A scientific guide to Gaia Astrophysical Parameters

Day 3

organized by Coordination Unit 8


ATHENS

Outlier Analysis (OA)

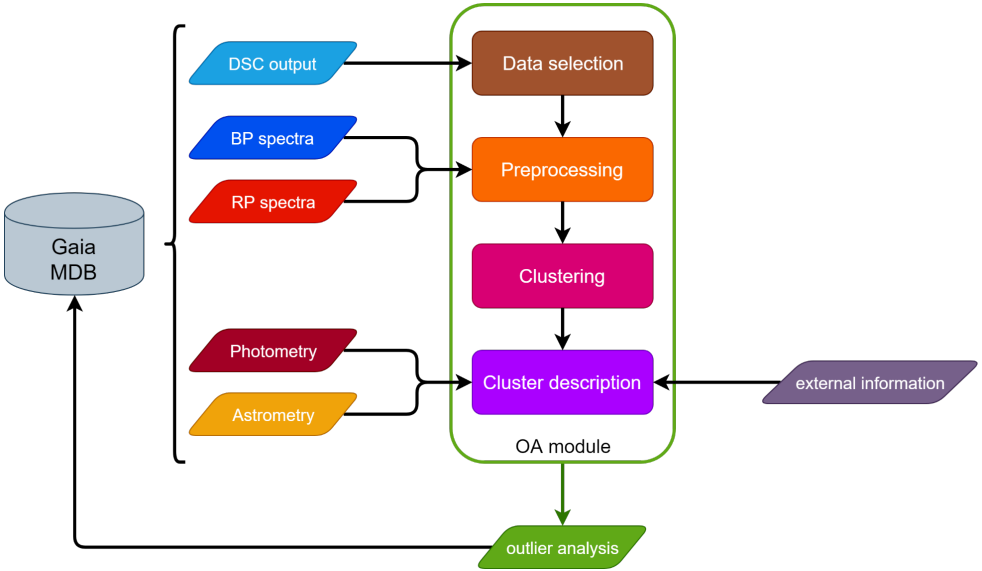
Daniel Garabato

University of A Coruna

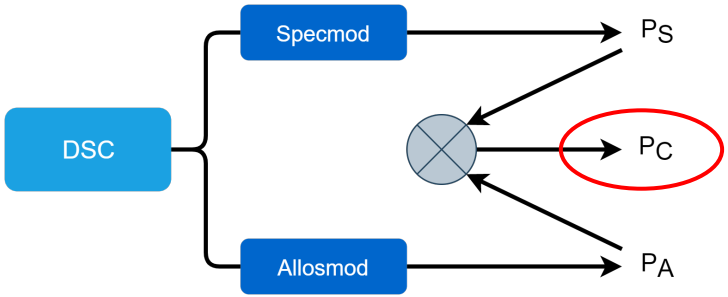
Objectives

- Process those sources with the lowest probability classification (outliers) from DSC, complementing its overall classification
  - This means dealing with weird and infrequent objects, low SNR objects, artifacts, etc.
- We pretend to group them according to their BP/RP spectra, so that similar ones should lie in the same group or a close one
  - Gaia observables are used to provide a statistical description
  - Quality indices are produced to measure the homogeneity of such groups
  - We try to give a hint on the astronomical type

Outline


DSC output
BP spectra
RP spectra
Gaia
MDB
Photometry
Astrometry
Data selection
Preprocessing
Clustering
Cluster description
OA module
external information
outlier analysis

