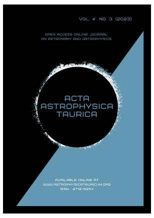
Open Access Online Journal on Astronomy and Astrophysics

Acta Astrophysica Taurica

www.astrophysicataurica.org



Acta Astrophys. Tau. 4(3), 5–10 (2023)

doi:10.34898/aat.vol4.iss3.pp5-10

Cluster analysis on the example of blazars from the Roma-BZCAT catalog

D.O. Kudryavtsev[1], Yu.V. Sotnikova[1], V.A. Stolyarov[1,2], T.V. Mufakharov[1,3], V.V. Vlasyuk[1], Yu.V. Cherepkova[1]

[1] Special Astrophysical Observatory of the Russian Academy of Sciences, Nizhny Arkhyz 369167, Russia
e-mail: dkudr@sao.ru

[2] Astrophysics Group, Cavendish Laboratory, University of Cambridge, Cambridge CB3 0HE, UK

[3] Kazan Federal University, Kazan 420008, Russia

Received 28 September 2023

ABSTRACT

Based on the collected multiwavelength data, we perform a cluster analysis for the blazars of the Roma-BZCAT catalog, selecting groups of blazars with similar properties. Using machine learning methods, we constructed an independent classification of the blazars and compared it with the known Roma-BZCAT classification. The clustering algorithms divide both BL Lac-type objects and flat-spectrum radio quasars (FSRQs) into two subclasses along with a separate group of mixed BL Lacs and FSRQs. The clustering did not reveal difference between the BL Lacs and galaxy-dominated BL Lacs, unlike in the Roma-BZCAT classification.

Key words: data analysis, active galactic nuclei, blazars, flat-spectrum radio quasars

1 Introduction

Development of astronomical instrumentation and data gathering techniques has been providing nowadays an overwhelming amount of observational data of different kinds: from large samples of point estimates for hundreds of characteristics to complex data such as images, spectra, or time series. Fortunately, along with this growing amount of information we witness the growing computational power and the development of sophisticated data analysis techniques based on machine learning methods.

In this paper we touch on a relatively simple case of investigation of multiparametric tabular data using cluster analysis, or clustering, a technique of unsupervised learning aimed at obtaining the most general properties of a dataset of some objects described by their characteristics. In our case we analyze the Roma-BZCAT catalog (Massaro et al., 2015) of blazars, complemented with observed data from other point-source catalogs.

We divide the blazars into five classes based on their multifrequency properties, show some preliminary results, and compare our classes with the Roma-BZCAT classification. Two clustering algorithms are implemented and compared. We also briefly consider the problem of imputing the missing measurements in the dataset.

The code with the implementation of some stages of this project is available in Jupiter notebooks on GitHub.[1] The web page also contains links to the used data sources, described further in Sect. 3, and a number of Python scripts used to collect the data.

2 The technique of cluster analysis

The basic idea of cluster analysis is that similar objects have similar numerical characteristics, which allows one to combine them in groups (clusters) based on some measure of their similarity and then investigate their statistical properties, which could give some insights into their nature.

Having the objects and their characteristics (features), we thus determine a feature space of dimension M, where M is the number of features. Each object is described in this space by a vector $\mathbf{x}_n$, where indices n range in a closed interval from 1 to N, $n = [1, N]$, and N is the number of objects. In our case the measure of similarity between any two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ could be the Euclidean distance:

$$d = |\mathbf{x}_i - \mathbf{x}_j|. \tag{1}$$

Now, if we describe our dataset as a set (or matrix) $\{X\} \in \mathbb{R}$ of dimension $N \times M$ and assign the clusters as a set $\{Y\} \in \mathbb{Z}$ of cardinality K, where K is the number of clusters, $\{Y\} = [1, K]$, then the solution of the clustering problem is finding an algorithmic function a: $\{X\} \rightarrow \{Y\}$ that assigns a singular label y_k, $k = [1, K]$ to each object $\mathbf{x}_n$, $n = [1, N]$ in such a way that the objects with similar properties (closer distances d) correspond to the same label (cluster).

