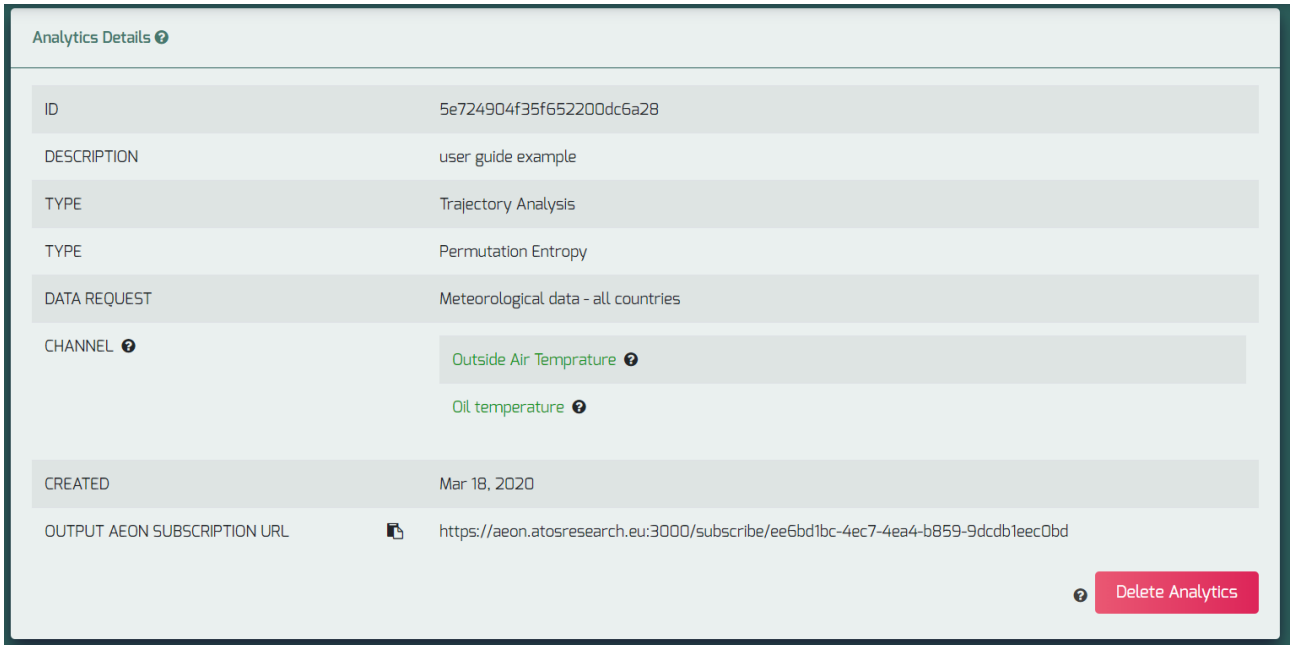


Figure 52. Time Series analytics configuration details

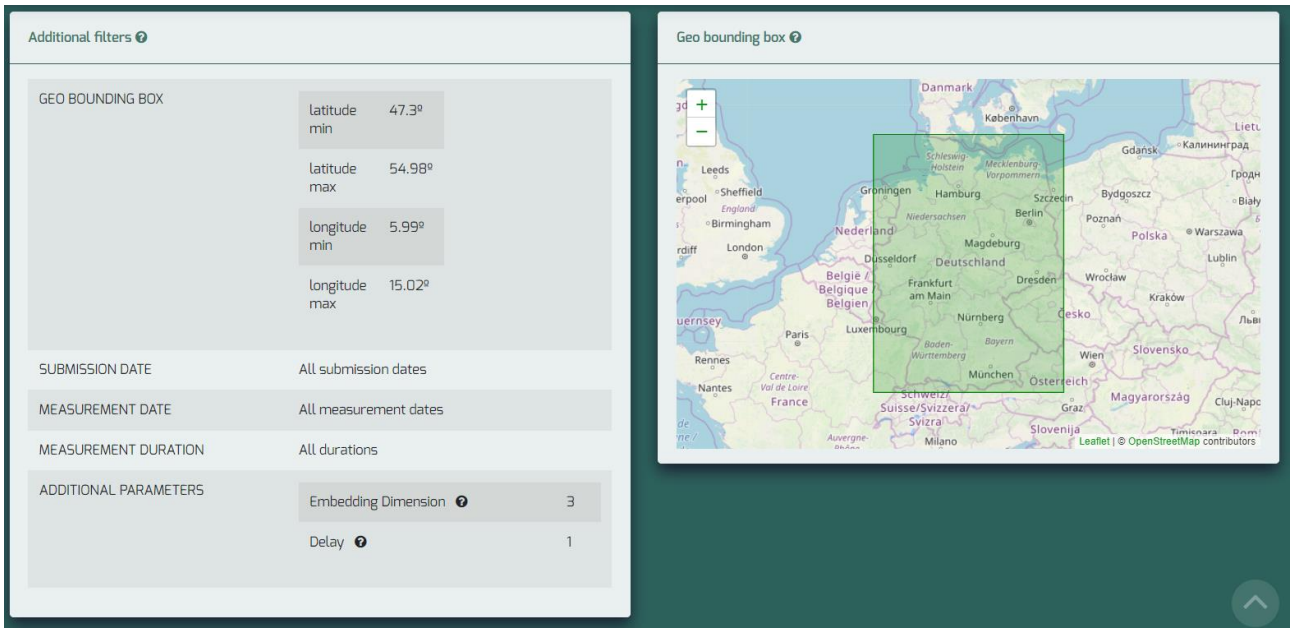


Figure 53. Time Series analytics filter details

4.2.3. Trajectory Analysis

Trajectory Analysis, understood as the management and treatment of GPS data (X-Y coordinates in which respects to this project), implies the necessity to operate with trajectories in different ways that enable the understanding of their main characteristics. With this purpose, the set of functionalities developed for the fulfilment of such requirements allow the user to evaluate the behaviour and properties of trajectories in different ways, including:

- Computation of simple statistics associated to a trajectory.
- Clustering of similar trajectories.
- Linear interpolation of a trajectory's GPS coordinates.
- Detection of anomalous coordinates within a trajectory.

These services are described in detail in the subsequent sections of this document.

4.2.3.1. Available analytics functions services and examples

4.2.3.1.1. Statistics

The statistics functionality provides a series of summarizing properties for a given trajectory, namely the total distance traveled in meters, its duration in seconds and its average velocity in meters per second.

The user needs to provide the trajectories for which to obtain these metrics, with no additional adjustable fields.

As noted in the previous description, there is only one requirement for this process' functioning:

- The coordinates of the trajectory to be resampled as pairs of (x, y) coordinates. Each of these points must have a timestamp associated indicating the date and time of occurrence.

4.2.3.1.2. Clustering

The clustering functionality enables the grouping of similar trajectories. Neighboring trajectories are labeled as belonging to the same cluster based on a metric of pair-wise distance between coordinates.

Since the adjustment of the parameters intrinsic to the algorithm is automatic, the user only needs to provide the trajectories to be grouped, with no additional adjustable fields.

In the example below (Figure 54) the left graphic represents the original scenario given as input to the algorithm, with the correspondent cluster centroids identified in the right part of the image.

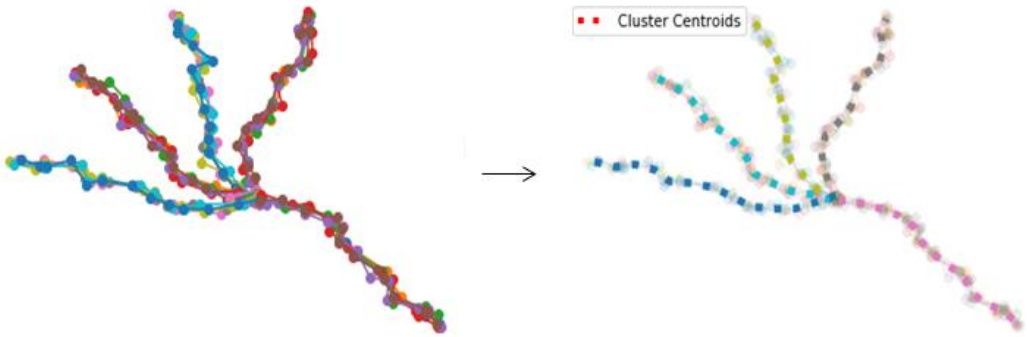


Figure 54. Example of a clustered trajectories.

With respect to this process' input, and as noted in the previous description, there is only one requirement for its functioning:

- The coordinates of the trajectories to be grouped as pairs of (x, y) coordinates. Each of these points must have a timestamp associated indicating the date and time of occurrence.

4.2.3.1.3. Interpolation

The interpolation functionality provides a mean of resampling a given trajectory. The type of interpolation performed is linear.

By providing a desired time resolution (in seconds) the user is able to set the frequency of appearance of each point in the new resampled trajectory.

In the example below (Figure 55) the solid blue circles represent the original coordinates, while the orange crosses are the resampled trajectory points (for a given time resolution).

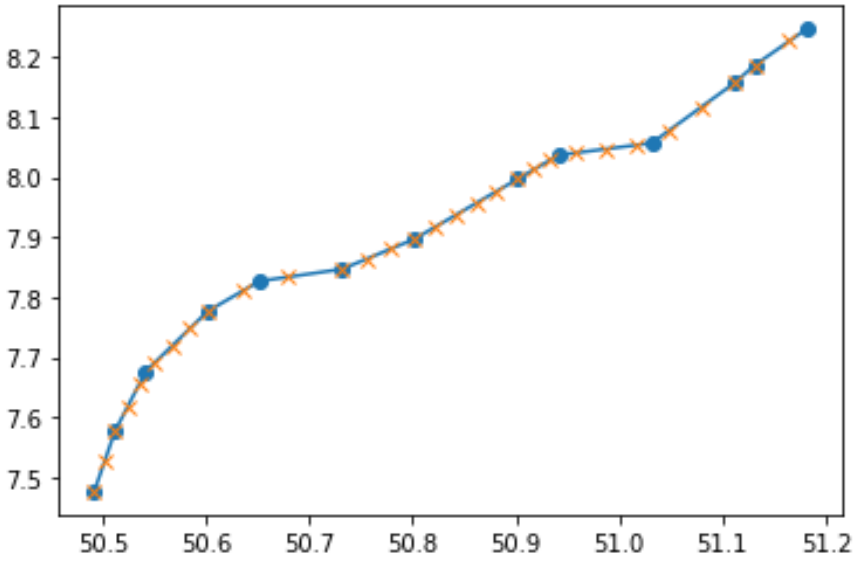


Figure 55. Example of a resampled trajectory.

The parameters required for the functioning of this process are:

- The coordinates of the trajectory to be resampled as pairs of (x, y) coordinates. Each of these points must have a timestamp associated indicating the date and time of occurrence.
- The time resolution (in seconds) to which the trajectory is intended to be resampled as a numeric value. This value should fulfil the following conditions:
    - Be an integer.
    - Be greater than 0.
    - Be smaller than the time difference between the initial and last coordinates' timestamps.

4.2.3.1.4. Anomaly detection

The anomaly detection functionality allows to identify abnormal points in a given trajectory. After associating the input trajectory to a cluster from the system's stored information, it compares the pair-wise distance between each of its coordinates and the group-centroid's. If this distance surpasses a certain threshold, the associated coordinates are marked as outlying.

The user has the option to change the type of threshold to apply in this process between a standard deviation protocol and an interquartile range one.

In the example below (Figure 56) the solid orange line and points represent the trajectory provided as input, with the outliers marked with black crosses.

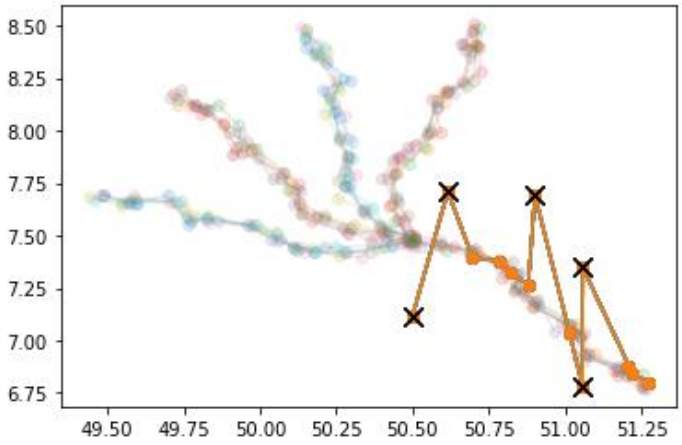


Figure 56. Example of positional anomalies' identification.

The parameters required for the functioning of this process are:

- The coordinates of the trajectory to be evaluated as pairs of (x, y) coordinates. Each of these points must have a timestamp associated indicating the date and time of occurrence.
- Optionally, the threshold method to apply for the identification of outliers. If the user does not provide this parameter, then the standard deviation protocol is applied by default. The available threshold metrics are standard deviation and interquartile range.

4.2.3.2. How to create a Trajectory Analysis

In order to perform analytics on a set of data, the Service Provider needs an active data request that has been approved for analytics by the CPP owner. Notice that only Data Requests containing measurement channel Position will be eligible, as that channel is the only one that includes the necessary data.