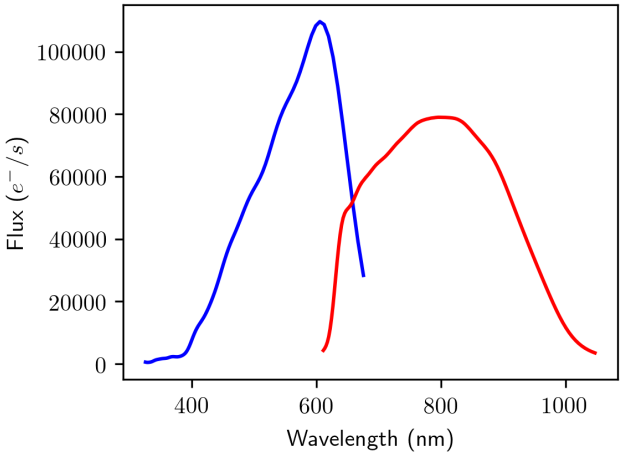
Stage 1: Data selection



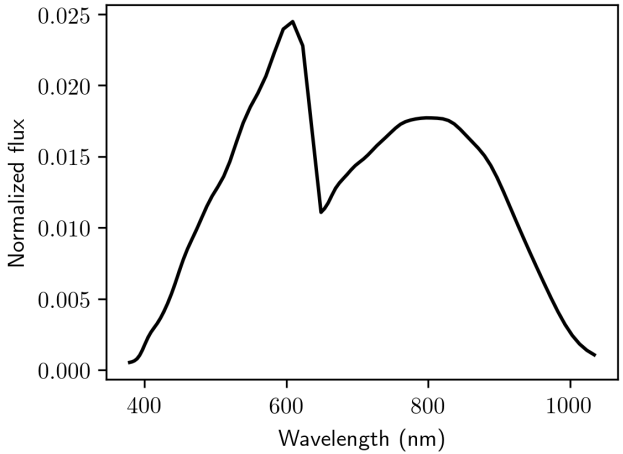
- Astrometric filters
  - Non duplicated sources
  - Converged solution
- Photometric filters
  - Sources without BP/RP spectra were discarded
  - Minimum of 5 BP/RP transits
- DSC combined probability (P_C)
  - $\max(P_C) < 0.999$

