Integration of GUASOM into the SPACIOUS Computing Platform

Ángel Regueiro, Marco A. Álvarez, Minia Manteiga, and Carlos Dafonte

Laboratorio Interdisciplinar de Aplicaciones de la Inteligencia Artificial (LIA2), Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Telemática, Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: a.regueiro.feal@udc.es

DOI: *https://doi.org/10.17979/spu.23.c34*

Abstract: The Science PlAtform Cloud Infrastructure for Outsize Usage Scenarios (SPACIOUS) project is developing a computational environment for astrophysical research using Big Data technologies, with the aim of enhancing the scientific exploitation of large volumes of data. To this end, data mining and Artificial Intelligence tools are being developed and integrated into the platform.
This paper presents the adaptation of GUASOM, a tool for the visualisation and analysis of Self-Organising Maps (SOM) aimed at analysing outliers from the Gaia mission. A Python application has been developed that allows these maps to be trained and loaded into GUASOM directly from SPACIOUS Jupyter Notebooks, facilitating their integration and use in the project ecosystem.

1 Introduction

The European Space Agency (ESA) missions generate large datasets, two examples are the Gaia and Euclid missions Prusti et al. (2016) Laureijs et al. (2010). To process and analyse these large volumes, Big Data and data mining technologies such as Spark are required Spark (2018), as well as machine learning libraries such as Tensorflow or PyTorch. The SPACIOUS project, Science PlAtform Cloud Infrastructure for Outsize Usage Scenarios, proposes a computing platform for astrophysics where researchers can use these techniques to work with different datasets.

SPACIOUS platform will be available to the scientific community through the Barcelona Supercomputing Centre (BSC) Martorell (2017), on the Google Cloud Platform (GCP), and with the option of installing and deploying it on the user's own infrastructure. This application will be based on the use of Jupyter Notebooks, which will allow users to interact with the data and applications available through Python, the most widely used language in the scientific field.

This paper details the adaptation of GUASOM (Gaia Utility for the Analysis of self-organizing maps), a web tool for visualizing large-scale Self-Organizing Maps (SOMs) Álvarez et al. (2022). A core component of this work involved migrating the original Java code for SOM training and data preprocessing to Python, enhancing accessibility for the scientific community. This new Python implementation is available through SPACIOUS, a platform we introduce to provide both analysis tools and tutorials. To demonstrate the complete workflow, we present a tutorial on training a SOM with data from Gaia Data Release 3 Babusiaux et al. (2023); De Angeli et al. (2023) and visualizing the result in GUASOM.

2 Methodology and Implementation

2.1 Self-organizing map

A self-organizing map (SOM) is an unsupervised neural network that reduces high-dimensional data to a low-dimensional map, typically 2D, preserving topological relationships so that similar data points are mapped to nearby locations on the map Kohonen (2012). The SOM's architecture, depicted in Figure 1, consists of cells (or neurons), each representing a prototype vector in the original data space. During classification, an input vector is assigned to the cell with the most similar prototype. In this work, we employ a numerical SOM where the input data are spectra from celestial objects.

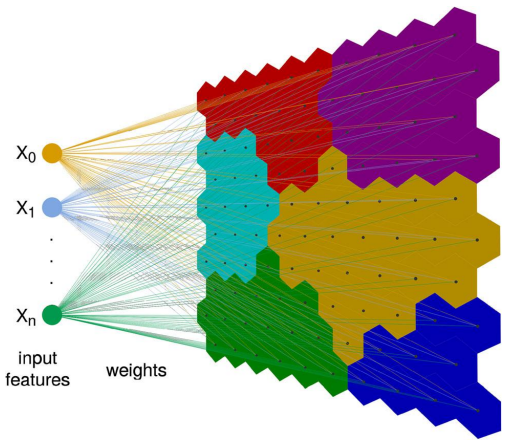

Figure 1: Self-organizing map architecture.

To train an SOM, we must first initialise its prototypes, specifying a map size NxM. These can be initialised randomly or using prior information from the input data. One method that improves the efficiency and effectiveness of the algorithm is to calculate the average of the input data and add random values to it in order to initialise each prototype differently. In this way, the initial values of the prototypes are already similar to the spectra being worked with, thus reducing the number of iterations needed to complete the training.