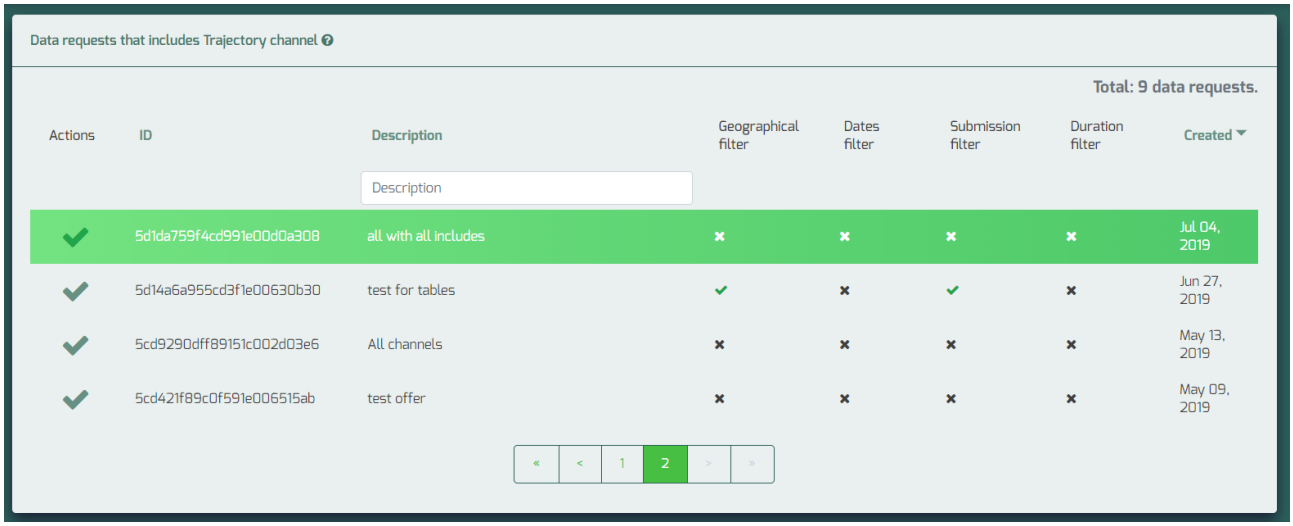


Figure 57. Create Trajectory Analysis - select Data Request

From the selected Data Request initial filters and geographical bounding (if any) data can be further filtered:

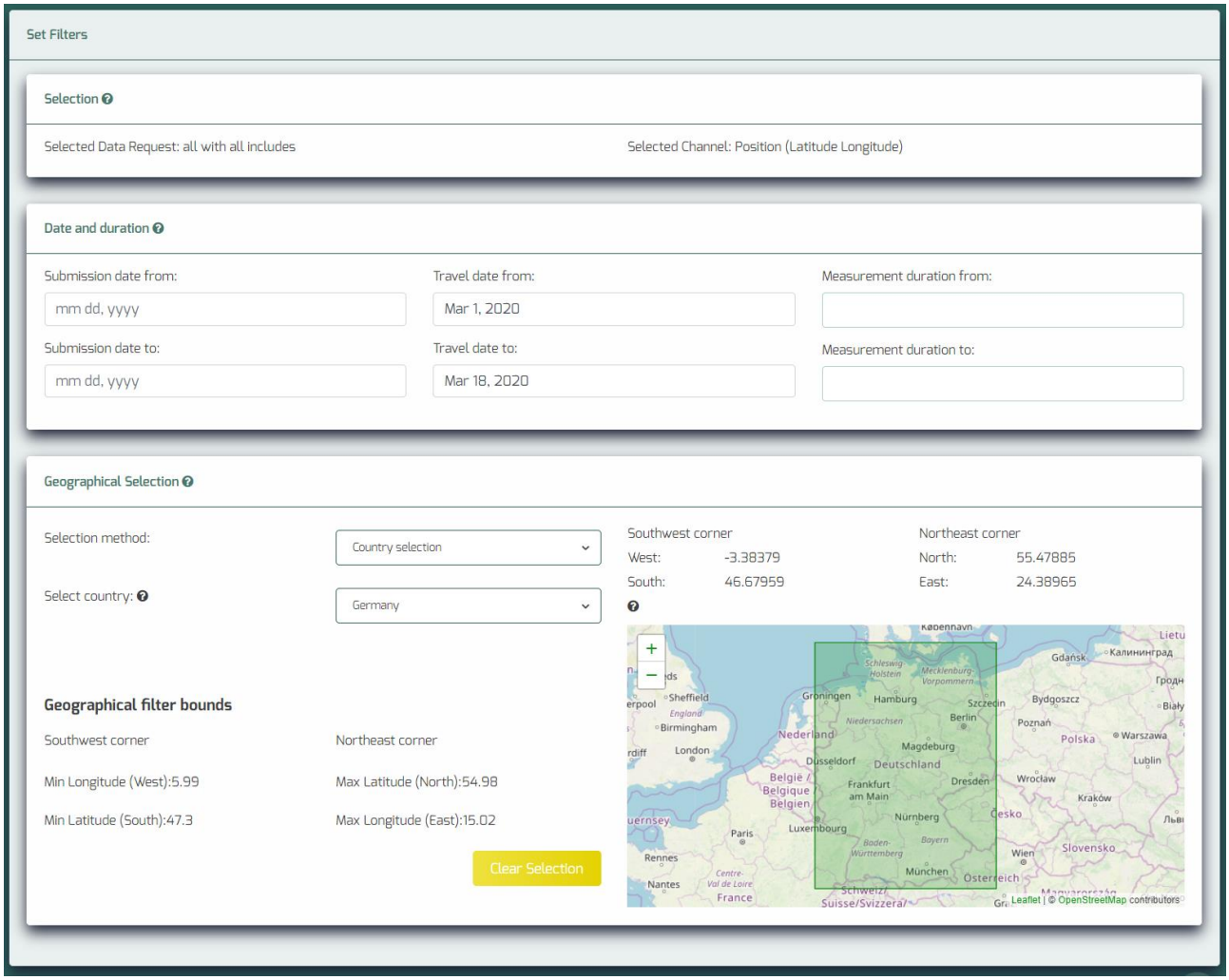


Figure 58. Create Trajectory Analysis - select filters

Next user has to select which type of analysis is desired to be performed in the trajectories contained in the data selected.

Depending on the analysis type additional parameters can be requested. A brief explanation of each parameter can be found in the question marks next to each of them:

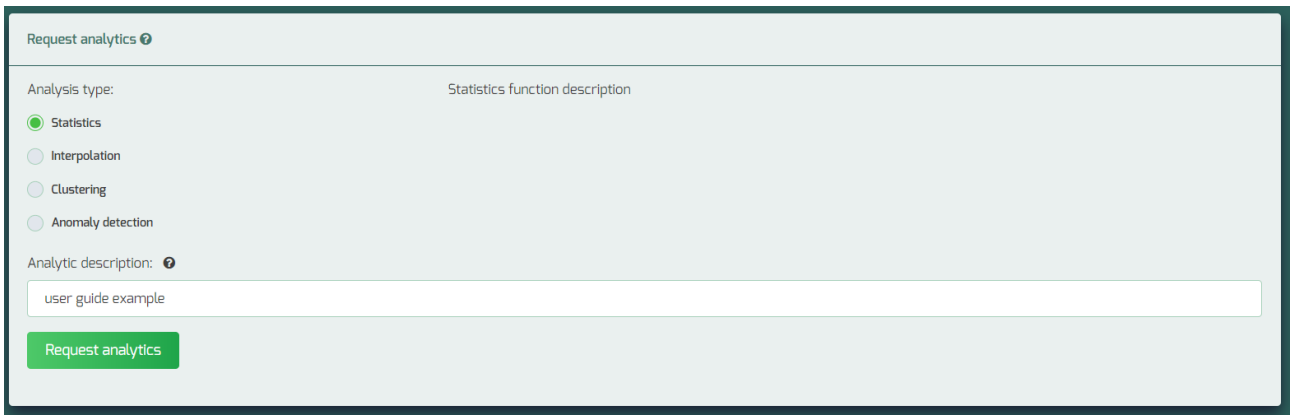


Figure 59. Create Trajectory Analysis - additional parameters

Finally, after labelling and describing the analysis, a message will be displayed leading to the trajectories list if the analysis was successful:

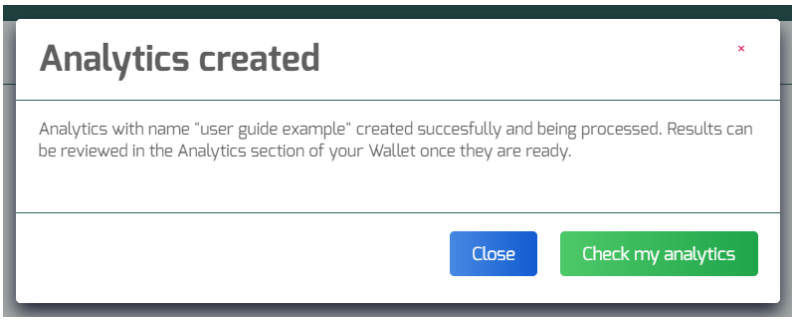


Figure 60. Create Trajectory Analysis - insert description

4.2.3.3. Review Trajectory Analysis results

Just clicking on the Trajectories menu will lead to the list view, showing a list of all Trajectory Analysis created by the Service Provider.

In this list users can filter the displayed results by description, analytics type, Data Request name or date.

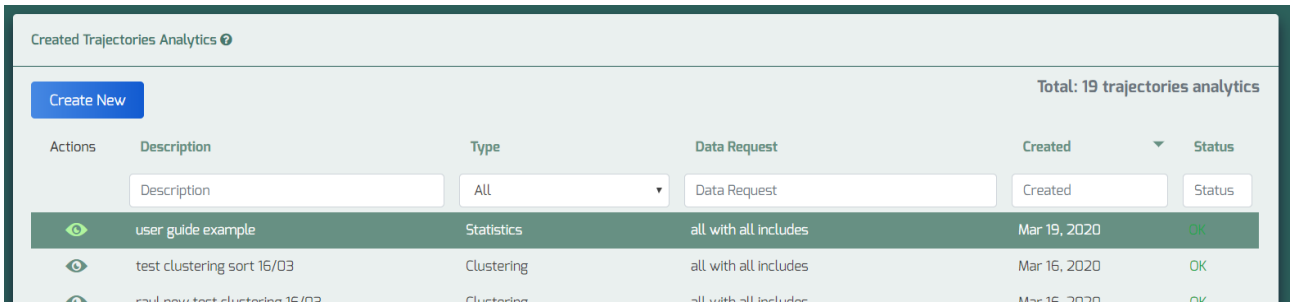


Figure 61. Created Trajectory Analysis list

Clicking on any of them will open the details view of that analytics, showing us the detailed configuration, including description, type and the results of the analysed data.

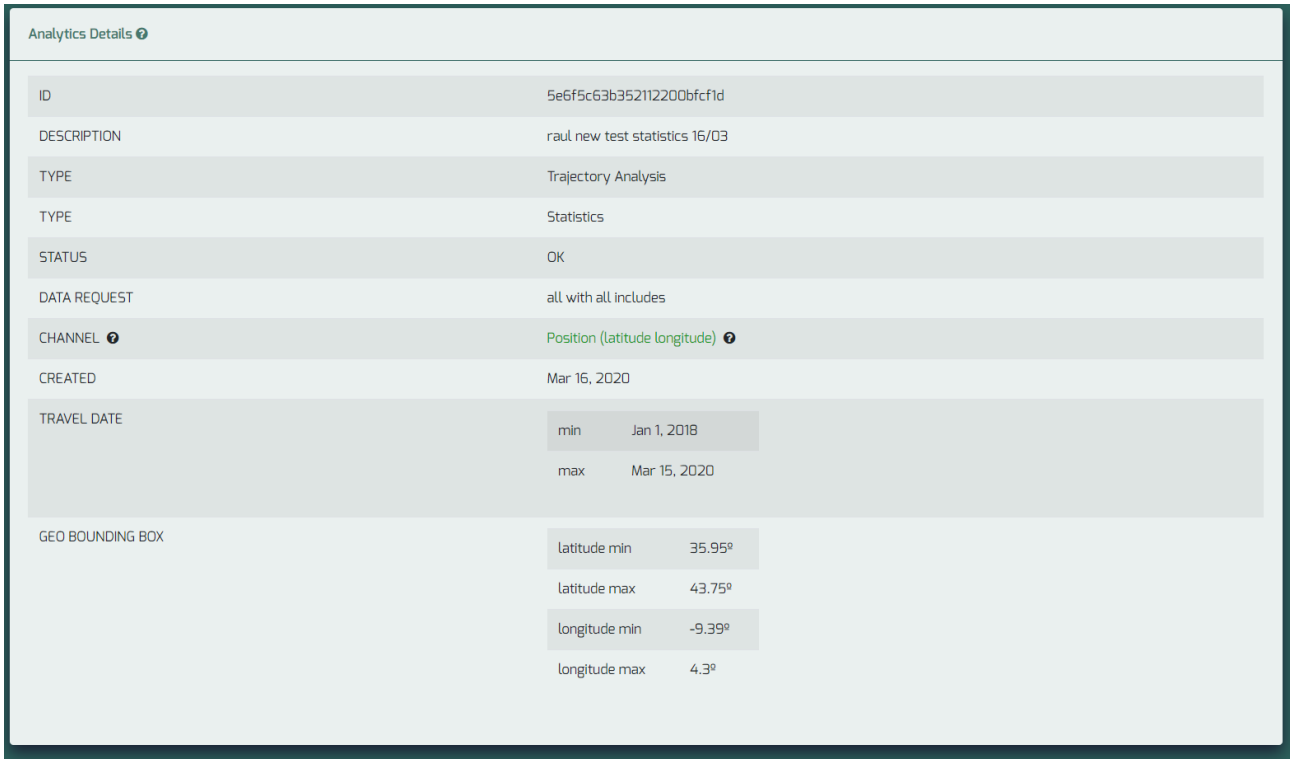


Figure 62. Trajectory Analysis configuration details

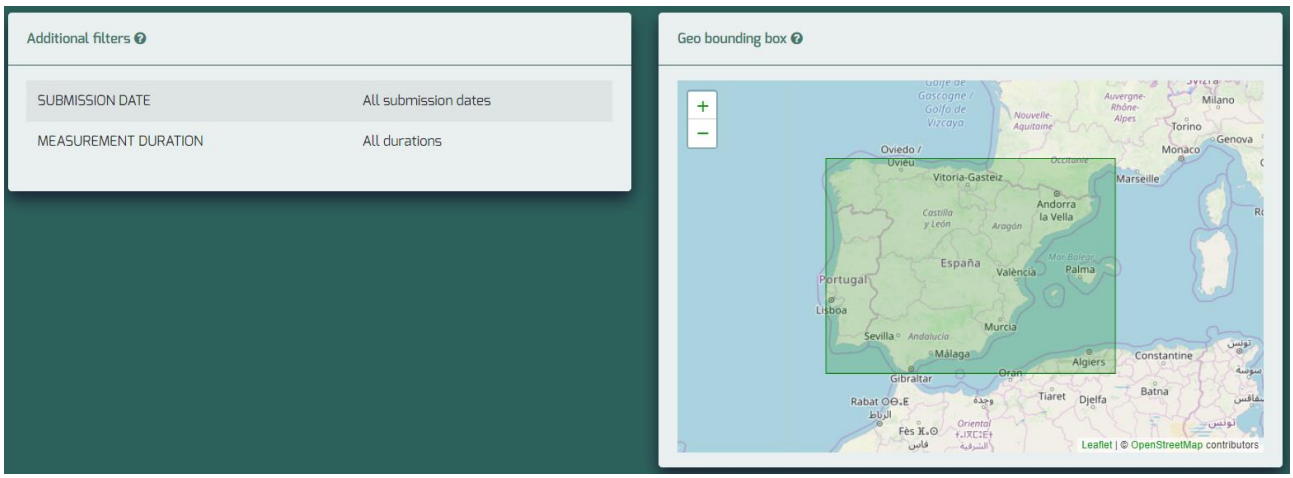


Figure 63. Trajectory Analysis filter details

Each type of Trajectory Analysis has its own kind of results and way to be displayed.