For DR3 we processed 56 million sources

Stage 2: Data preprocessing

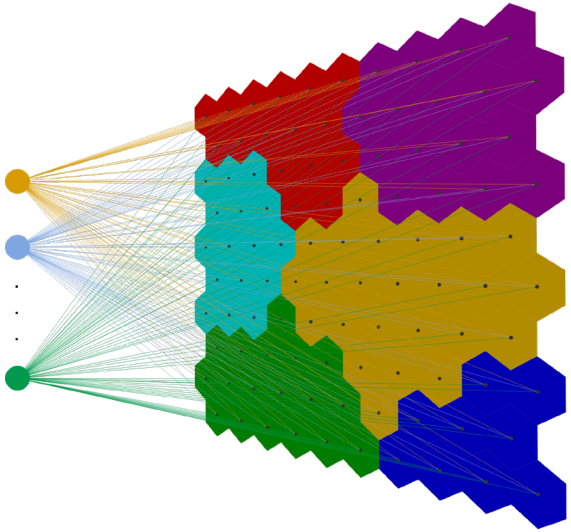


(a) Original BP/RP spectra



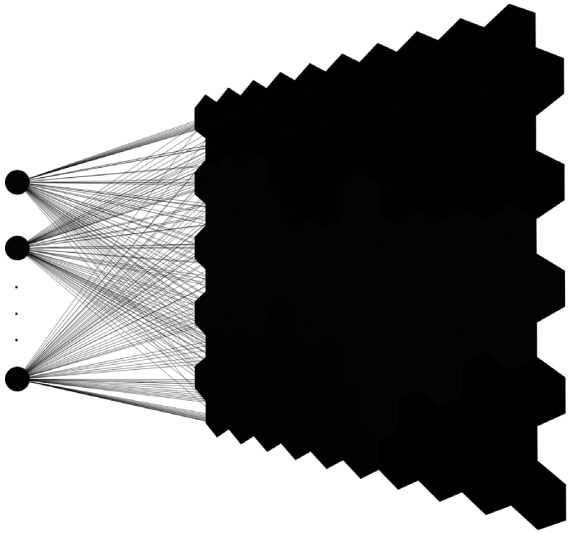
(b) Preprocessed spectra

Stage 3: Clustering



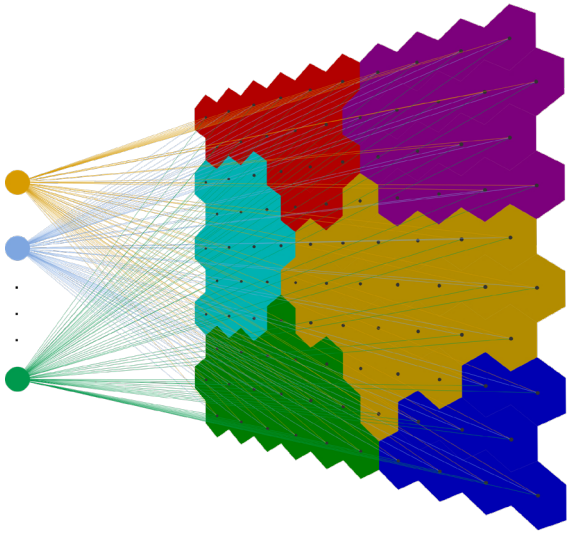
- We use **Self Organized Maps** (SOM)
  - Unsupervised and competitive ANN
    - Similarity function: **Euclidean distance**
  - Dimensionality reduction
    - 2D grid (**30x30**)
  - Topological order preservation
    - Neighborhood function: **Gaussian**
- Each neuron has a **prototype**, which is a virtual pattern that represents the sources assigned to it
- Specialized tools allow for visual exploration
  - GUASOM DR3: https://guasom.citic.udc.es

Stage 4: Description



- The actual product of a SOM is the grouping itself
  - Where does a processed source belong to?
  - Sources within a neuron have similar XP spectra

Stage 4: Description

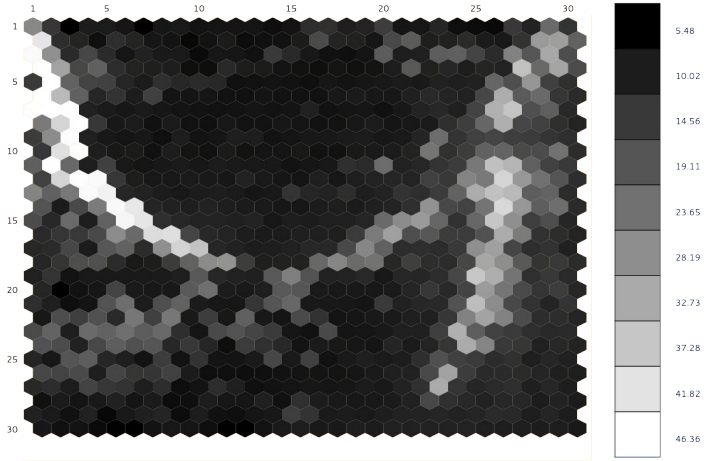


- The actual product of a SOM is the grouping itself
  - Where does a processed source belong to?
  - Sources within a neuron have similar XP spectra

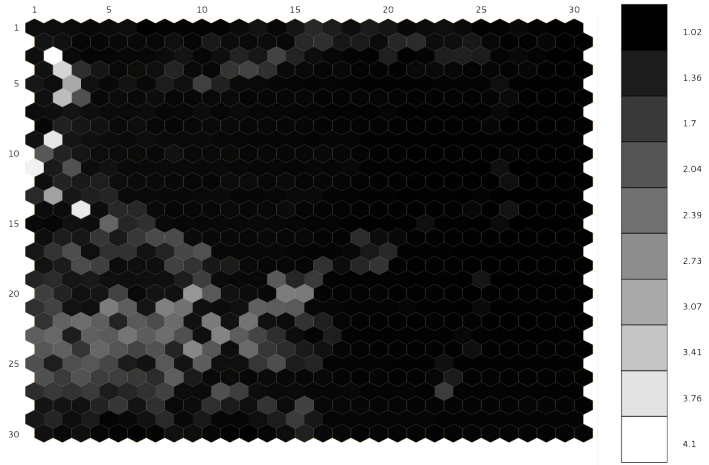
- Can we provide some **extended features**?
  - Statistical descriptors: G, G_{BP}, G_{RP}, parallax…
  - Indices to measure the homogeneity of each neuron (quality)
  - Hint on the astronomical type for best quality neurons

