

AGINFRA+ data science workshop

PLANT PHENOTYPING EXPERIMENTS ANALYSIS

1. Accessing the Food Security VRE
2. Exploring phenotyping studies
3. Retrieve data
4. Visualize data
5. Analyze data

1. ACCESSING THE FOOD SECURITY VRE

To access the VRE, open a web browser and go to the website : <https://aginfra.d4science.org/> On this AgInfra+ Gateway site, please log in with your account details.

You should see the VREs that are accessible to you. Select the Food Security VRE. This will open the website in your browser.

The screenshot shows the Food Security VRE dashboard. At the top, there is a navigation bar with links to Food Security, Administration, Members, Analytics, Semantics, Discovery, Visualization, and Plant Phenotyping installations. Below this is a 'Statistics' section titled 'Your Stats in FoodSecurity' with a user profile icon and activity counts. To the right is a 'Share updates' section with a text input field and a 'Share' button. Below that is a 'News feed' section showing a post from Panagiota Koltida dated February 07, 2:16 PM. The post text mentions updates to the portal and provides a link to documentation. To the right of the news feed is an 'About' section with the Food Security logo and a description of the VRE's purpose. At the bottom right, there is a link to 'VRE Managers and Groups'.

Name	Owner	Last modified
GetPlantHeight_Fr...	me	05 Apr 14:37 18
plant_height_viz	me	13 Feb 15:23 18
Preinstalled esrd	me	09 Jan 19:48

2. EXPLORING PHENOTYPING STUDIES AND RETRIEVING DATA

Open the Studies Exploration. Select the server PHIS_SANDBOX.

BRAPI Compliant Servers Exploration Studies Exploration Study Variables Preview Observations Data

Choose the BRAPI server to look in:

PHIS_SANDBOX

Season (Enter a year to filter)

Study Name (search with the name of the study)

Choose a specie:

all

Find

Studies Exploration

This application is meant to explore open data stored in BRAPI compliant databases.

PHIS_SANDBOX studies

Show 25 entries Search:

studyDbId	studyName	startDate	endDate	active
http://www.opensilex.org/demo/DIA2017-1	MAU17-PG	2017-05-19	2017-09-22	false
http://www.opensilex.org/demo/DMO2012-1	ZA12	2012-01-01	2012-12-31	false

studyDbId studyName startDate endDate active

Showing 1 to 2 of 2 entries Previous 1 Next

You should see at least 2 studies. In this exercise we will work with the study ZA12 data. We are interesting in plant height data. Let's check if there are plant height data on this study.

Copy the ExperimentURI and go to the tab "Preview study observations". You can see the first observations measurements of the study.

BRAPI Compliant Servers Exploration Studies Exploration Study Variables Preview Observations Data

Enter the studyDbId

http://www.opensilex.org/demo/DMO2012-1

See Data Download the table

Study Observations

Here you can retrieve 20 first observations data from one study

This study contains 506 observations

Show 25 entries Search:

observationDbId	observationLevel
http://www.opensilex.org/demo/id/data/tdohn3xthfwxbgkav2hg35ct6upj5rxu7drj2t7oalg64pnqf2e9c12181584b0c8ebe904e1694a311	http://www.opensilex.org/vocabulary/oeso#Pl
http://www.opensilex.org/demo/id/data/svmz52tcypbh2ziukpw46ouxmwzgjytcabdlts4vweedsrkunhqa2ac74a7f1034a859da09b14dde7b27e	http://www.opensilex.org/vocabulary/oeso#Pl
http://www.opensilex.org/demo/id/data/cclsgjtw4fyagqbpfd27fsmjky7jq3n6hawwdt6felt2zjqgecq6970e1d22cf8460a8bf0f996f7ca14ea	http://www.opensilex.org/vocabulary/oeso#Pl
http://www.opensilex.org/demo/id/data/zdg7nkwejkfhkdgxz34ae5q6cxzxw3d6vk22u5bxhqn2jalix2q706009e5a7cd444db4f59855129d3476	http://www.opensilex.org/vocabulary/oeso#Pl
http://www.opensilex.org/demo/id/data/77tzcpglwq5vbqeomlfr2wqfsftawb36lbqfk4s7vxf64smzczyad4617f025f8147b38ee1a6aca11973b4	http://www.opensilex.org/vocabulary/oeso#Pl
http://www.opensilex.org/demo/id/data/ae34ggmzh3mbofvt4pmjpr234t62jxhad5qrd6dr264q3vdfeaf3773a64cc0d4e0cb5f280f326d01906	http://www.opensilex.org/vocabulary/oeso#Pl