4.2.3.3.1. Statistics

First a table with all analysed trajectories is shown (Figure 64). Selecting one will display another box containing the trajectory statistics and map representation (Figure 65).

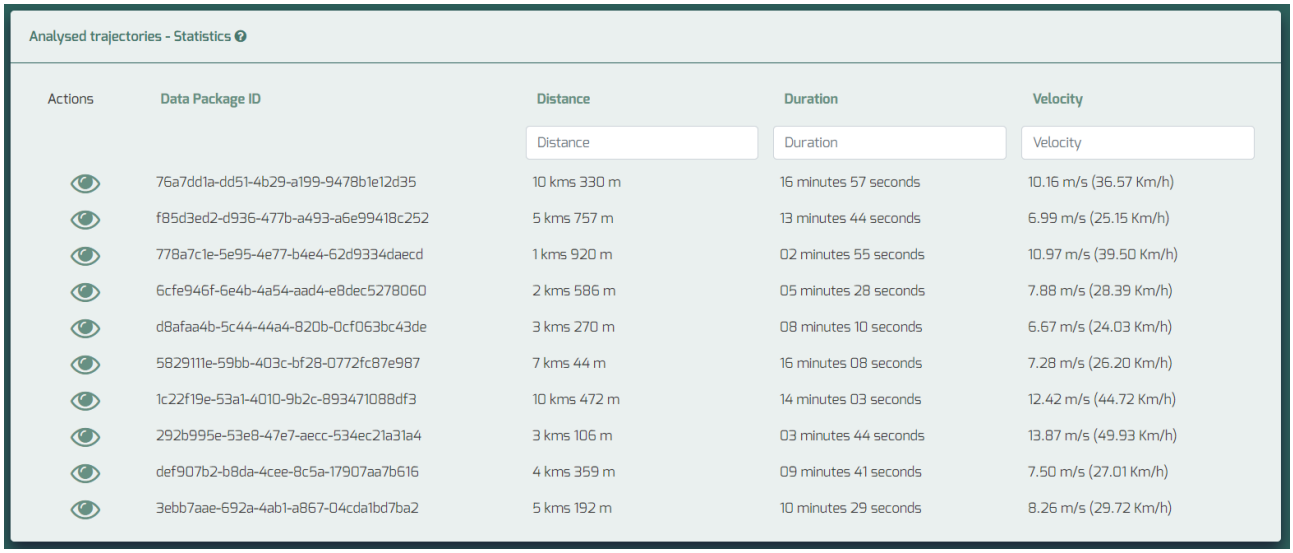


Figure 64. Trajectory Analysis - statistics results list

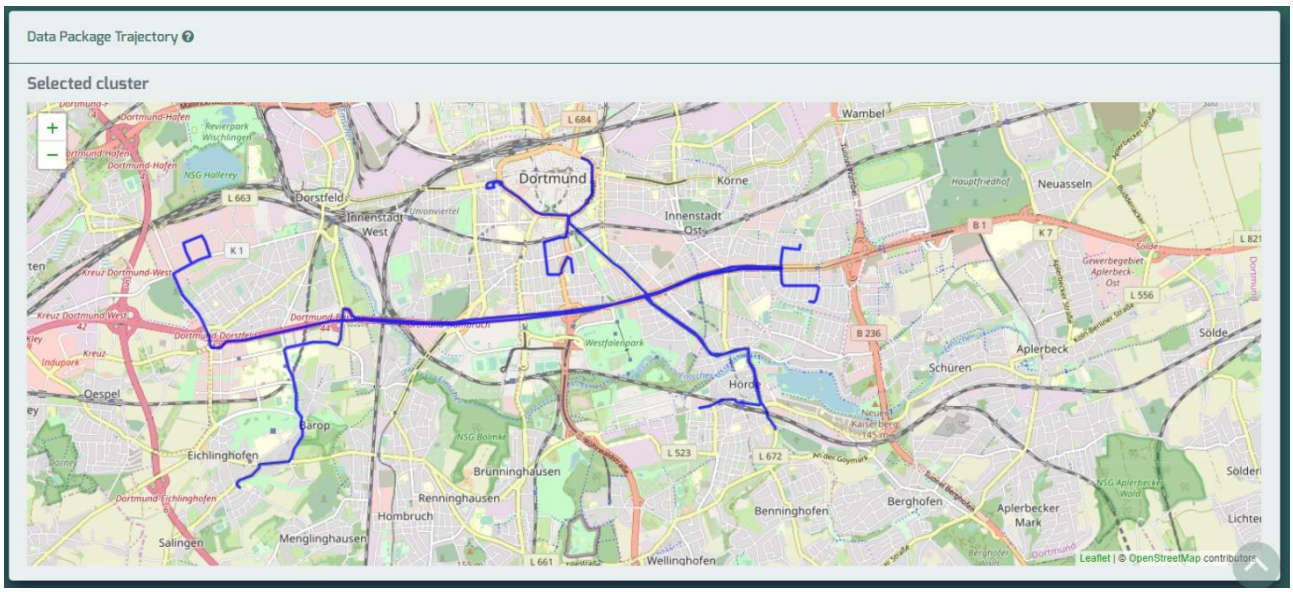


Figure 65. Trajectory Analysis - statistics results display

4.2.3.3.2. Clustering

First a table with resulting clusters from all analysed trajectories is shown (Figure 66). Selecting one will display another box containing the trajectory statistics and map representation (Figure 67).

Analysed trajectories - Clustering

Actions	Cluster №	DP Amount	Data Packages
	0	5	76a7dd1a-dd51-4b29-a199-9478b1e12d35,f85d3ed2-d936-477b-a493-a6e99418c252,d8afaa4b-5c44-44a4-820b-0cf063bc43de,1c22f19e-53a1-4010-9b2c-893471088df3,292b995e-53e8-47e7-aecc-534ec21a31a4
	1	1	6cfe946f-6e4b-4a54-aad4-e8dec5278060
	2	1	5829111e-59bb-403c-bf28-0772fc87e987
	3	3	778a7c1e-5e95-4e77-b4e4-62d9334daecd,def907b2-b8da-4cee-8c5a-17907aa7b616,3ebb7aae-692a-4ab1-a867-04cda1bd7ba2

Figure 66. Trajectory Analysis - clustering results list

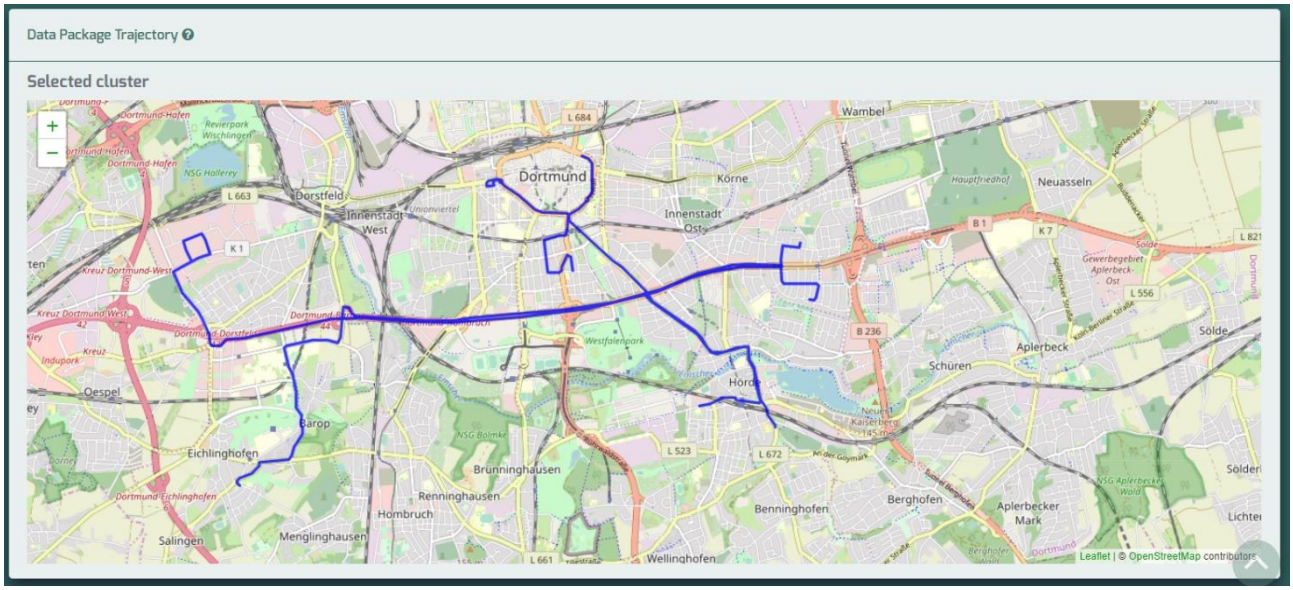


Figure 67. Trajectory Analysis - clustering results display

4.2.3.3.3. Interpolation

First a table with all interpolated trajectories is shown (Figure 68). Selecting one will display another box containing map representations of both the original trajectory and the resulting one from the interpolation function (Figure 69).

Analysed trajectories - Interpolation

Actions	Data Package ID	Analysis ID	ID	Packages
	76a7dd1a-dd51-4b29-a199-9478b1e12d35	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb749	1018
	f85d3ed2-d936-477b-a493-a6e99418c252	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb74a	825
	778a7c1e-5e95-4e77-b4e4-62d9334daecd	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb74b	176
	6cfe946f-6e4b-4a54-aad4-e8dec5278060	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb74c	329
	d8afaa4b-5c44-44a4-820b-0cf063bc43de	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb74d	491
	5829111e-59bb-403c-bf28-0772fc87e987	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb74e	969
	1c22f19e-53a1-4010-9b2c-893471088df3	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb74f	844
	292b995e-53e8-47e7-aecc-534ec21a31a4	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb750	225
	def907b2-b8da-4cee-8c5a-17907aa7b616	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb751	582
	3ebb7aae-692a-4ab1-a867-04cda1bd7ba2	5e6f5c3eb352112200bfb73e	5e6f5c42b352112200bfb752	630

Figure 68. Trajectory Analysis - interpolation results list

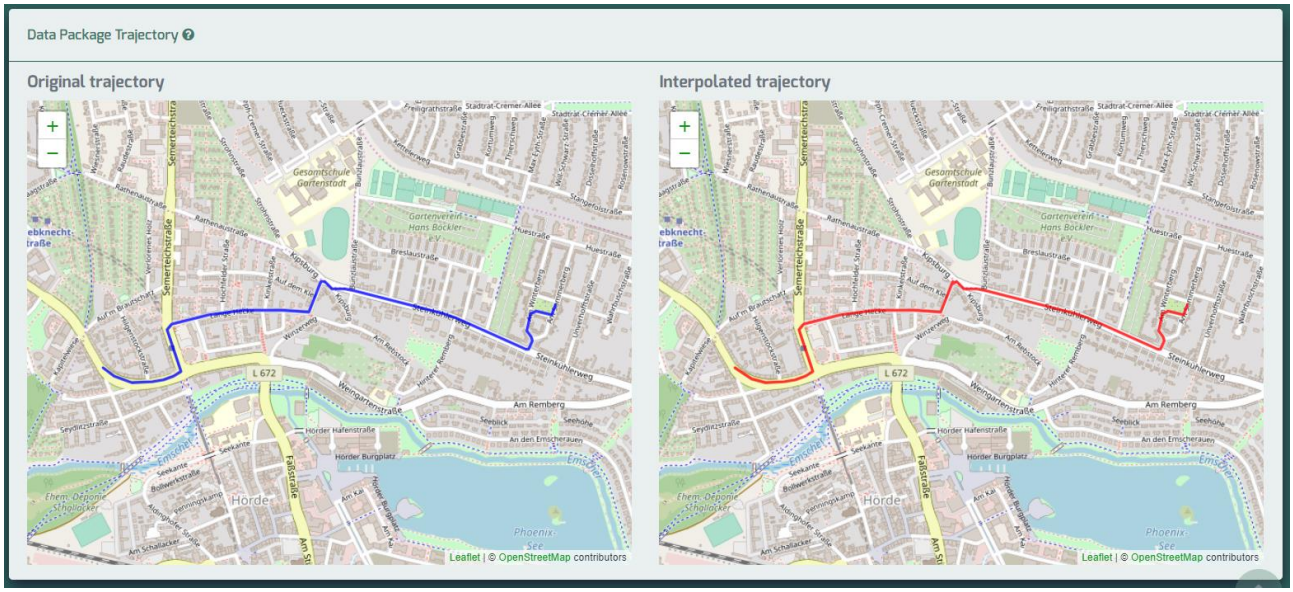


Figure 69. Trajectory Analysis - interpolation results display

4.2.3.3.4. Anomaly detection

First a table with all analysed trajectories is shown (Figure 70). Selecting one will display another box containing the map representation of the anomaly detection of the trajectory (Figure 71).