Stage 4a: Statistical description

Example of statistical description for Gaia observables



RP transits



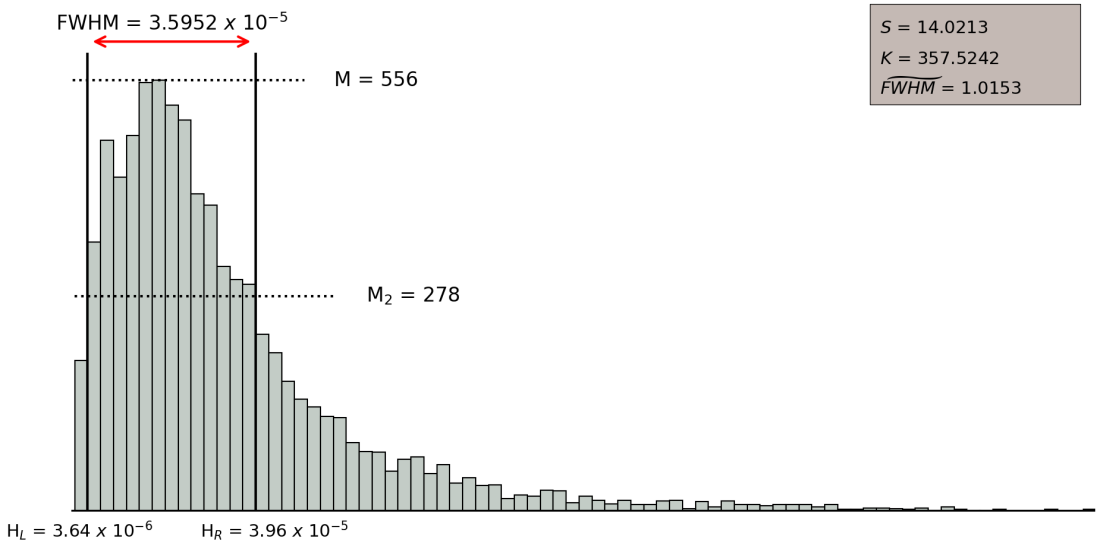
RUWE

Stage 4b: Quality assessment (I)

- Based on the intra-neuron distance distribution
  - Full Width at Half Maximum (FWHM)
  - Skewness
  - Kurtosis

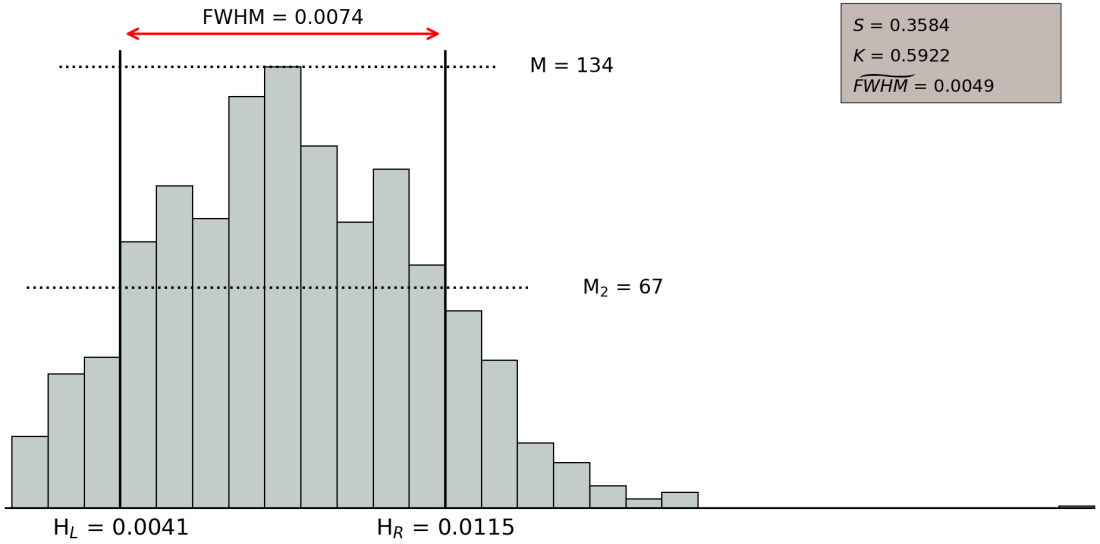
- We derived a categorical index to rank the neurons
  - Percentile values on the above quantities were used to determine its value
  - Six quality categories were established from 0 (best ones) to 6 (worst ones)

Stage 4b: Quality assessment (II)