3. RETRIEVE DATA

You will use a dataminer algorithm to get all data of the ZA12 study. Go to *Analytics/Dataminer* to execute the algorithm **Brapi Get Studies Observations**. (this algorithm is inside the category "Data extraction")

Fill the different parameters as presented below (the login and password parameters can be used to access to private studies):

The screenshot shows the Dataminer web interface. On the left, the 'Operators' panel lists several operators: CHARTS (1), CURVE FITTING (2), and DATA EXTRACTION (5). Under DATA EXTRACTION, there are three operators: 'Brapi Get Studies', 'Brapi Get Studies Observations', and 'Brapi Get Variables'. The 'Brapi Get Studies Observations' operator is selected. On the right, the 'Operator' configuration panel is shown. It has a 'Tools' section with 'Remove All Operators' selected. Below this, a description states: 'Returns all observations where there are measurements for the given study and if so the given observation variables (Published by Alice Boizet (aliceboizet) on 2019/09/16 14:18 GMT)'. The 'Parameters' section includes: 'DBServerURL' set to 'PHIS_SANDBOX', 'studies' set to 'i.org/demo/DMO2012-1', 'variables' set to 'all', 'login' set to 'none', and 'password' set to 'none'.

Then *Start Computation*

You should see this execution loading bar :

The screenshot shows the Dataminer web interface with the 'Computations Execution' panel active. It shows the execution of the 'Brapi Get Studies Observations' computation. The status is 'Computation Complete'. The text indicates: 'The computation Brapi Get Studies Observations finished.' and 'The algorithm produced Multiple Results.' There is a 'Download File' button for the log of the computation.

This algorithm creates a csv file and a json file with the study data. They are stored in your workspace under the dataminer folder.

4. VISUALIZE DATA

There are 2 ways to visualize the data. You can do it by using the visualization tools or you can use Rstudio to generate graphs.

A. Visualize data in the visualization tool

The workspace is not integrated with this tool yet. So you will first need to download the csv file created by the dataminer. If you didn't do it at the end of the dataminer computation, you can go to your workspace:



button at the upper left corner. The files are stored in DataMiner/Output Data Sets.

When you have downloaded the csv file, you can go to the *visualization/Create Graphs* to create a new graph.

In Data, upload the csv file you got from the studies exploration tool.

Fill the General tab this way :