The training process is iterative. For each element in the dataset, we find its Best Matching Unit (BMU), which is the most similar prototype neuron according to a distance metric. A neighborhood function then determines how the BMU and its adjacent neurons should be adjusted to be more like the input element. After a full pass over the entire dataset, the positions of all prototypes are updated collectively. This cycle is repeated for a maximum number of iterations or until an early-stopping condition is met.

2.2 Python frameworks

Before using GUASOM, it is necessary to obtain a map and its data in a specific format. This was done using two Java apps: *clustering_toolkit*, which trained the network, and *preprocess*, which converted the result into GUASOM input for visualisation. These applications and GUASOM communicated using serialised Java files with the extension .ser. For compatibility with the SPACIOUS platform, the two Java applications have been migrated to Python, as shown in Figure 2.

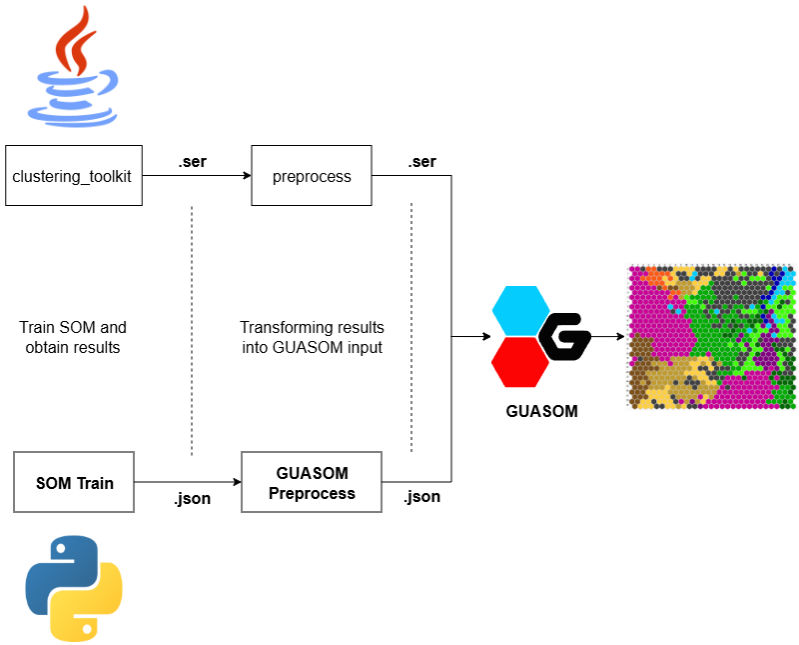

Figure 2: Workflow for training and pre-processing a SOM map, to finally visualise it with GUASOM.

The new Python applications for training and preprocessing are *SOM Train* and *GUASOM Preprocess*, respectively. In addition, the data format for communication between applications has been changed to JavaScript Object Notation (JSON) files. This language-agnostic format allows us to perform training and preprocessing in Python. To make this data compatible with GUASOM, we modified its Java backend to directly read the JSON files.

3 Validation and Results

To validate the new implementation of the SOM algorithm and its integration with GUASOM, a reference dataset with Gaia spectra will be used. This set is called the Validation Source Table (VST), which allows for agile validation runs Garabato Míguez (2020). From this dataset, we will work with the stars, for which we will have their BP/RP spectrum and their type label. Table 1 shows the basic subtypes of stars that will be used, dividing some of them into specific subtypes. On the other hand, Figure 3 shows a spectrum for each possible subtype.