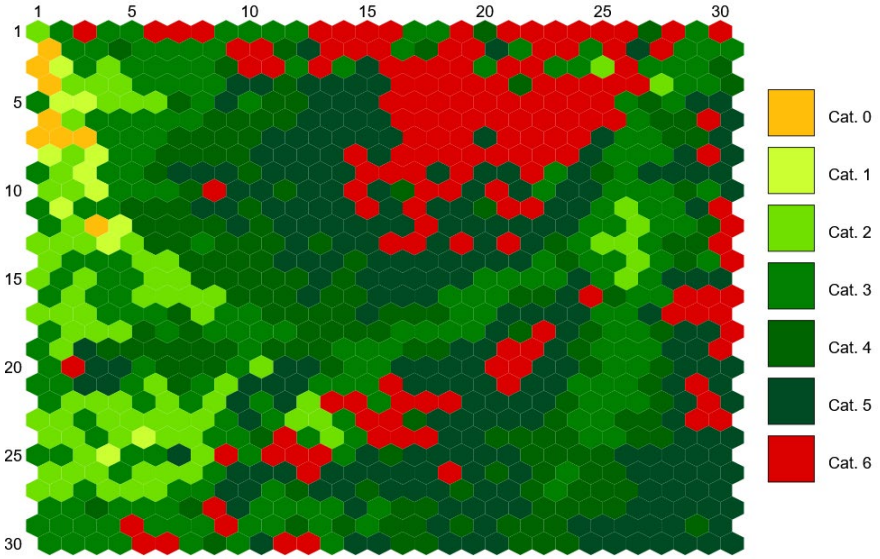
Example of a good quality neuron

Stage 4b: Quality assessment (III)

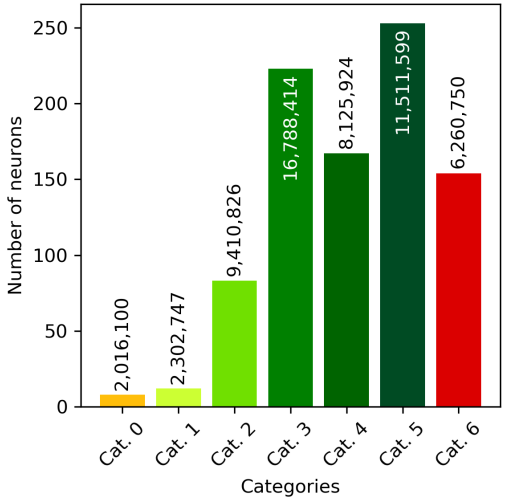


Example of a bad quality neuron

Stage 4b: Quality assessment (IV)



Categorical index



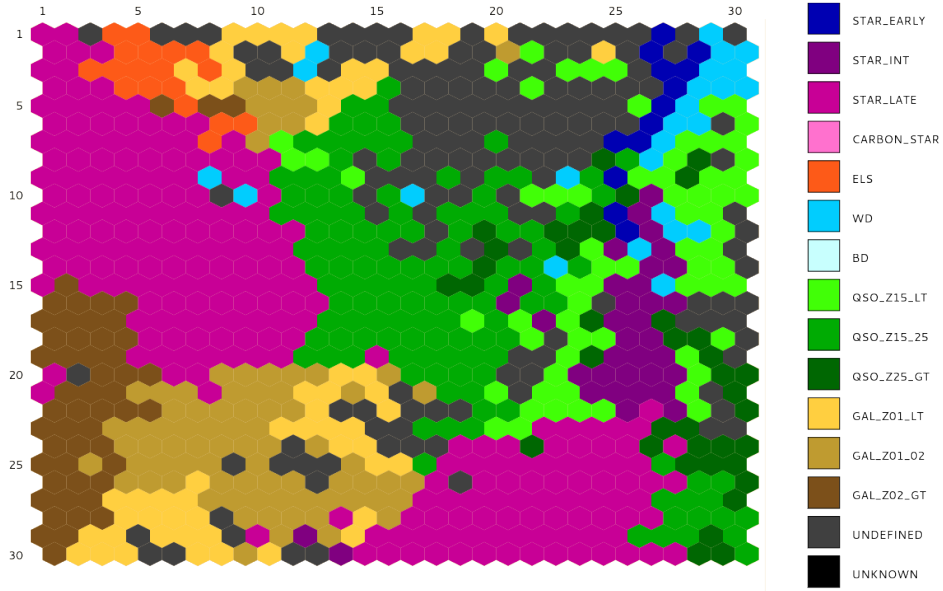
Categorical index distribution

Stage 4c: Labeling (I)