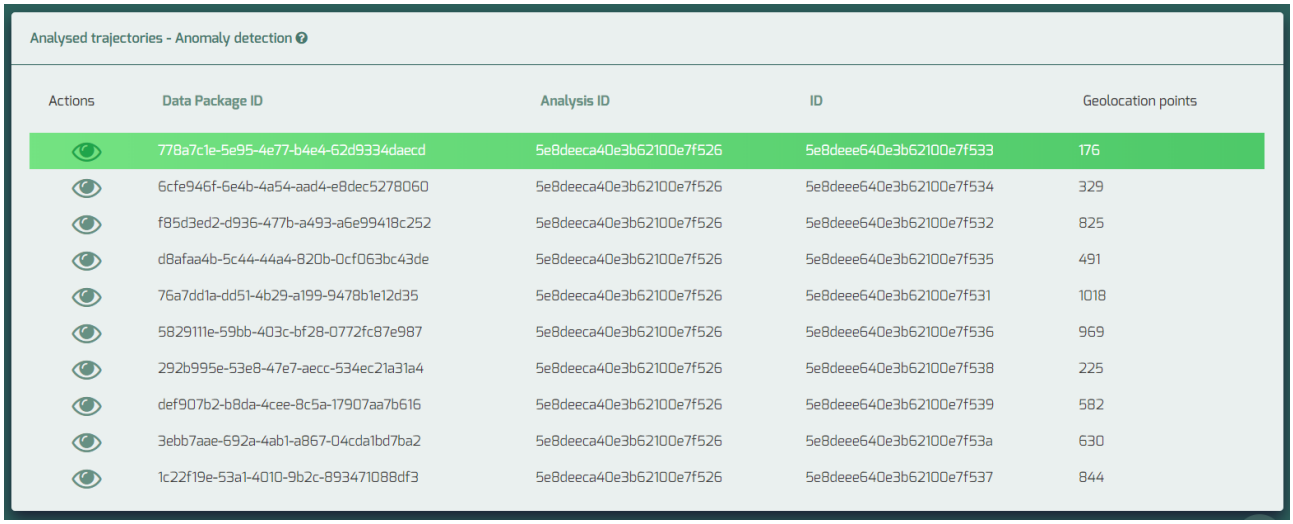


Figure 70. Trajectory Analysis – anomaly detection trajectories list

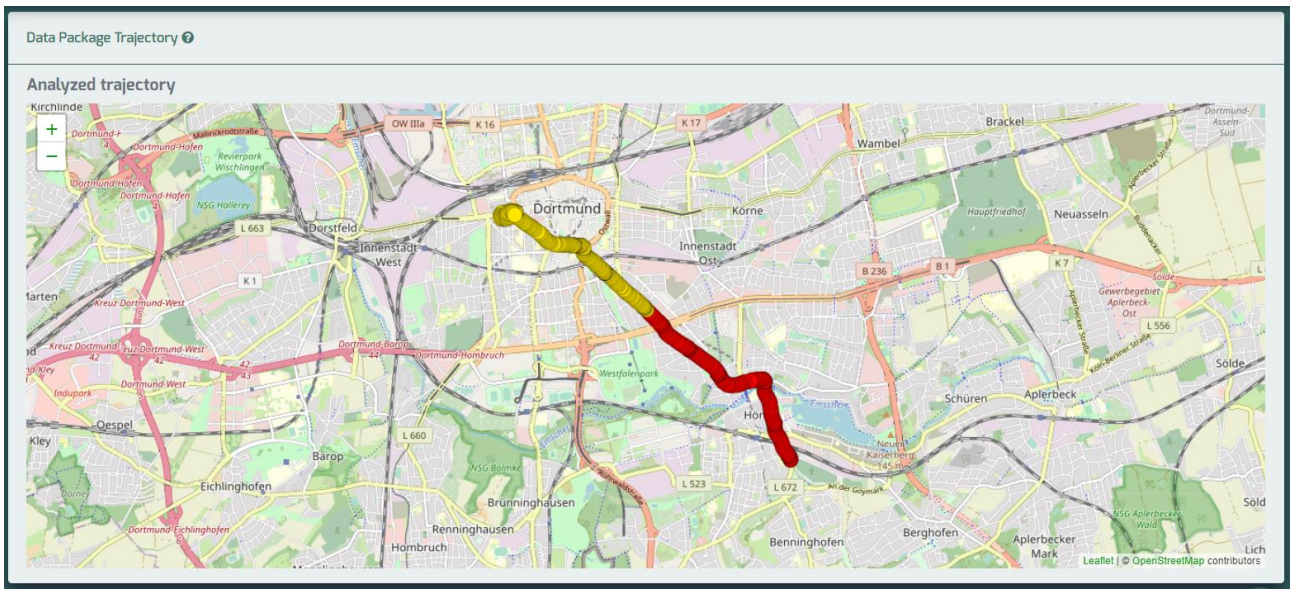


Figure 71. Trajectory Analysis – anomaly detection results display

4.2.4. Network Analysis

Networks can be understood as the graphic representation of relationships between entities. Each of these entities would be a node, connected to other similar ones by a link. The following figure depicts this concept.

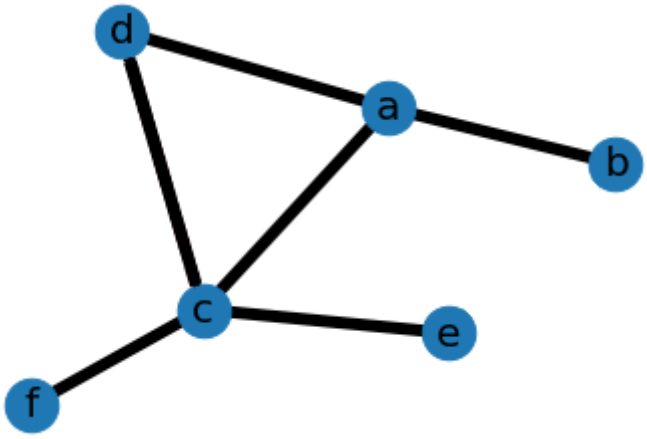


Figure 72. Networks - example of a simple network connecting entities (labeled with an alphabet letter)

In the CROSS-CPP ambit, a node would be a uni-dimensional signal from an available Measurement Channel (e.g. vehicle speed, temperature…). A link would simply be an indicator of the similarity between two signals (which is the reason why these links are called "weighted", because they are associated to a metric telling us how similar it is to its neighbour).

The specific shape of this type of graphs encloses information with respect to each specific node and its links (e.g. position with respect to other nodes, measurements of how outlying each node is…), as well as more general features referred to the arrangement of the network (e.g. how sparse or well-connected it is). This data can, in turn, be interpreted to provide the user with valuable insight with respect to the signals of the channel being analysed.

With this purpose in mind, the API described in this document provides a set of functionalities that enable the user to create and inspect a network's properties over time. Specifically, the developed functionalities allow to generate a network based on the available signals of a measurement channel. After the network is created, it will be continuously updated as new data from the analysed signals arrives, providing with updated information of their status.

Each of these services are described in detail in the forthcoming sections.

4.2.4.1. How to create a network

In order to perform analytics on a set of data, the Service Provider needs an active data request that has been approved for analytics by the CPP owner.

In the example the Data Request with name "all with all includes" is selected.

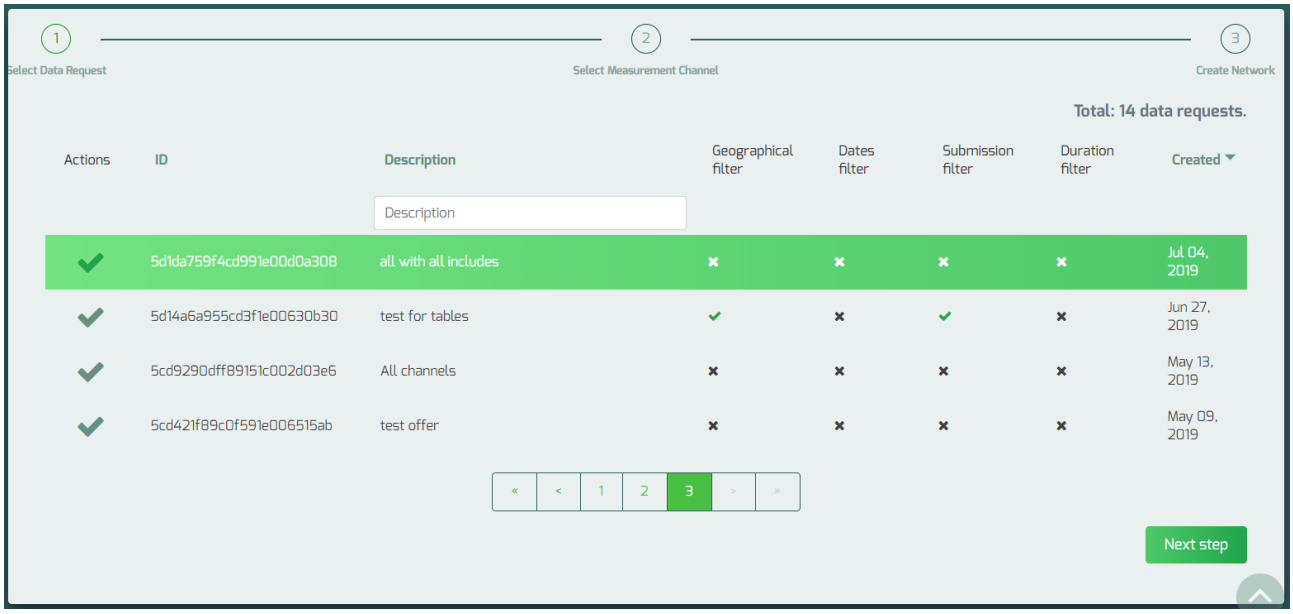


Figure 73. Create Network - select Data Request

Every Data Request comprises one or more measurement channels from which the request receives data. Next step is to decide from which of those measurement channels data analysis is desired. Every CPP in the network will conform a node with its value of the selected measurement channel.

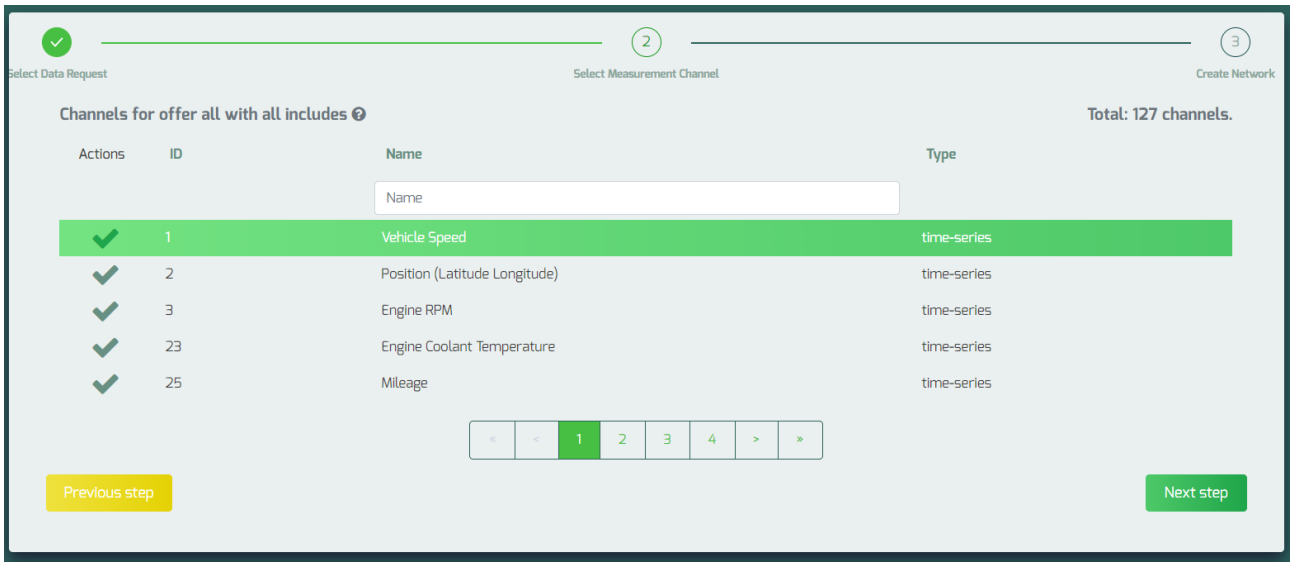


Figure 74. Create Network - select Measurement Channel

When selecting the Measurement Channel, a check of the network availability is performed with three possible outcomes.

Networks are unique, so if there is already an existing network for the selected Data Request and Measurement Channel a message will be shown.

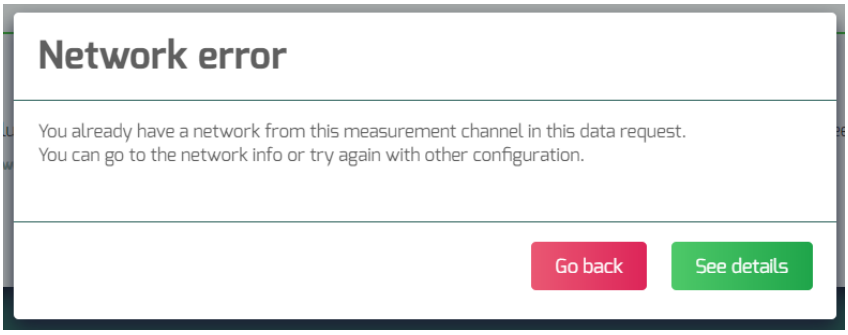


Figure 75. Create Network – Network already exists

If the Network does not exist but there are not enough available nodes a message will be displayed explaining it.

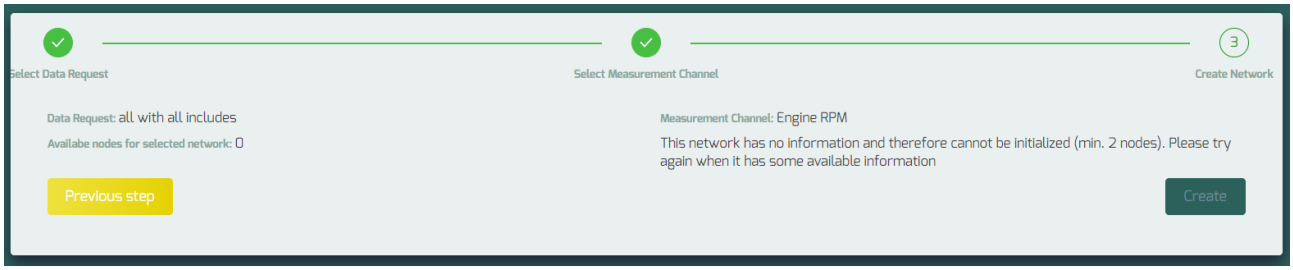


Figure 76. Create Network – no nodes available

If there is enough number of nodes creation will be enabled.

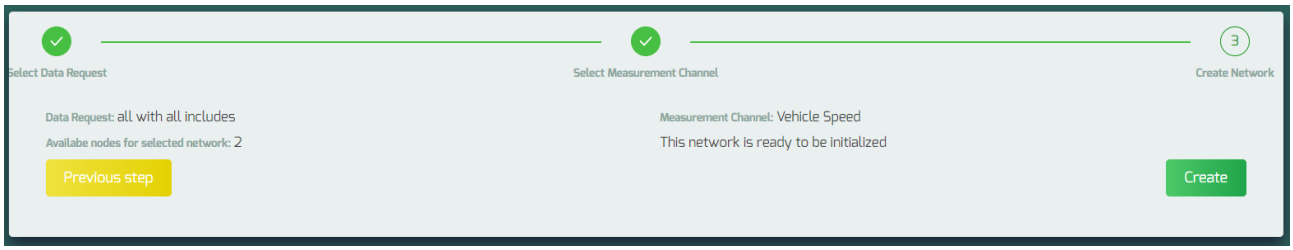


Figure 77. Create Network – Network ready

Once created a message will be displayed letting the user review the details and see the status of the network.