Ideally, these clusters of similar objects could form localized groups in the feature space, but this is not always true, and especially in our case of blazars, which are by themselves

[1] https://github.com/DKudryavtsev/BZCAT-Clustering

are a very particular group of AGNs distinctive by a specific peculiarity: the orientation of the jet toward the observer.

In general, the clustering algorithm builds an independent unsupervised classification that is based almost solely on the multiple properties of the objects under consideration. The method, nevertheless, have its own hyperparameters (the parameters that are set by the researcher rather than learned by the model from data):

- the feature space, which determines the characteristics relevant for the scientific scope of the problem;
- the algorithm implemented to find the clusters, usually several algorithms are selected based on the data distribution and then evaluated from internal clustering metrics (e.g., the silhouette coefficient, Calinski–Harabasz index, and Davies–Bouldin index; Caliński, Harabasz, 1974; Davies, Bouldin, 1979; Rousseeuw, 1987);
- the number of clusters, a trade-off between uniformity and individuality, which to a certain degree can also be evaluated from metrics (look for, e.g., the "elbow" method).

3 The dataset and feature space

The original dataset, which is going to be published in a separate paper currently in preparation, is based on the Roma-BZCAT catalog (Massaro et al., 2015) and complemented with multifrequency data from various other catalogs. The total number of collected characteristics is over one hundred.

For the feature space we tried to take the maximum number of characteristics related to the physical properties of the objects. Leaving the detailed description of feature selection and transformations, let us go directly to the result. For modeling we chose:

- flux densities:
  - radio range represented by the Roma-BZCAT 1.4 GHz data (NVSS, FIRST) and by the CATS database[2] (Verkhodanov et al., 1997, 2005) and BLcat RATAN-600[3] (Mingaliev et al., 2014; Sotnikova et al., 2022) measurements at 4.7 GHz;
  - IR range from WISE[4] W1–W4;
  - optical range from Pan-STARRS[5] *grizy*;
  - UV range from GALEX[6] FUV and NUV;
  - X-rays from the Roma-BZCAT 0.1–2.4 keV data (ROSAT, Swift-XRT);
- a set of "hardnesses:"
  - radio-to-optical spectral index from Roma-BZCAT;
  - radio–IR, $\log_{10}(\nu F_{1.4\mathrm{GHz}} / \nu F_{\mathrm{W2}})$;
  - radio–UV, $\log_{10}(\nu F_{1.4\mathrm{GHz}} / \nu F_{\mathrm{NUV}})$;
  - radio–X, $\log_{10}(\nu F_{1.4\mathrm{GHz}} / \nu F_{\mathrm{X}})$;
  - IR–opt, $\log_{10}(\nu F_{\mathrm{W2}} / \nu F_{i})$;
  - IR–UV, $\log_{10}(\nu F_{\mathrm{W2}} / \nu F_{\mathrm{NUV}})$;
  - IR–X, $\log_{10}(\nu F_{\mathrm{W2}} / \nu F_{\mathrm{X}})$;
- monochromatic (4.7 GHz) radio luminosity corrected for the redshift z (transformed to the rest frame):
  $$L_{4.7} = 4\pi D_L^2 S_{4.7} (1+z)^{-\alpha-1}, \tag{2}$$
  where D_L is the luminosity distance, $S_{4.7}$ is the flux density, and α is the radio spectral index defined as
  $$\alpha = \frac{\log F_2 - \log F_1}{\log \nu_2 - \log \nu_1}, \tag{3}$$
  where F_1 and F_2 are the flux densities at frequencies ν_1 and ν_2, respectively. Here the frequencies were 4.7 and 11.2 GHz or, where the measurements were absent, 4.7 and 7.7 GHz;
- frequency of the synchrotron peak determined by polynomial fitting[7] of the spectral energy distributions (SEDs) derived from the SED Builder[8] tool of the Italian Space Agency Space Science Data Center;
- spectrum slopes in the WISE and Pan-STARRS ranges calculated from linear regression; these values replace in our case the optical colors, providing rougher but more robust estimates of the SED;
- comoving distance calculated from the ΛCDM cosmology with the Planck parameters (Planck Collaboration et al., 2020).