Table 1: Validation Source Table (VST) dataset with 163919 samples

Basic subtype	Specific subtype	Number of samples
Star Early	Star O	1077
	Star B	11434
	Star A	30392
Star Intermediate	Star F	32467
	Star G	30011
Star Late	Star K	40219
	Star M	5067
White Dwarf		2887
Emission Line Star		7428
Carbon Star		1393
Physical Binary		1544

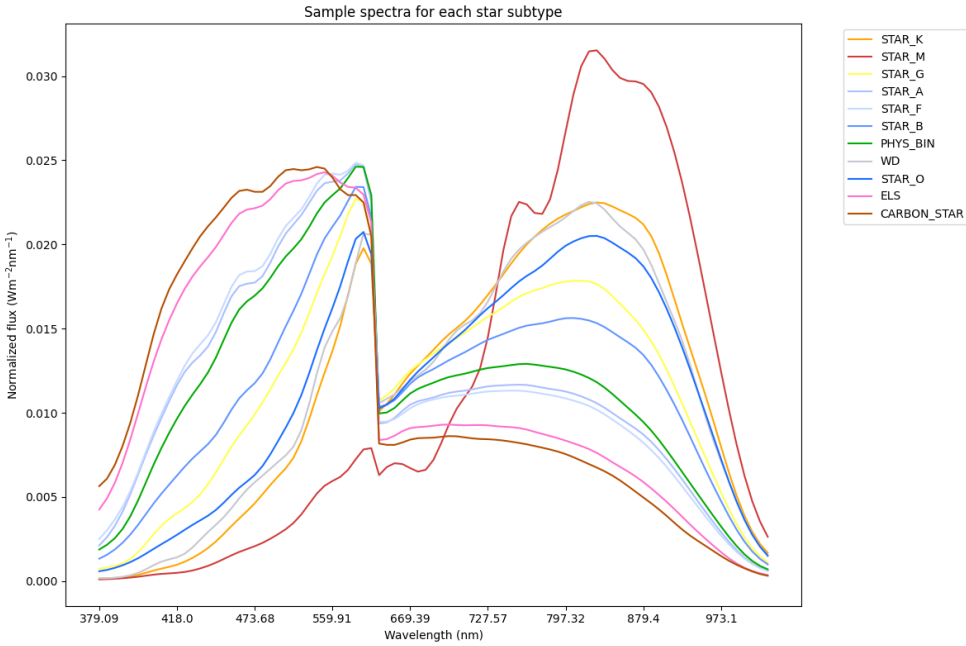

Figure 3: Example spectrum for each star subtype in the Validation Source Table (VST) dataset.

A 16 x 16 SOM, a total of 256 neurons, was used to group the 163919 spectra. Squared Euclidean distance has been selected as the distance metric for calculating the BMU. The Gaussian function has been selected for the neighbourhood function with an initial radius of 9 and a decay factor of 16.0. The decay factor governs the rate at which this radius exponentially shrinks as training progresses, gradually localizing the influence of input spectra to smaller regions of the map. The maximum number of iterations was set to 45. This decision was based on an analysis of the map's convergence, where both the Quantization Error, which reached a stable

value of 0.0018, and the Topological Error, which fell to 0.1318, indicated that further training would not significantly improve the map's quality.

Once the map has been trained using *SOM Train* in Python, it must be converted to the GUASOM input format and the labels for each neuron must be obtained. For this purpose, the *GUASOM Preprocess* programme in Python is used. To label each neuron in the SOM, all elements in the dataset classified in them will be taken into account, assigning the most common label. This will allow us to view the map labelled in GUASOM in Figure 4. This map shows a good distribution of classes, with good topology, i.e. neurons of similar classes are grouped together. No neurons with a majority of Physical Binary or Carbon Star are found, which is normal, since, as can be seen in Table 1, there are very few samples.