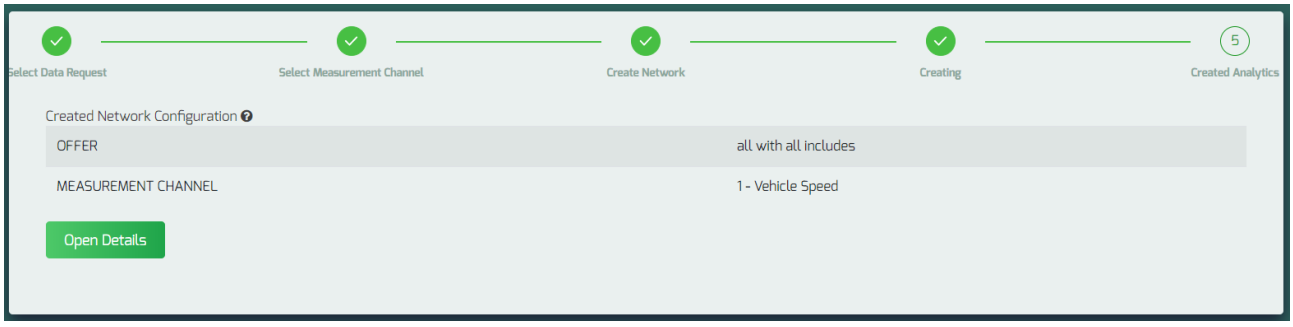


Figure 78. Create Network – Network created successfully

4.2.4.2. Get network status

Just clicking on the Networks menu will lead to the list view, showing a list of all Networks created by the Service Provider.

In this list users can filter displayed results by Data Request and Measurement Channel of the network.

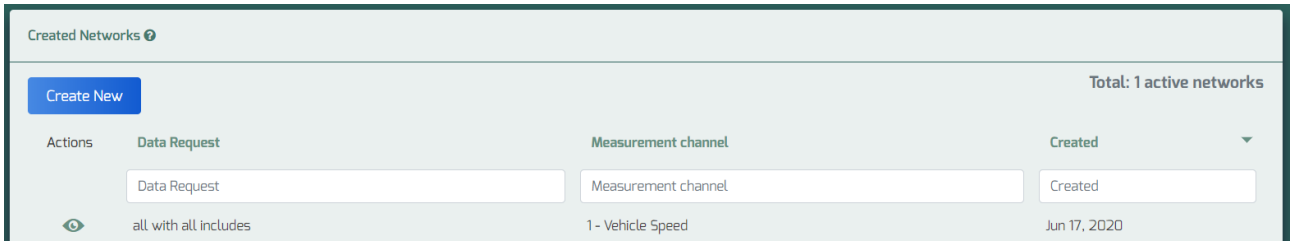


Figure 79. Networks list

Clicking on any of them will open the details view of that network, showing us the detailed configuration.

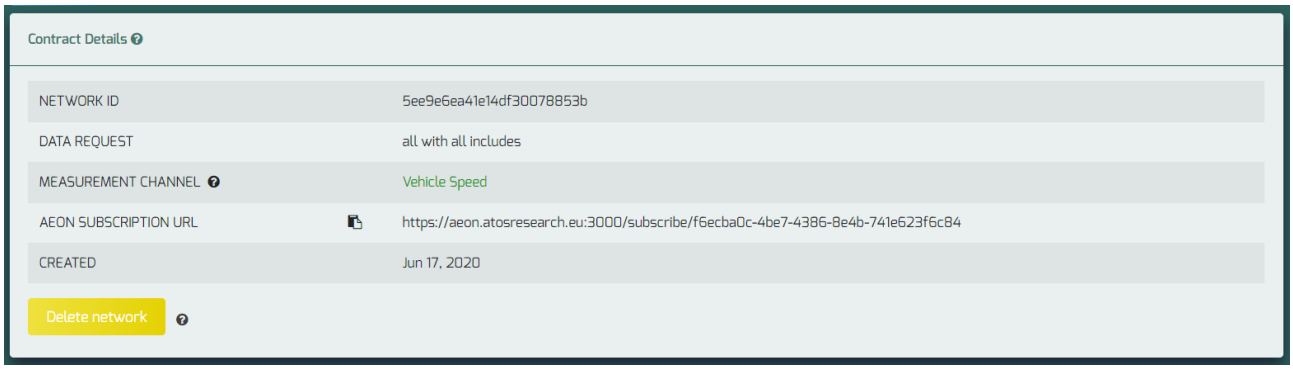


Figure 80. Network configuration details

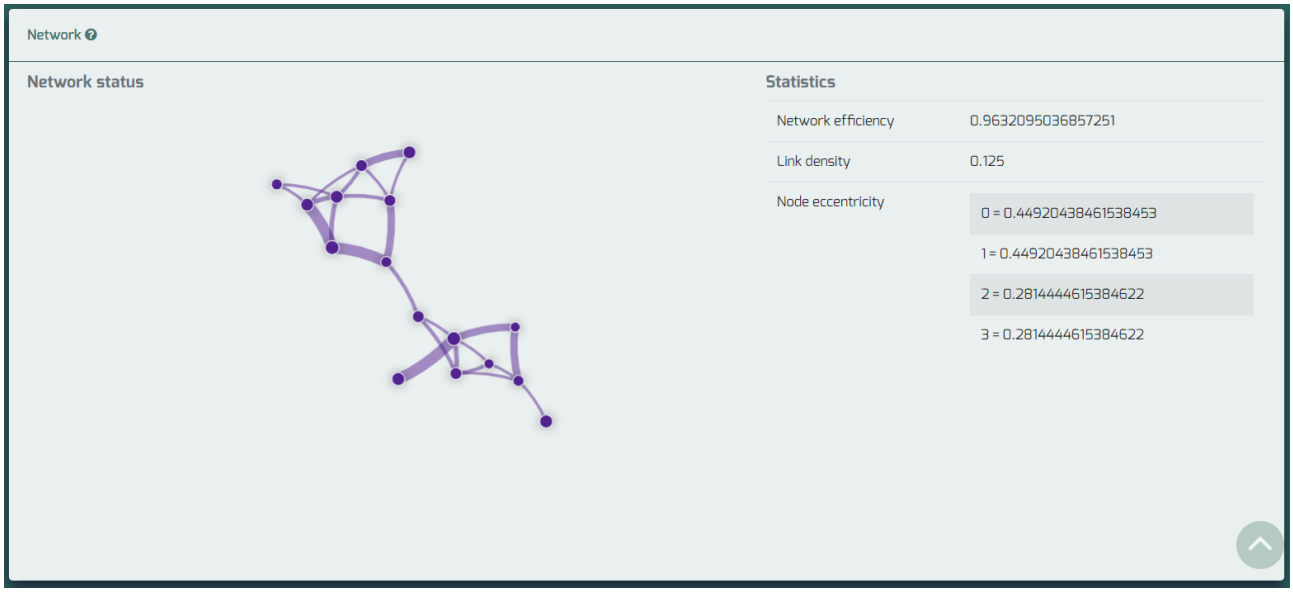


Figure 81. Network status

Network status include the following information:

- **Map of the network**: network representation on how nodes are connected.
- **Network Efficiency**: measure of how well-connected a network is. This metric is representative of the number of nodes that are connected between themselves by a direct link or through their connection to a path of nodes.
  Consequently, this metrics provides insight into how well connected the different groups present in a network are. This is, how similar are each group's signals between themselves (a desirable occurrence in well-formed networks). The higher it is, the more connections between nodes there are.
- **Link Density**: ratio expressing the amount of links present in the network with respect to its nodes. Homogeneous networks are typically link-dense, this is, they have a high amount of links connecting most of its nodes. A high value for this metric would be indicative of a network with its nodes closely related or, in other words, the corresponding signals being highly related.
  As this metric becomes smaller, the corresponding network is less well-connected and would thus imply the presence of separated groups of signals and/or outliers.
  This metric is generally directly correlated to network efficiency.
- **Node Eccentricity**: as an additional list containing the indices of the nodes from which to obtain the eccentricity. It can be thought of as a measure of how eccentric a node is. Thus, nodes with high eccentricity would represent a potentially outlying signal that may require further analysis or filtering.

4.2.5. Machine Learning

The last module of the Data Analytics Toolbox provides a generic interface for building advanced machine learning models from large data available in the Cross-CPP Data Marketplace. It

interconnects the Marketplace with existing frameworks (currently supported are Scikit-learn, Tensorflow, and PyTorch).

Note, that the processing pipeline differs from that of other analytics components as the training phase can take a significant time for large data. Consequently, the endpoints mostly work in an asynchronous manner. For example, the invocation of the ML model building just initiates the process and returns an identifier of this task. The caller can check the status of the process regularly. When the model is ready, it can be applied to new data. The basic scheme of the interaction with the machine learning components is shown on the following figure.

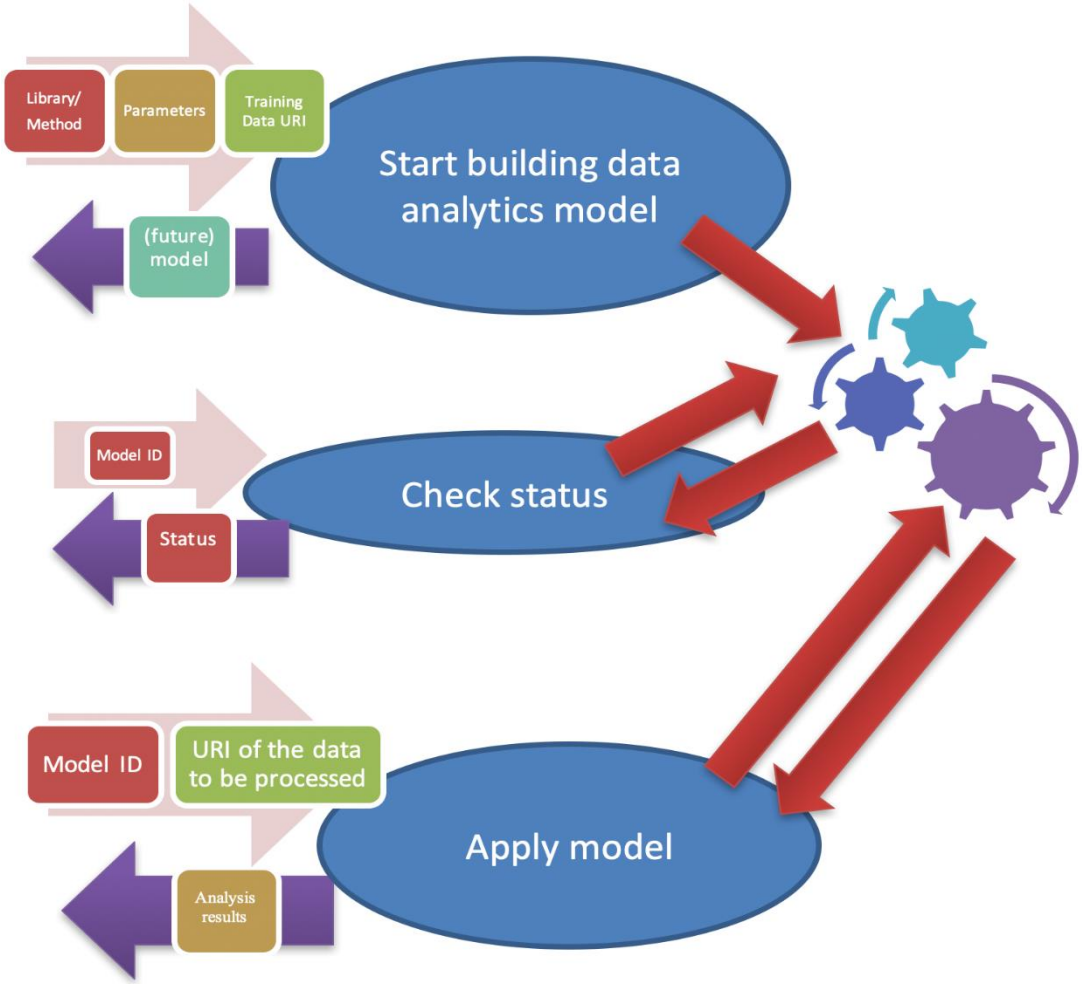


Figure 82. Machine Learning operation workflow

To link the scheme with the full workflow and functionality in the Marketplace, let us identify necessary components and actors in the process. A service provider interested in the use of the ML functionality identifies channels to be considered as inputs for the ML model first. In the example mentioned below, which classifies driver behavior patterns, the input channels consist of the speed, the location, and the rain intensity. The service provider defines a data view (see Section 6.4.1) and lists the selected channels in the call.

The machine learning functionality is employed to create a model from annotated data. The annotation of the data needs to be created in the next step. The service provider can add the

expected category to each row (to each combination of input channel values stored) in the data view created in the previous step. This is done by invoking the endpoint /sei/category/create first and naming the category to be predicted by the future ML model (for example, driver_behaviour_ml_prediction_output in the mentioned example). Then, the actual category values for the data view rows (for example, normal/risky/dangerous) are assigned by the endpoint /sei/category/assign to as many input data as possible.

Having the annotated data in the form of the data view, the service provider invokes the ML model training (model creation/building) process by calling endpoint /ml_model/batch/build with the data view ID as a parameter. When the subsequent call of endpoint /ml_model/status indicates that the model is prepared and ready, the model can be used to classify new data.

The created model is used on new data (to be automatically classified) by invoking the /ml_model/batch/apply endpoint. The service provider can create a new data view with the same (input) channels as the data view used for the model training, which corresponds to the new data to be automatically classified. Alternatively, the original data view can be used and the endpoint ignores the manually added categories (the last column) in the view. The response to the /ml_model/batch/apply request contains the predicted category (e.g., normal/risky/dangerous) for the data view rows provided in the parameter.

Service providers can be also interested in evaluating the quality of the ML model in question. According to the machine learning methodology, this should be done on a separate set of annotated data – a test set containing data not used in the model training). In a standard setting, the user creates a new data view, manually adds categories (in the same way as described for the training data view discussed above) and invokes endpoint /ml_model/batch/evaluate. The response contains a score of the model on the given data view – an accuracy of the prediction compared to manually provided categories.

4.2.5.1. How to create a Machine Learning model

In order to create a Machine Learning model, the Service Provider needs an active Data View that has been assigned with a category, and that has at least 10 rows with an assigned value.