Among these 25 characteristics, 13 features are the flux densities in different ranges of the electromagnetic spectrum: from radio to X-rays. The ratios between different kinds of electromagnetic radiation is already described in our dataset by the hardness parameters, so these flux densities are redundant and even adverse for the clustering algorithms due to the so-called "curse of dimensionality". As a first step of dimensionality reduction, we convolved them into two metafeatures characterizing flux densities in the radio and shorter wavelengths. The choice of these two metafeatures is based on the simple core–jet model of AGNs. In this model the radio emission is unambiguously related to the synchrotron radiation from the jet, while emission in other electromagnetic ranges can be generated by both the core regions and the jet (via its synchrotron radiation and inverse Compton scattering). The model is confirmed by the correlations observed in our data: while the radio emission is not correlated with the emission in other ranges (Kendall's $\tau \leq 0.1$), we can see weak correlations for the X-rays relative to the UV, optical, and IR ranges ($\tau = 0.1$–0.3) and notable correlations between flux densities in the UV, optical, and IR ranges ($\tau = 0.3$–0.7). Obviously, the correlations between flux densities of the same kind (e.g., between the optical *grizy* bands) are very strong ($\tau > 0.7$).

The two metafeatures were obtained as the PCA[9] first principal components for each of the ranges (radio and shorter wavelengths). After that, the resulting model dataset used for the clustering comprised 14 features, forming correspondingly a 14D feature space.

We should notice that this feature space is subjected to some effects that are negative for interpretation of the results that could be obtained from the clustering. In the first place, all selection effects are preserved, and almost all characteristics are dependent on the distance to the blazars (or

[2] https://www.sao.ru/cats/
[3] https://www.sao.ru/blcat/
[4] https://irsa.ipac.caltech.edu
[5] https://outerspace.stsci.edu/display/PANSTARRS/
[6] https://archive.stsci.edu/missions-and-data/galex
[7] https://github.com/DKudryavtsev/BZCAT-SED-Viewer
[8] https://tools.ssdc.asi.it/SED/
[9] Principal component analysis

redshift z). For example, the redshift-corrected radio luminosity nevertheless shows strong dependence on z due to the Malmquist effect. Even the flux density ratios are dependent on the distance because of the cosmological rest frame drift, which could not be corrected due to the absence of an accurate SED model for each of the blazars. At the same time, these effects can be considered useful for the clustering because they could potentially help separate classes that naturally demonstrate different distance distributions (because of the selection in the data or not). It is for this reason that we leave the comoving distance and raw flux densities as model characteristics. This dependence on distance must be kept in mind during further analysis of the obtained groups. The second nuisance is the fact that blazars are variable, therefore in some cases the characteristics may be measured in different states of blazar activity (active/quiescent). This restricts our results to only the groups' statistical properties. Finally, the BZCAT catalog is not a complete flux-limited list of blazars, but it is the largest collection of the well-known blazars selected using experimental data from different surveys in a wide wavelength range from radio to gamma-rays, hence containing a large amount of various data crucial for performing our task. The incompleteness of the blazar sample does not affect the main objective of this study, tests of clustering algorithms, and the analysis of the observed differences; however, it could influence the distribution of blazars within the clusters, i.e., the population of certain groups (borders of the clusters in the feature space) may change for a more complete sample.

4 Clustering

There are a lot of approaches to the clustering problem. As for the tabular data, many of them are implemented in the Python machine learning library `scikit-learn` (Pedregosa et al., 2011). The choice of the particular algorithm is dependent on data distribution and found by trial and error. We tested several algorithms, evaluating the results by internal clustering metrics, and found that a combination of PCA dimensionality reduction with k-means (Arthur, Vassilvitskii, 2007) provided the best result in terms of cluster separation. The PCA+k-means results have been controlled using the second, non-linear, approach: Kohonen's Self-Organizing Maps (SOMs, Kohonen, 2001), which are based on a neural network with competitive learning. In this paper we used the `Somoclu` implementation of SOMs (Wittek et al., 2017).