General

Data

Filters

Transformations

Documents

Label*

ZA12_plantHeight

Description

Description

Type

Scatter

Available Types

Select other chart types that will be available as options

Group By

observationUnitName

X Axis*

observationTimeStamp

X Axis Label*

date

Y Axis*

value

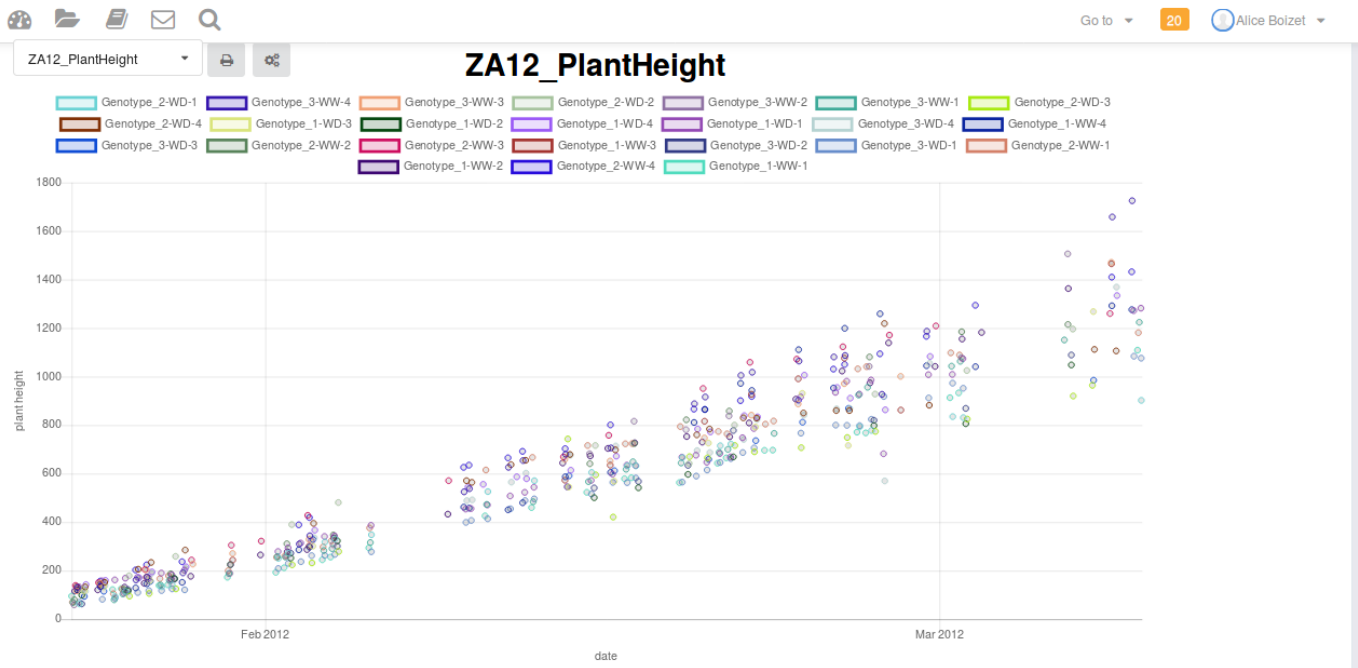
Y Axis Label*

plant height

Color

observationUnitName

You can go to *visualization/Create Graphs* and see your chart.

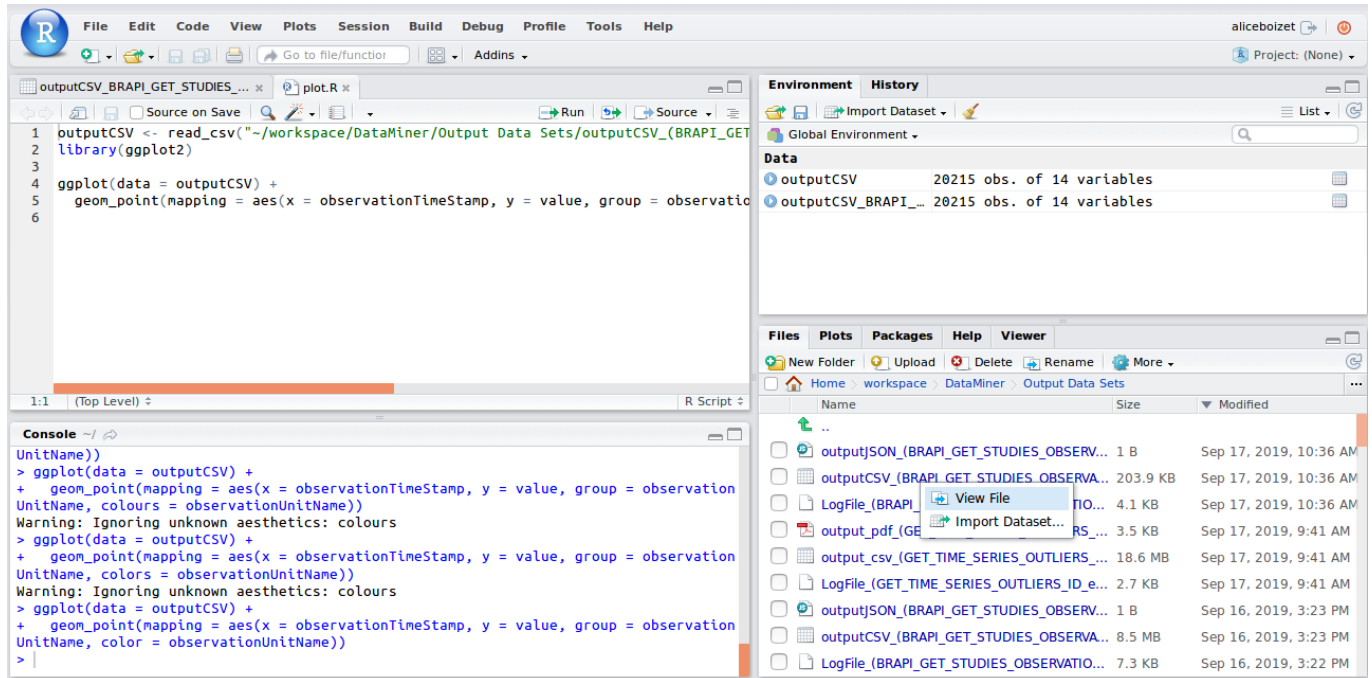


B. Visualize data in Rstudio

An other good way to visualize data is to use R. Indeed there are great packages such as ggplot2 to build charts. There are also very helpfull galleries with a lot of code examples (<https://www.r-graph-gallery.com/>)

Go to *Analytics/Rstudio* to open **Rstudio**. Find your data csv file (created by the dataminer) in the *Files* tab at the bottom right corner. It is stored in your workspace under *DataMiner/Output Data Sets* folder.