In the example the Data View with name “Temperature, humidity and wind speed to assess outside weather” is selected.

First we get the Select Data View step.

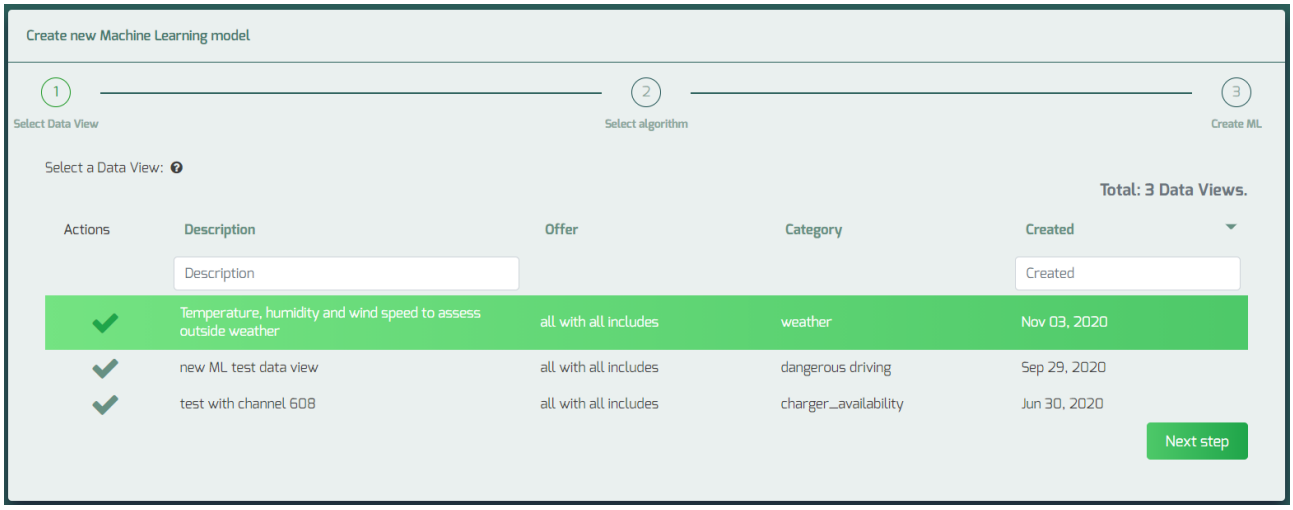


Figure 83. Create ML model - select Data View

Once a Data View is selected it is checked to know if it fulfil the necessary requirements mentioned before.

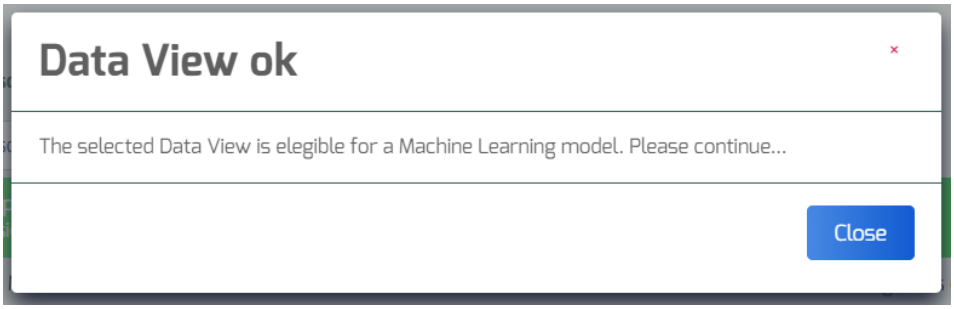


Figure 84. Create ML model - selected Data View ok

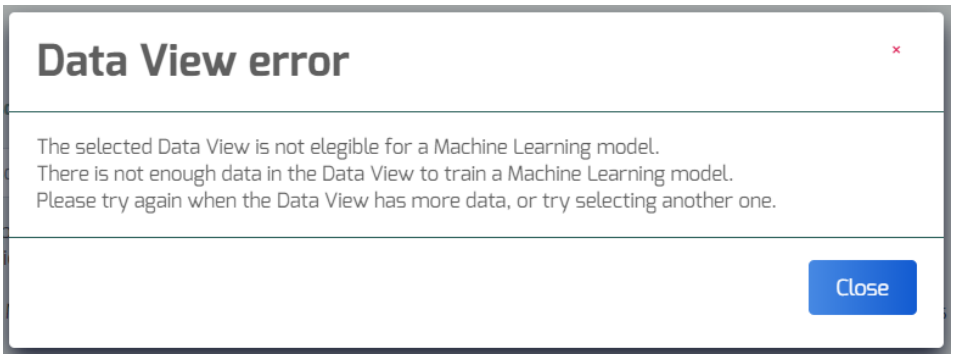


Figure 85. Create ML model - selected Data View not eligible

The next step is to select the type of algorithm to use when training the model with the selected Data View collected data.

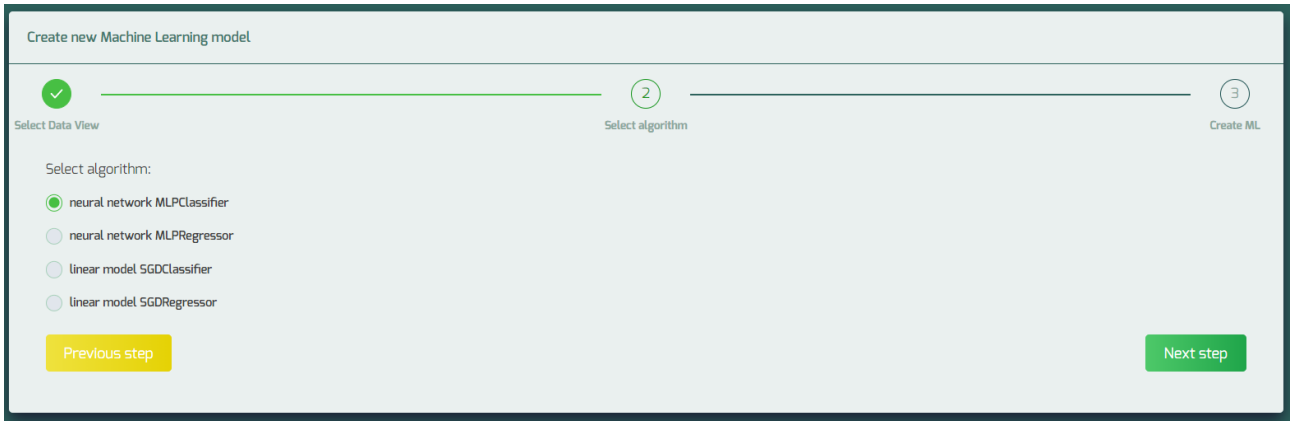


Figure 86. Create ML model – select algorithm

Finally, we get a quick review on the selections made and get the choice to create the model or return to change something.

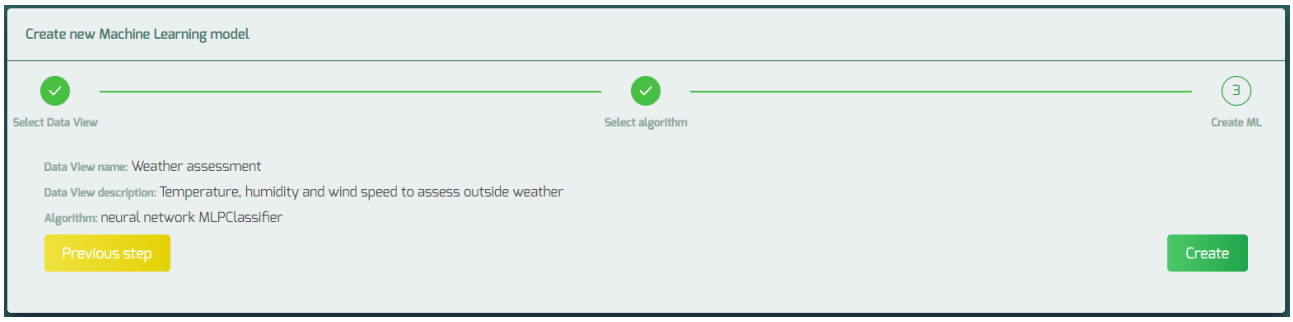


Figure 87. Create ML model – create step

Once created we get the choice to go directly to the details screen:

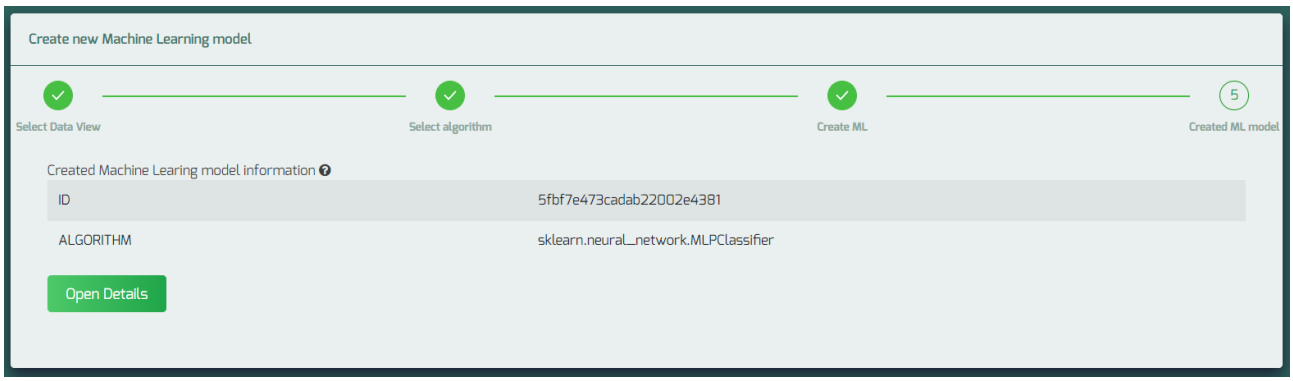


Figure 88. Create ML model – ML model created

4.2.5.2. Review Machine Learning model

Just clicking on the Machine Learning menu will lead to the list view, showing a list of all ML models created by the Service Provider.

In this list users can filter displayed results by algorithm of the model.

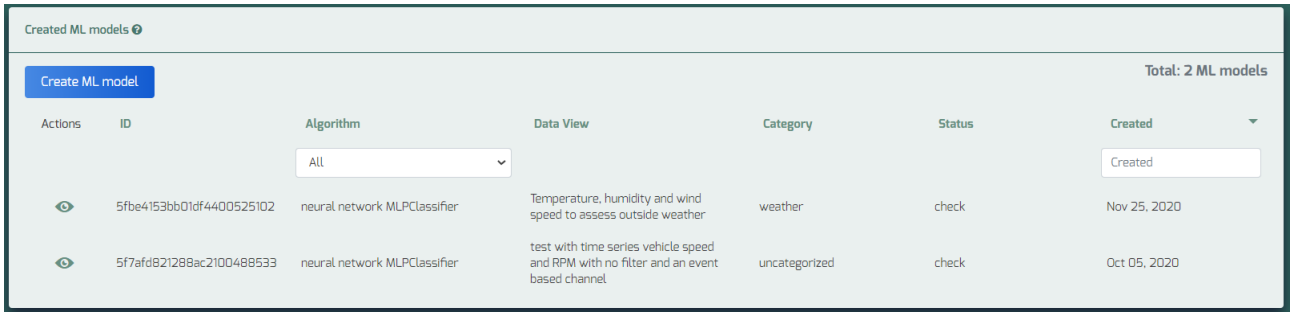


Figure 89. Machine Learning model list

Clicking on any of them will open the details view of that network, showing us the detailed configuration, including the ML model configuration along with the associated Data View and the currently known status of the model training.

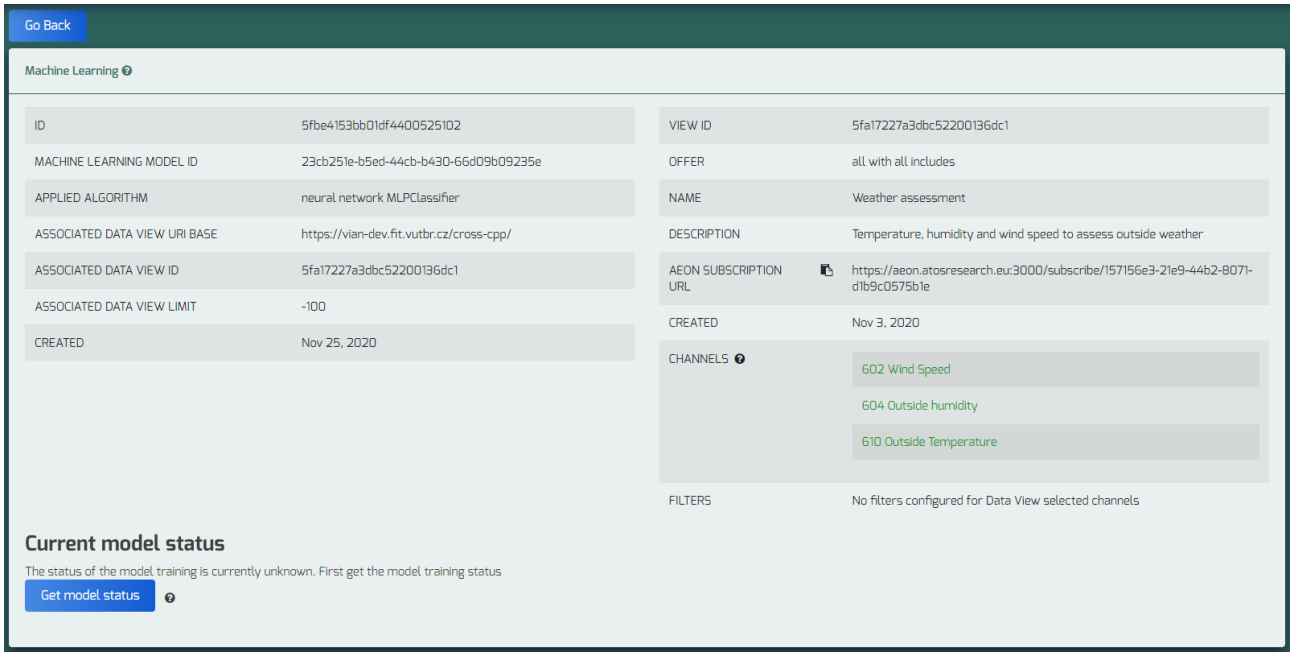


Figure 90. Machine Learning model details view

4.2.5.3. Operate with a Machine Learning model

At first every ML model will have the assigned status of "check", which means the current training status of the model is unknown. To start operating with the model first is needed to know the status of the model training. To do that just click on the "get status" button.