Type	Basic label	Specific label
Star	Star Early	Star O
		Star B
		Star A
	Star Intermediate	Star F
		Star G
	Star Late	Star K
		Star M
	White Dwarf	
	Emission Line Star	
	Carbon Star	
	Physical Binary	
Quasar	Quasar Z25_GT	Quasar Z45_GT
		Quasar Z35_45
		Quasar Z25_35
	Quasar Z15_25	
	Quasar Z15_LT	Quasar Z05_15
		Quasar Z05_LT
Galaxy	Galaxy Z02_GT	Galaxy Z025_GT
		Galaxy Z02_025
	Galaxy Z01_02	Galaxy Z015_02
		Galaxy Z01_015
	Galaxy Z01_LT	Galaxy Z005_01
		Galaxy Z005_LT

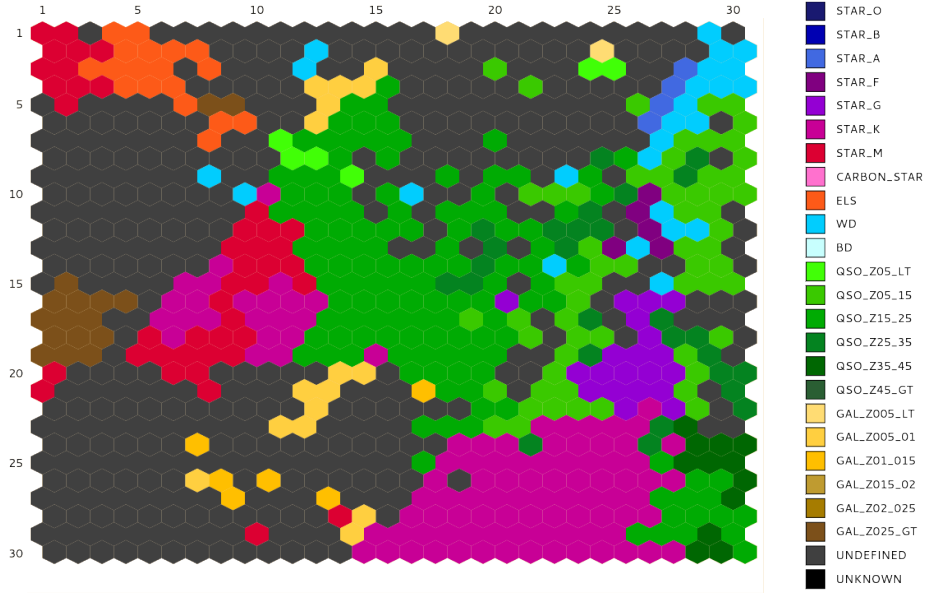
- We gather a reference set of sources for all the considered types
  - Two label types: **basic** and **specific**
  - Templates were extracted from isolated SOMs built on these sources
- The class labels are assigned using a template matching procedure
- We label just good quality neurons
  - If $QC < 6$, then non label is assigned at all
- The neuron is labeled, not the sources
  - A label for a given source can be inferred from the neuron label, if desired

Stage 4c: Labeling (II)



Basic set of labels

Stage 4c: Labeling (III)



Specific set of labels

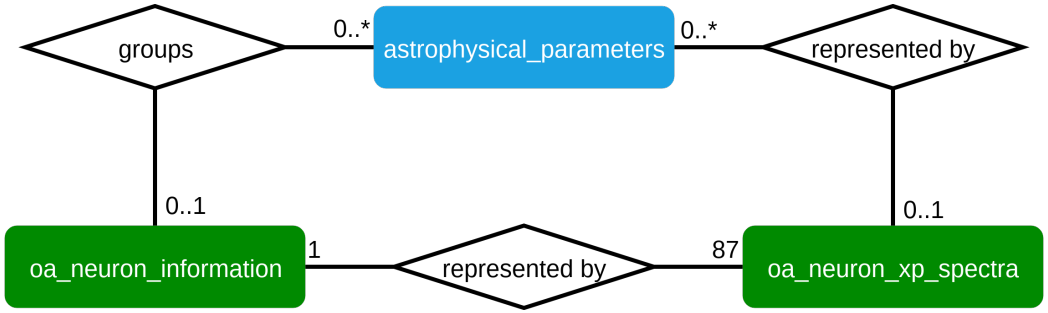
Limitations in DR3

- No outlier detector was used, we just selected those sources with lowest classification probabilities from DSC
- The reference templates used were not representative for some types
  - No templates were built for Physical Binaries
  - Some templates for quasars were not accurate enough
- Many sources were discarded due to issues in their XP spectra
  - These mostly correspond to sources with a few number of transits (<5)