Click on the file to "import Dataset".



Now you can use the R console to visualize your data. To create a simple chart, you can use the code bellow which creates the following chart:

```
library(ggplot2)
ggplot(data = outputCSV) +
  geom_point(mapping = aes(x = observationTimeStamp, y = value, group = observationUnitName, color = observationUnitName))
```

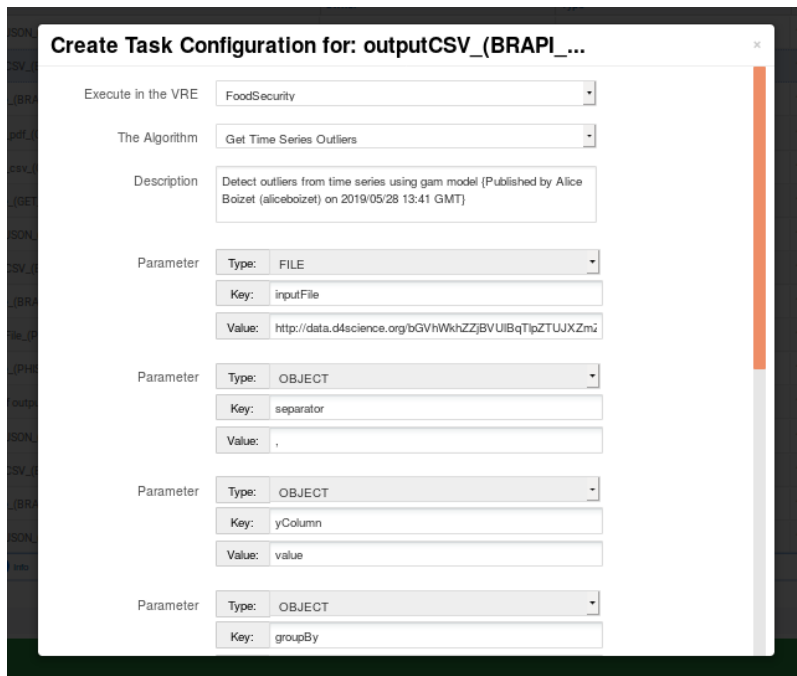


5. ANALYZE DATA

A. Reuse an algorithm to detect outliers

You would like to detect the outliers on your data. To do that, you could use Rstudio (or even Jupyter) and make a script. But you are very lucky, another VRE member has already done a similar script and he has imported it in the dataminer tool as a black box (the algorithm name is **Get Time Series Outliers**). So all you need is to go to your workspace and right click on your data csv file and select "execute DM task".

Fill the parameters as bellow:



Create Task Configuration for: outputCSV_(BRAPI_...