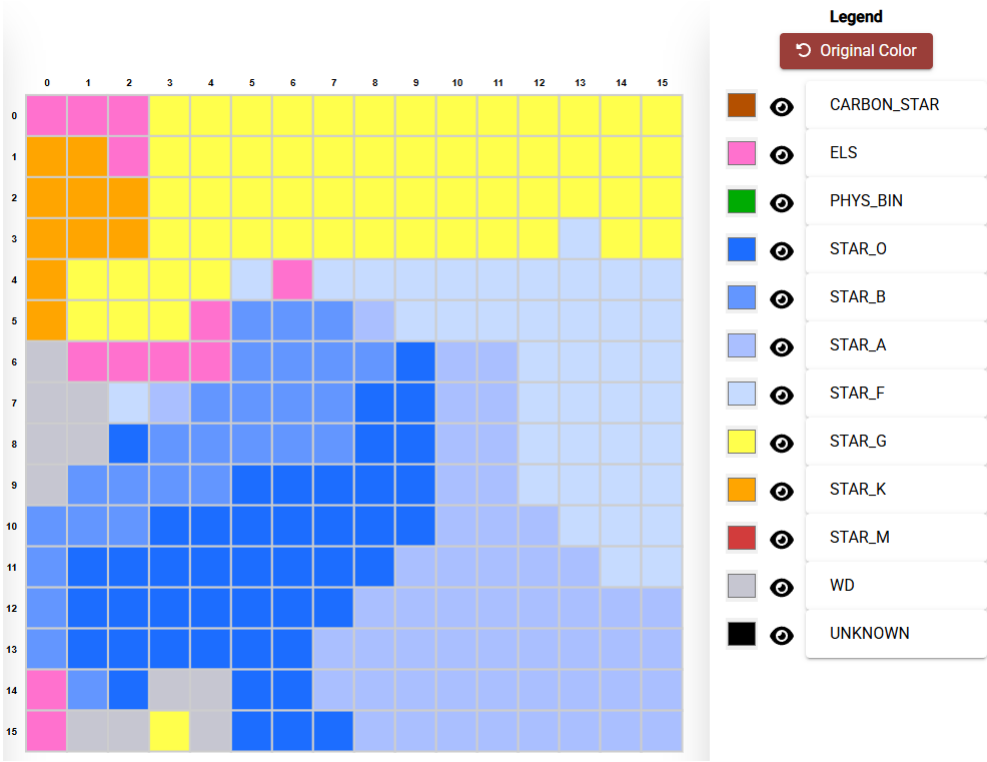

Figure 4: Map trained with VST visualized in GUASOM. Each neuron is colored according to the most common stellar subtype.

GUASOM contains different visualisations that we can explore. For example, we can view the *Hits*, which indicate the number of elements in the Dataset grouped in each neuron. In addition to using 2D visualisations, other 3D visualisations are also available. Figure 5 shows this visualisation, called *Catalogue labels + Hits*, where the height of each neuron indicates the number of elements grouped within it.

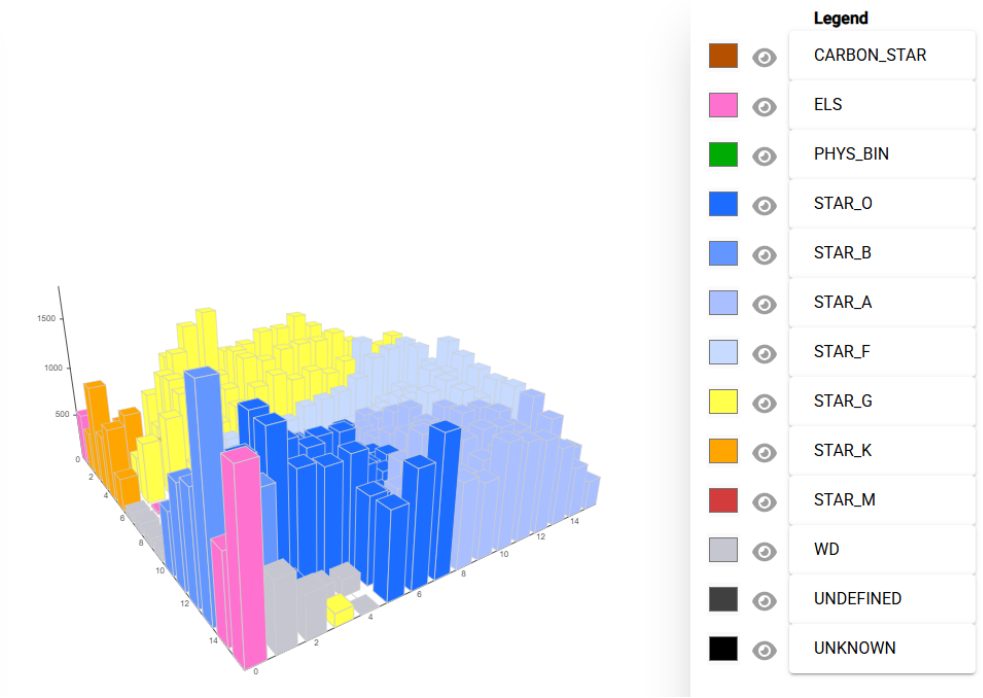

Figure 5: A 3D visualization of the same VST-trained map. The color represents the dominant stellar subtype, while the height of each column corresponds to the number of hits (i.e., the number of spectra) assigned to that neuron, highlighting densely populated regions of the map.