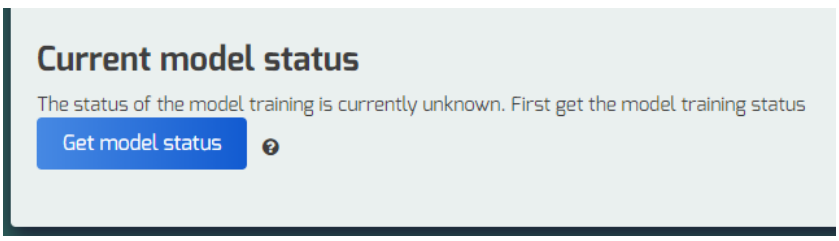


Figure 91. Get Machine Learning model status button

Once we have the status of the model, we have three possible outcomes that gives us different possibilities to operate with the model:

- The training has been initialized or is running at the moment. In this case we will have to wait for the training to be completed. We can also stop the training by killing the model, resulting in a failed training.
- The training has been completed successfully. This is the desired outcome and gives us the opportunity to actually work with the model.
- The training has failed. An error message is displayed, and we can delete the model.

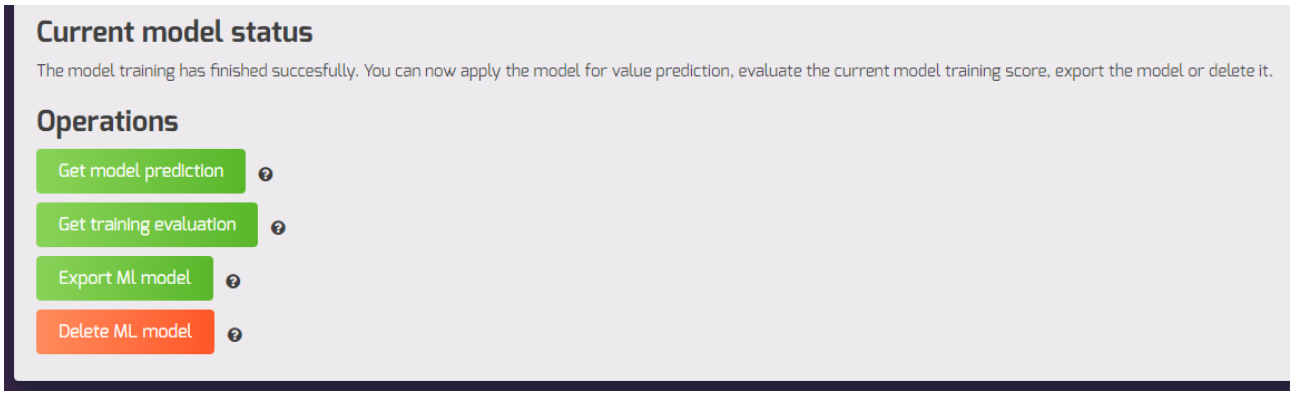
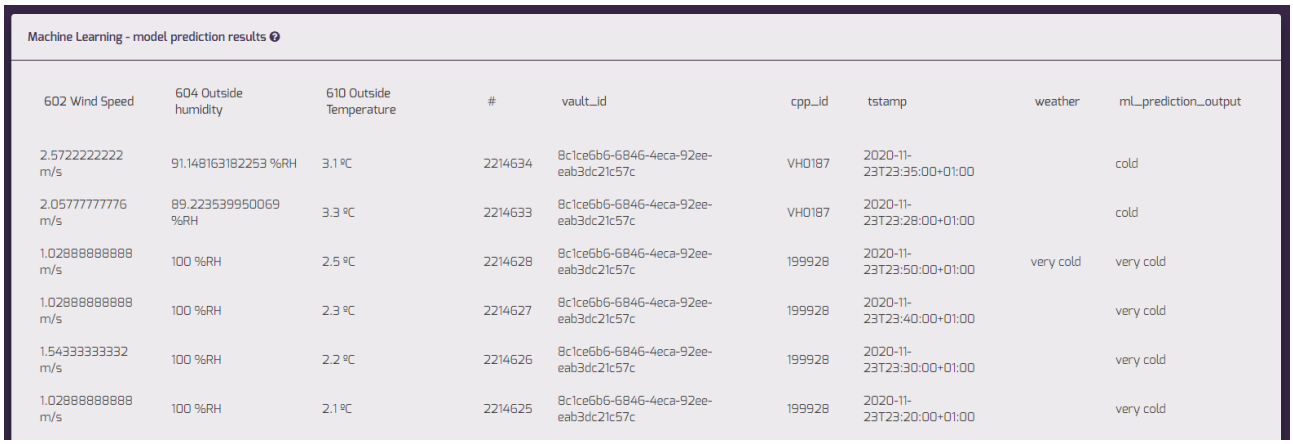


Figure 92. ML model status operations example

4.2.5.3.1. Get model prediction (apply ML model)

You can get a prediction of category values for the uncategorized rows of the Data View by clicking on the "get model prediction model" button.

A table will be displayed with the allocated values.



602 Wind Speed	604 Outside humidity	610 Outside Temperature	#	vault_id	cpp_id	tstamp	weather	ml_prediction_output
2.5722222222 m/s	91.148163182253 %RH	3.1 ºC	2214634	8c1ce6b6-6846-4eca-92ee-eab3dc21c57c	VH0187	2020-11-23T23:35:00+01:00		cold
2.05777777776 m/s	89.223539950069 %RH	3.3 ºC	2214633	8c1ce6b6-6846-4eca-92ee-eab3dc21c57c	VH0187	2020-11-23T23:28:00+01:00		cold
1.02888888888 m/s	100 %RH	2.5 ºC	2214628	8c1ce6b6-6846-4eca-92ee-eab3dc21c57c	199928	2020-11-23T23:50:00+01:00	very cold	very cold
1.02888888888 m/s	100 %RH	2.3 ºC	2214627	8c1ce6b6-6846-4eca-92ee-eab3dc21c57c	199928	2020-11-23T23:40:00+01:00		very cold
1.54333333332 m/s	100 %RH	2.2 ºC	2214626	8c1ce6b6-6846-4eca-92ee-eab3dc21c57c	199928	2020-11-23T23:30:00+01:00		very cold
1.02888888888 m/s	100 %RH	2.1 ºC	2214625	8c1ce6b6-6846-4eca-92ee-eab3dc21c57c	199928	2020-11-23T23:20:00+01:00		very cold

Figure 93. ML model value prediction table

4.2.5.3.2. Get training evaluation (evaluate ML model)

You can get an evaluation score of the model training. Click the "Get training evaluation" button to get the numeric score.

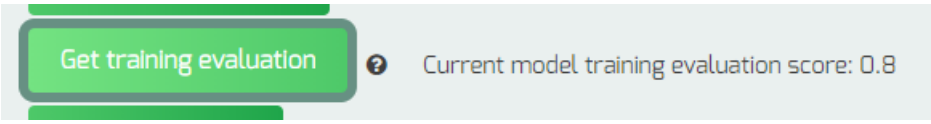


Figure 94. ML model evaluation score

4.2.5.3.3. Export ML model

You can export the current ML model by clicking the "Export ML model" button and downloading the file in your device.

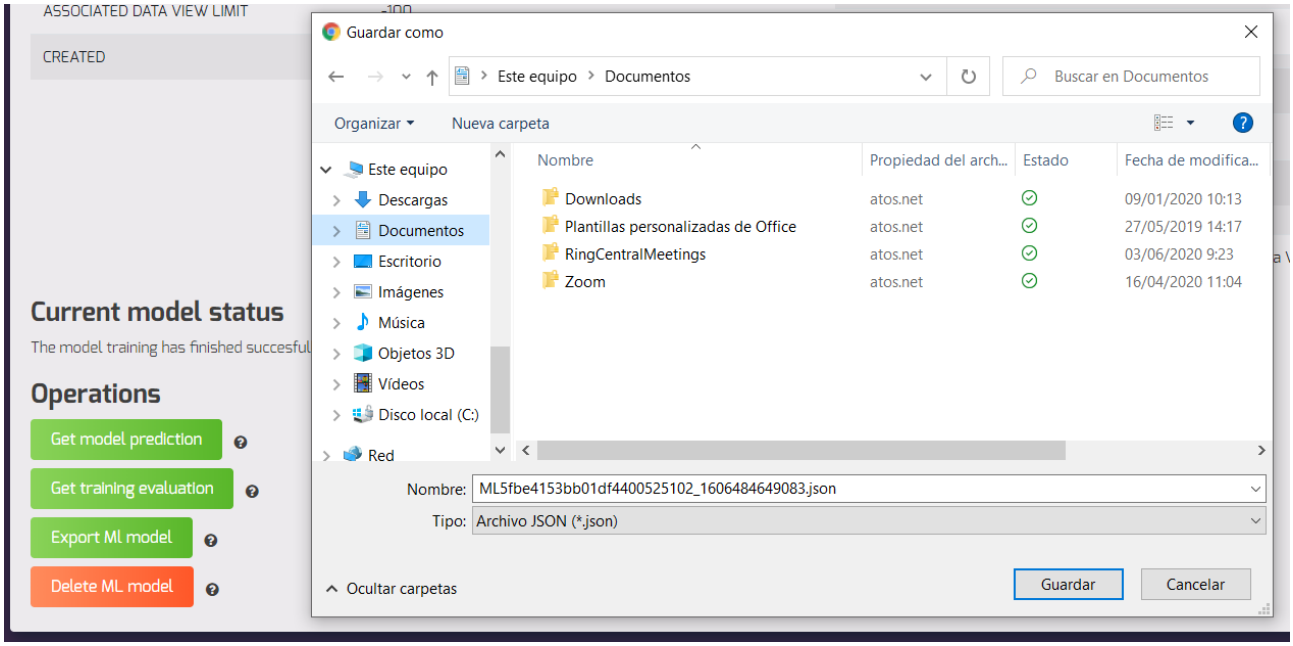


Figure 95. export ML model

4.2.5.3.4. Delete ML model

Once the training has finished, whether it has failed or has been completed, we can delete the model at any time by just clicking in the "Delete ML model" button and confirming. If the model is still running, we first have to kill it to stop the process, and then delete it.

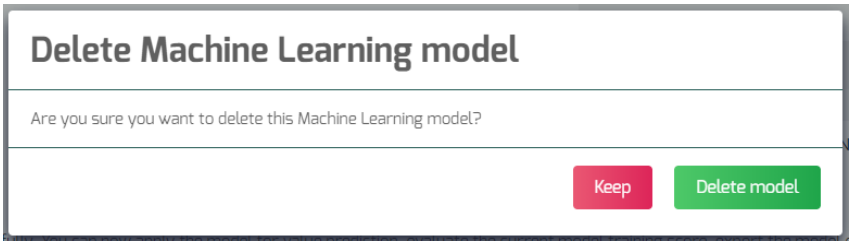


Figure 96. Delete ML model confirmation

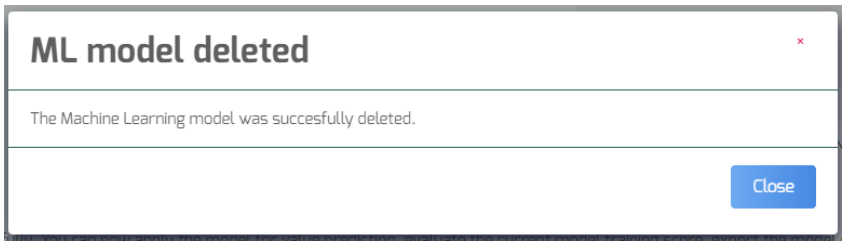


Figure 97. ML model deleted response

5. Security and authorization

5.1. Context Sensitive Security

The Cross-CPP project aims to provide a cross-sectorial Ecosystem, where unified quality of data, commercial confidentiality, privacy, IPR and ethical issues may arise in dynamically changing and varied combinations, requiring the services provided to be adapted to the specific needs of the users and of the providers of CPP data streams under the prevailing contextual conditions.

More detailed explanation of the Context Sensitive Security module can be read in the Cross-CPP Ecosystem guide (to be found under www.cross-cpp.eu).

5.1.1. How does the CSS module work?

To cope with the complexity of these needs, Cross-CPP has an innovative solution that employs a novel combination of two key technologies:

- A characterization for various information about the Cross-CPP Ecosystem environment, a means to extract that information from events occurring in the environment, and to make the corresponding conditions (such as time of day, calendar dates and events, location, user identity, and other specified conditions) available to other processes that can usefully apply that information to adapt services. This capability is embodied as the Context Monitoring and Extraction (CME) module.
- A flexible way to express confidentiality, privacy, IPR and ethical constraints in the form of precise rules that are subject to variable interpretation according to the current contextual conditions, or that are subject to modification in response to changes in the context. This capability is embodied as the Context Sensitive Security (CSS) module.

The CSS capabilities are used to represent rules corresponding to the contracts made between the Service Providers that make a Data Request, and the CPP owner (or a particular CPP, since a CPP owner can possess more than one CPP). Upon acceptance of the Data Request, an access control policy is constructed that codifies the access rights of Service Providers to data from various CPPs and this policy is used during runtime to grant or deny access to the data.

The CSS can function both in the mode of taking CPP extracted context into account or, in the case that the CPP owner does not allow the extraction of context data, a mode that is context independent, as reflected in the configuration of the CSS and CME modules.

5.2. Authorization

Authorization to access CPP data is achieved through an appropriate configuration in the CSS module, including the access control policy and declaration of the conditions and context variables upon which the policy is dependent. The Marketplace acts as the enforcer of the policy decisions rendered by the CSS module in response to access queries. An access query specifies the user requesting data access, the data to be accessed, and the operation requested on the

data item. The CSS module interprets its current policy for the given parameters and provides a grant or deny response to the Marketplace, which the Marketplace in turn carries out.

The access control policy specified in the Declarative Policy Language extended for context sensitivity is contained in a file that the CSS Policy Server is instructed to load and use as the current policy. The CSS Policy Tool is used for the development and testing of a policy file. The declaration of conditions and context variables is done in the two files conditions.pl and context.pl respectively. These files are generated by the Context Sensitive Security Setup tool, which provides a simple graphical user interface for configuring context sensitivity.

F.A.Q.

Cross-CPP data-marketplace

Q: What is Cross-CPP data-marketplace?

A: Cross-CPP data-marketplace connects CPP Owners and Service Providers for selling and acquiring Connected Vehicle and Home Building data under the Common Industrial Data model (CIDM). It offers a secure and privacy preserving experience when selling or buying sharing big data, by having the full control over your data shared, to whom and for what purposes.