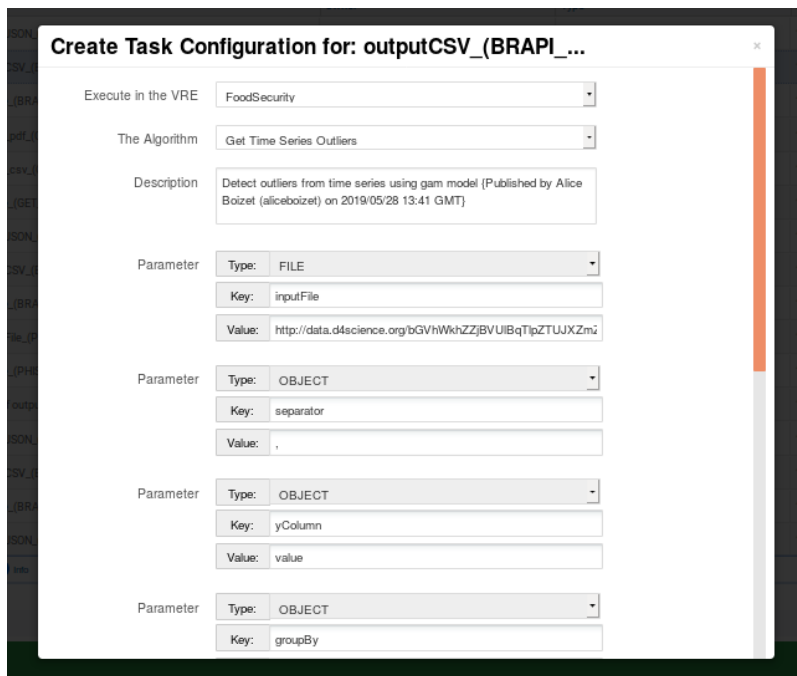
Execute in the VRE: FoodSecurity

The Algorithm: Get Time Series Outliers

Description: Detect outliers from time series using gam model (Published by Alice Boizet (aliceboizet) on 2019/05/28 13:41 GMT)

Parameter

- Type: FILE
 - Key: inputFile
 - Value: http://data.d4science.org/bGVhWkhZZjBVUIBqTlpZTUjXZm2
- Type: OBJECT
 - Key: separator
 - Value: ,
- Type: OBJECT
 - Key: yColumn
 - Value: value
- Type: OBJECT
 - Key: groupBy



Create Task Configuration for: outputCSV_(BRAPI_...

Execute in the VRE: FoodSecurity

The Algorithm: Get Time Series Outliers

Description: Detect outliers from time series using gam model (Published by Alice Boizet (aliceboizet) on 2019/05/28 13:41 GMT)

Parameter

- Type: FILE
 - Key: inputFile
 - Value: http://data.d4science.org/bGVhWkhZZjBVUIBqTlpZTUjXZm2
- Type: OBJECT
 - Key: separator
 - Value: ,
- Type: OBJECT
 - Key: yColumn
 - Value: value
- Type: OBJECT
 - Key: groupBy

Parameters description : If removeOutliers=0 the outliers will be kept and identifiable. If createPDFwithPlots=1 a pdf file with plots of each time serie will be created

Click on *Create configuration*, then run the algorithm.

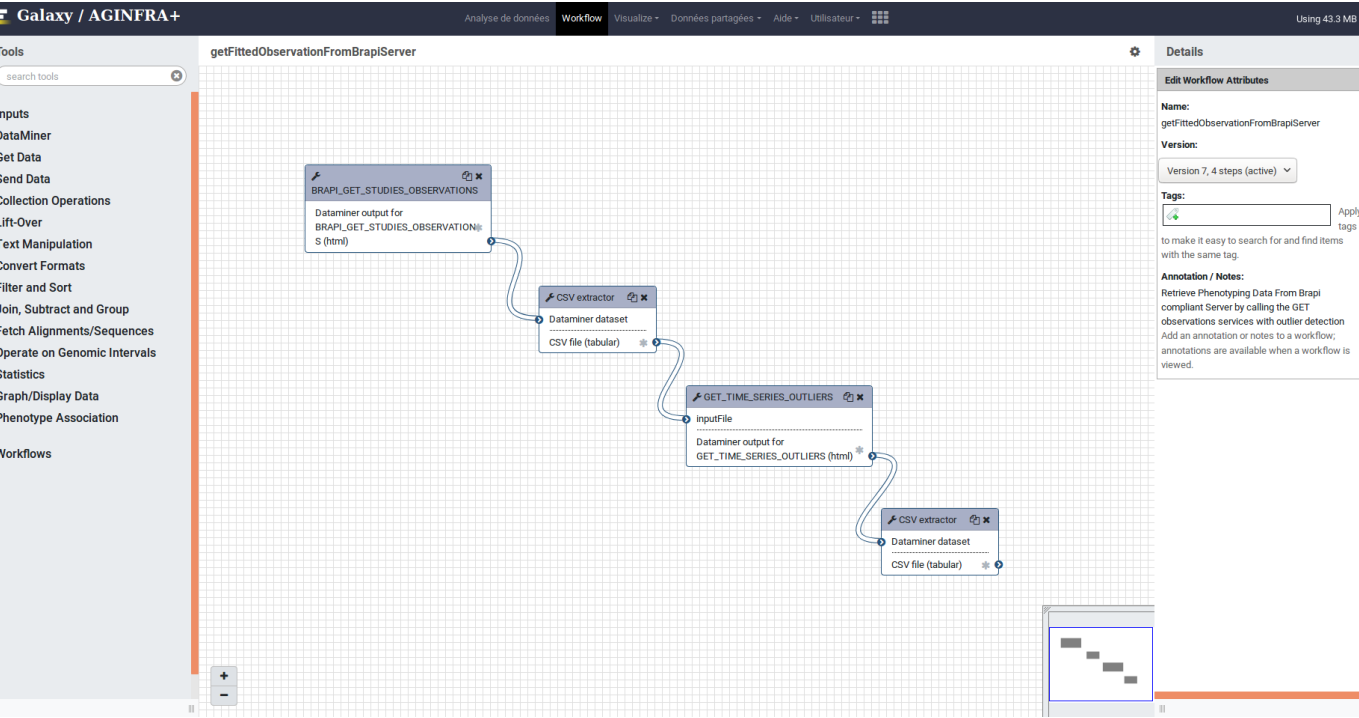
After the execution, you should find the output files in your workspace (*dataminer/output data sets folder*)