Recommendations

- The class labels provided by OA are not ground truth
  - The user should rely on the given quality measurements to decide whether to trust a label or not
  - In order to infer the class label of a source from the neuron, the classification distance (or percentile) should also be taken into account

OA products in the Archive (I)



OA products in the Archive (II)

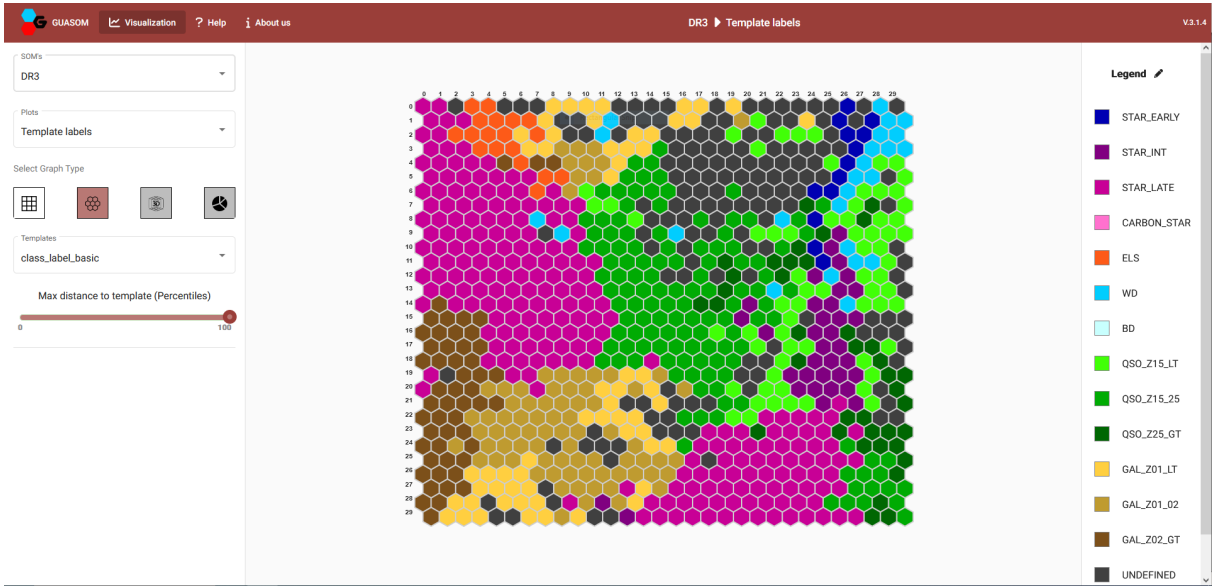
- astrophysical_parameters
  - Correspondence between the sources and the neurons
  - Distance between a source and the prototype of the neuron it belongs to
- oa_neuron_xp_spectra
  - Multidimensional data related to each neuron
    - Prototype
    - Template, if a class label was assigned
- oa_neuron_information
  - Non multidimensional data related to each neuron
    - Position within the map
    - Statistical descriptions on Gaia observables
    - Quality measurements

Querying the Gaia Archive

- To retrieve all the information the 3 tables mentioned before should be joined on ‘neuronId’
- The official documentation / papers give additional examples on how to perform common operations
- Use our visualization tool which also provides built-in ADQL queries to retrieve information for each neuron

GUASOM visualization tool

Available at: https://guasom.citic.udc.es



Update for DR4

- Outlier Analysis (OA)
  - Improve outlier detection/selection with DSC
  - Improve the preprocessing of the data
  - Update the set of templates used to label the neurons
- Unsupervised clustering analysis (UCA)
  - New CU8 module, currently under development
  - Perform a clustering on the entire survey, not just outliers using a similar procedure than in OA
    - We will use pre-trained maps on well known sets of sources to speed up the execution