Cross-CPP offers to cross-sectorial Service Providers, the possibility to search for more than 200 sensor signals, display advance visualization representations (such as Histograms, Geo-Histograms, Time Series) and retrieve those datasets in a seamless experience thanks to the open SDK-API created.

Q: How do I, as Service Provider, register into Cross-CPP data-marketplace?

A: You can find the registration form by clicking the "Sign on!" button in the landing page. Select "Service Provider" role and fill the fields to request your registration. Once your registration is validated by a system administrator an email will be sent to you to confirm your access, Then, access the link in your email, login and accept the consent to start using the Cross-CPP Marketplace.

Q: What do I have to do in order to start working with Cross-CPP data-marketplace?

A: Once registered you must be familiar with the CIDM, as it is the format in which you will receive the data you request. You must also be familiar with AEON, as it is the communication channel used to send the data. You can find information for both in the Service Providers Developers Guide.

Cross-CPP data model

Q: What is the Common Industrial Data Model (CIDM)?

A: The CIDM is a standardized data model for industrial data-driven services. You can find extended information about the CIDM in the Service Providers Developers Guide

Q: Which are the benefits and advantages of using the CIDM model for data -driven services:

A: -The CIDM constitute a major business and technical advantage for Service Providers:

- The CIDM provides a brand-independent and transparent data model, which harmonizes proprietary data into generic datasets independently of any cross-sectorial Industry
- It is built on an open and highly scalable automotive big data format (JSON Schema).
- Active community of service providers increasing the number of signals available from vehicles and Smart Buildings to be recorded as well as the type of measurement channels can be modified or extended

- The Company Backend also provides an origin certification as a CIDM feature to support the validation and verification of origin, integrity and completeness of data. The intention is to protect the data inside the Data Package against manipulation.

Q: What is a signal?

A: A signal is the information provider of each CPP. They are the perception organs of CPPs, and it is their main duty to detect physical phenomenon and chemical quantities. They observe the environment and generate data in the CIDM format. An example could be "speed" or "latitude"

Q: What is a channel?

A: A channel is the way the physical signals and their sampled measurements are implemented and represented in the CIDM format. Some examples could be "Vehicle Speed" using the signal "Speed" in a time-series or in a histogram format, or "Position" using both "Latitude" and "Longitude" signals.

Q: Can I request a new signal or channel?

A: Cross-CPP data-marketplace offers a wide variety of signals provided by the manufacturers. The catalogue is extensive and can be filtered in many ways. If even then you can't find the signal that you need and/or think can be provided by any of our company backends, please contact us in: support@datagora.eu

Cross-CPP marketplace components

Q: What is the Data Discovery component and how does it work?

A: The Data Discovery component is a tool that allows you to find what data you can access through the marketplace. There you can use the filters provided to narrow or enlarge your results and create Data Request based on the configured search. You can think of it like a test of what would you receive if you publish that request.

Q: What is the Context Monitoring and Extraction module and how can help my Organization?

A: The Context Monitoring and Extraction module allows Cross-CPP to suggest you signals to add to your current Data Discovery signals, based on the context model of the CPP and signals already selected. This might help you find data of interest that you would miss otherwise.

Q: How does the Suggestion component work?

A: The Context Monitoring component allows Cross-CPP to suggest you signals to add to your current Data Discovery filters, based on the context data of the signals already selected. This might help you find data of interest that you would miss otherwise.

Q: How can I use the Context Component in my Data Requests?

A: During the creation of a Data Request the Cross-CPP Marketplace gives the Service Provider the choice of making the Data Request context sensitive and configure the available context filters.

Q: What does it mean my Data Request is context sensitive?

A: The Data Request is context sensitive in case the Service Provider selects some of the CME module data filters to restrict the data packages that it wants to receive. For instance, if the Service Provider just wants to receive data that is provided by vehicles that are driving in the highway and this type of context data is extracted then one case say that the Data Request (and data packages that are produced from this request) is context sensitive.

Q: Which are the consequences of making my Data Request context sensitive?

A: Since you can do this by selecting several context information that could be used to filter the Data Request, the consequence of adding such a filter will be that you receive a smaller set of data packages (filtered by the selected context information) than you would usually receive in the case that no context information is selected.

Data Requests

Q: What is a Data Request?

A: A Data Request is a set of filters that defines which type of data would you like to receive. You would receive data from CPP owners that have accepted these requests through each request unique AEON channel.

Q: How do I create a Data Request?

A: You must get to the Data Discovery and define the data you are interested in through the filters given. Once set you give it a descriptive name, so CPP owners guess in a glance the nature of the request.

Q: How do I receive data from a Data Request?

A: In the very moment a Data Request is created an AEON channel is assigned. You can find the channel configuration in each data request details view. Data from CPP owners that have accepted the request will be sent through each AEON channel assigned to each eligible data request, meaning you can receive data from the same user and signal from more than one request.

Q: Can I modify my Data Request once created?

A: No. The acceptance of a request by a CPP owner implies a consent from its side. Modifying the request would make invalid such consent. Therefore, you can create another Data Request with the new desired configuration.

Q: How can I delete the Data Request published in the Marketplace

A: - You can delete any published Data Request in the details view of that request, or through the provided API endpoint (see Service Providers Developers Guide). This action cannot be undone, and all acceptances of the request will be automatically terminated.

Q: Can I use the data collected for other purposes not described in my Data Request?

A: No. The acceptance of a request by a CPP owner implies a consent from its side. That user allows certain usage of the data given and only for the purposes described in the request by the Service Provider (such as use for analytics or evaluate its context).

Toolbox

Q: What is the toolbox?

A: The Toolbox is a set of tools that offers Service Providers a way to generate analytics from the data obtained from Data Requests or further filter this data in order to get exact measurements or use the service as a notification system.

Q: How can I request analytics?

A: Any analytics uses data packages received from a Data Request, meaning that first the data Request must have been accepted by CPP Owners and started receiving data. Then, on the Cross-CPP Marketplace a Service Provider can create analytics based on those Data Requests.

Q: How do I get my analytics results?

A: Depending on the analytics type results can be instantly shown on the screen, as a diagram, chart, map and so on, or a new AEON channel is provided in order to subscribe to upcoming analytics results. Every analytics type explains the way to get the results on screen.

Q: How can I consult my analytics?

A: Any created analytics can be consulted on the Cross-CPP Marketplace under Analytics section. There a Service Provider can get the AEON channel for subscription, see results of one-time analytics, or even delete them.

Q: Which Time Series forecasting method should I use for my time series predictions?

A: Prior to attempting to predict future values of a time series, you should verify the assumptions your selected algorithm does on the input data. For example, ARIMA will assume that your data is autoregressive and Regression trees will not be able to detect trends on data (thus not being advisable for a time series in which trend is a significant feature). If you are unsure about the underlying patterns of your data, Neural Networks may work best for you, as it does not rely on any a-priori assumption.

Q: How does the embedding dimension parameter “m” affect the entropy metrics?

A: The time-series are defined by sequences of points. The embedding dimension refers to the number of points use to evaluate. So then, smaller embedding dimensions would yield more

detailed information while larger numbers tend to be more general. Bear in mind if the embedding dimension is too low maybe the results may be chaotic.

Q: How do I choose between a Pearson and a Spearman's correlation coefficient?

A: This decision will depend on the nature of your data, as these coefficients measure different types of association between variables (Pearson quantifies linear relationship, while Spearman does so with non-linear patterns). Their use is not exclusive, so even both could possibly be valuable to your analysis.

Q: What are the units of the trajectories module?

A: For distances, the meter, and for times the second. Then the velocity is expressed in m/s.

Q: What is the interpretation of the results of clustering the trajectories?

A: Each trajectory included in the data is assigned to a numerical ID that refers to the cluster that is being assigned. By instance, all trajectories of similar length from east to west are grouped in the cluster number 1, so the results are 1 for each trajectory.

Q: What do the nodes represent in the networks module? And the links?

A: Each node is an object. It could represent a house, building, a car. The links are an abstract representation of some magnitude that is being measured. Links could be based on distances by instance.

Q: What is a Data View?

A: A Data View is a configuration to get data filtered by specific values constraints for one or more Measurement Channels included in a Data Request.

Q: How can I create a Data View?

A: Go to Data Views section under Toolbox in the Cross-CPP Marketplace. There a step by step guide will be offered.

Q: How do I get the data view results?

A: Service Providers can consult their generated Data Views through Data Views under Toolbox section. There the configuration can be consulted, as well as the AEON channel to subscribe to. Also, the options of retrieving the latest data or even deleting the Data View are offered there.

Q: How can I use a created ML model?

A: A machine learning model that was built in a previous step can be applied on new (unseen, unannotated) Data View rows by invoking the Apply function. The service estimates the category on the given data rows and outputs it as the response.

Q: My service processes a lot of data, is the ML component ready for big data application?

A: Yes, even a large neural network model loads in less than 480 ms and it can apply the category prediction with the speed of more than 5000 rows per second on the Cross-CPP testing infrastructure.

Q: What ML methods should I try first?

A: Although the ML components support all new and fancy neural network methods available in Scikit-learn, Google TensorFlow, and Facebook PyTorch libraries, we suggest to start with the simple linear method of Stochastic Gradient Descend or the basic Multi-Layer Perceptron neural network for your initial experiments. Often, the quality of results provided by these models is satisfactory and performance gains can be brought by additional annotations of the data.

AEON

Q: What is AEON?

A: AEON is a cloud platform to create applications with real time communications channels. You can find extended information about AEON in the Service Providers Developers Guide

Q: How can I use AEON to subscribe to my data requests and data analytics?

A: AEON provides an SDK (Node.js, JavaScript and Java) that encapsulates the complexity of connecting to a socket server. Please refer to section 6.4, document examples and online page for extended documentation.

Q: How do I create or configure an AEON channel?

A: You don't have to create or configure any channel. All needed AEON channels, such as for data requests or analytics, are created and assigned by Cross-CPP. You only have to use the channels given. The channels configuration can be found in the details view of each data request or analytics.

Glossary

Administrator: Cross-CPP marketplace system administrator

Autoregressive data: In a time series domain, it refers to data which values depend on prior data points from the same time series.

AEON: AEON application

AEON application: publication/subscription based communication application

AEON channel: set configuration for communication between two actors through AEON application

Analytics Toolbox: set of available analytics functions to be requested by the Service Provider

CIDM: Common Industrial Data Model

CIDM model: standardized data model for industrial data-driven services

CME: Context Monitoring and Extraction

Company Backend: system of an OEM that provides its users data to the Cross-CPP marketplace

Contract: entity that resumes the acceptance of a data request from a CPP owner

CPP: cyber-physical product

CPP Data: data created by a CCP and sent to the system by the Company Backend

CPP owner: CPP owner which CPP is registered in the Cross-CPP data-marketplace

Cross-CPP: System

CSS: Context Sensitive Security

Data Request: set of configurations that define a scope for CPP Data to be received by a Service Provider

Data View: set of configurated filters to receive specific values from a Data Request through a different notification channel

Entropy: It is usually explained as the order of a system. It is more accurate to understand the entropy as the lost information of a system. This definition for data classification problems implies the algorithms search for the variables that reduces the lost information of the system, those are the best classifiers.

Homoscedasticity: property of a multivariate domain in which the variance of each variable's error term is equal.

Machine Learning Analytics: A type of Analytics included in the toolbox

Marketplace: Marketplace Web Application

Measurement Channel: sampler of the data the signals process

Monotonic relationship: A type of association between variables that occurs when two variables tend to increase or decrease in the same direction, but not following a linear pattern (linear relationship).

MP: Marketplace

Multilayer Perceptron: A type of Artificial Neural Network with a varying number of hidden (processing) layers.

Networks Analytics: A type of Analytics included in the toolbox

Network Diameter: Value indicating the shortest distance between the two most distant nodes in a network.

Network Efficiency: Measure of how well information is exchanged between the nodes of a networks.

Node Eccentricity: Value representing the centrality of a network's node, or how close it is to the other nodes in the network.

OEM: Original Equipment Manufacturer

Rank Correlation: A type of correlation measure that quantifies ordinal association between two variables.

Service Provider: actor who receives the data created by owners to use it on the creation or improvement of services

Service Provider Wallet: group of MP functionalities for Service Providers

Signal: information provider of the data the CPP sensors generate

Stationarity: Property of a time series indicating that its statistical properties (e.g. mean, variance…) do not change over time.

System: the whole lot of applications that conforms Cross-CPP, including Marketplace Web Application and Marketplace Server.

Time Series Analytics: A type of Analytics included in the toolbox that analyses drifts in the data flow of time-series type channels.

Time Series Complexity: Measure of the presence of nonlinear patterns that explain the behaviour of a time series' data.

Trajectories Analysis: A type of Analytics included in the toolbox that uses trajectory related signals.

Figures

About Cross-CPP

The objective is to establish an IT environment for the integration and analytics of data streams coming from high volume (mass) products with cyber physical features, as well from Open Data Sources, aiming to offer new cross sectorial services and focusing on the commercial confidentiality, privacy and IPR and ethical issues using a context sensitive approach. The project addresses cross-stream analysis of large data volumes from mass cyber physical products (CPP) from various industrial sectors such as automotive, and home automation. The business objective of the research is to allow for analyses of such data streams in combination to other (non-industrial, open) data streams and for the establishment of diverse enhanced sectorial and cross-sectorial services. The project will develop: (i) New models for integration and analytics of data streams coming from multi-sectorial CPP, including shared systems of entity identifiers applicable to multi-sectorial CPP (as well as the definition of agreed data models for data streams from multiple CPP aiming at defacto standard; (ii) Ecosystem, including a common Marketplace, and methodology to use such models to build multi-sectorial cloud based services, (iii) Toolbox for real-time and predictive cross-stream analytics, context modelling and extraction, and dynamically changing security policy, privacy and IPR conditions/rules and (iv) set of services such as services based on a combination of data streams from home automation and (electrical) vehicles to pro-vide enhanced local weather forecast and predict and optimise energy consumptions in households. The project will build upon the results from past and current projects, where results from the project AutoMat, addressing services developed based on data streams from vehicles, will be used as a basis for further development aiming to extend it to integrated, cross-sectorial data streams analytics. More information is available at https://cross-cpp.eu

Funded by the Horizon 2020 Framework Programme of the European Union

Every effort has been made to ensure that all statements and information contained herein are accurate, however the Cross-CPP Project Partners accept no liability for any error or omission in the same.