B. Make a Galaxy workflow

You would like to be able to access data from PHIS studies and detect outliers in a single process as a routine, so that you could easily do the same process on next studies. To do that, you can build a workflow in Galaxy. Galaxy can be found in the analytics tab. The dataminer algorithms are automatically transferred as galaxy tools. Build a workflow which first Retrieve observation data of the study ZA12 and then detects outliers.

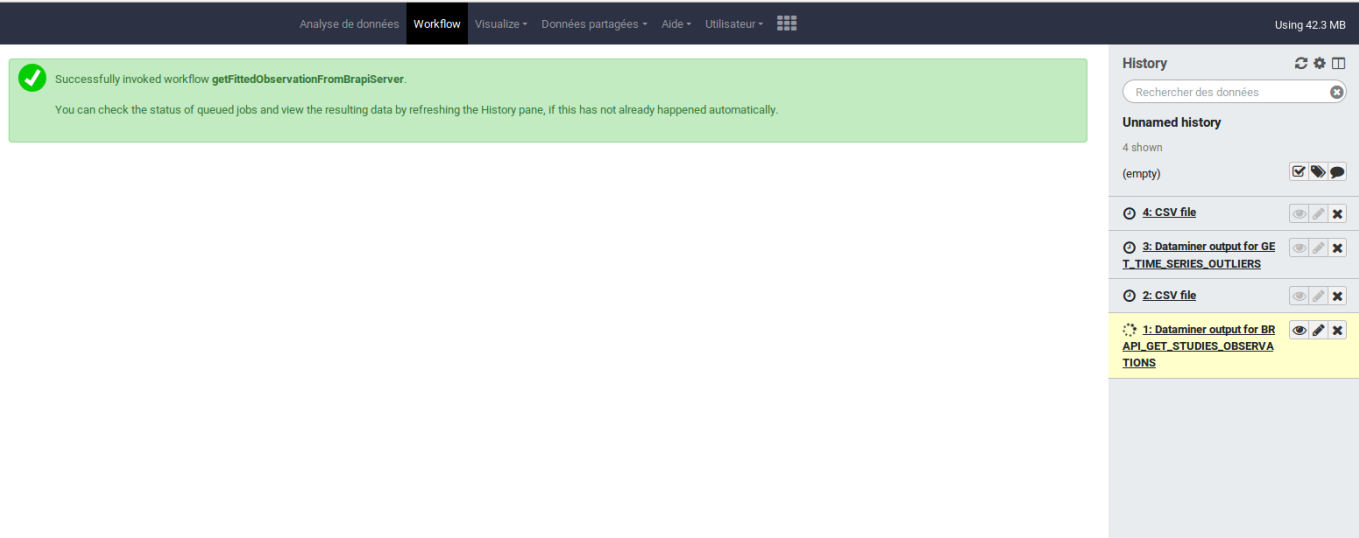
Tip : you need to use *node csv_extractor* in order to convert the dataminer tools output into csv

You should have a workflow looking like this :



Don't forget to save you workflow by clicking on 

When your workflow is ready, you can run it. You can see each step execution on the right of the screen.



After the workflow execution, you can view the outputs. You can click on the view button on the 3rd step (Dataminer output for get_time_series) where you can click on the 2 output files to download them.

The csv validator tool enables to convert the dataminer csv output into a galaxy datatable file which means that it can be reused as an input on another galaxy tool and you can visualize the data directly in Galaxy. To do that, you can click on the view button of the last step of the workflow.