4 Conclusions

This paper details the successful migration of the Self-Organizing Map (SOM) training and preprocessing toolkit from Java to Python, ensuring its seamless integration into the SPACIOUS (Science PlAtform Cloud Infrastructure for Outsize Usage Scenarios) platform. The primary goal of this work was to modernize the existing workflow, making it more accessible and compatible with the preferred technologies of the scientific community.

The effectiveness of the new Python implementation was validated using a substantial dataset of 163919 stellar spectra from the Gaia Validation Source Table (VST). The resulting 16x16 SOM successfully clustered the spectra according to their physical properties, effectively separating different star types (Early, Intermediate, Late, White Dwarfs, etc.) across the map. The visualizations in GUASOM confirm that the topological preservation characteristic of SOMs was achieved, grouping similar stellar spectra into adjacent neurons.

By integrating these tools into the SPACIOUS platform via Jupyter Notebooks, we have significantly lowered the barrier to entry for researchers wishing to apply unsupervised machine learning techniques to large-scale astrophysical datasets. The availability of this workflow, complemented by tutorials, empowers the scientific community to perform complex data mining tasks efficiently. Future work will focus on expanding the tutorials, integrating additional datasets, and further optimizing the tools for distributed computing environments to handle even larger data volumes from upcoming surveys.

Acknowledgements

The Horizon Europe Programme is funding this research through the [HORIZON-CL4-2023-SPACE-01-71] SPACIOUS project, Grant Agreement no. 101135205. We also acknowledge support from the Xunta de Galicia and the European Union (FEDER Galicia 2021-2027 Program) Ref. ED431B 2024/21, CITIC ED431G 2023/01.

Bibliography

M. A. Álvarez, C. Dafonte, M. Manteiga, D. Garabato, and R. Santoveña. Guasom: an adaptive visualization tool for unsupervised clustering in spectrophotometric astronomical surveys. *Neural Computing and Applications*, 34(3):1993–2006, 2022.

C. Babusiaux, C. Fabricius, S. Khanna, T. Muraveva, C. Reylé, F. Spoto, A. Vallenari, X. Luri, F. Arenou, M. Alvarez, et al. Gaia data release 3-catalogue validation. *Astronomy & Astrophysics*, 674:A32, 2023.

F. De Angeli, M. Weiler, P. Montegriffo, D. W. Evans, M. Riello, R. Andrae, J. Carrasco, G. Busso, P. Burgess, C. Cacciari, et al. Gaia data release 3-processing and validation of bp/rp low-resolution spectral data. *Astronomy & Astrophysics*, 674:A2, 2023.

D. Garabato Míguez. *Análisis no supervisado de observaciones atípicas en la misión espacial Gaia; optimización mediante procesamiento distribuido e integración en Apsis*. PhD thesis, Universidade da Coruña, 2020.

T. Kohonen. *Self-organizing maps*, volume 30. Springer Science & Business Media, 2012.

R. J. Laureijs, L. Duvet, I. E. Sanz, P. Gondoin, D. H. Lumb, T. Oosterbroek, and G. S. Criado. The euclid mission. In *Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave*, volume 7731, pages 453–458. SPIE, 2010.

J. M. Martorell. Barcelona supercomputing center: Science accelerator and producer of innovation. *Contributions to science*, 12(1):5–11, 2017.

T. Prusti, J. De Bruijne, A. G. Brown, A. Vallenari, C. Babusiaux, C. Bailer-Jones, U. Bastian, M. Biermann, D. W. Evans, L. Eyer, et al. The gaia mission. *Astronomy & astrophysics*, 595:A1, 2016.

A. Spark. Apache spark. *Retrieved January*, 17(1):2018, 2018.