In the PCA+k-means approach, we first transform the data into a new coordinate system where the principal components (directions representing the maximum variance in the feature space) constitute an orthonormal basis. After that, we can

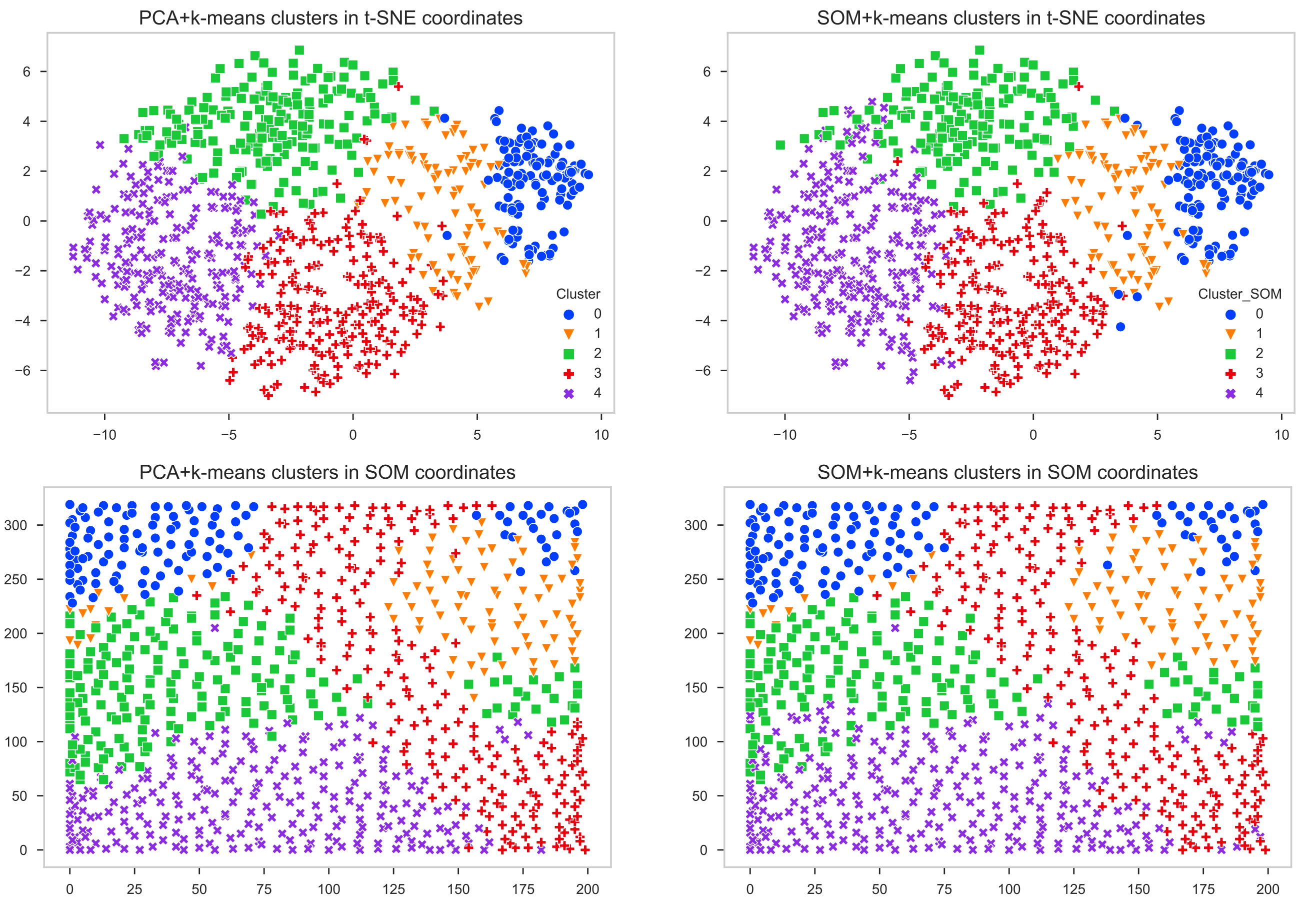


Fig. 1. Comparison of the PCA+k-means (left) and SOM (right) clustering. The coordinates are the conditional 2D t-SNE coordinates (top) and the output map of the SOM method (bottom). The points are the blazars colorized according to their cluster membership.

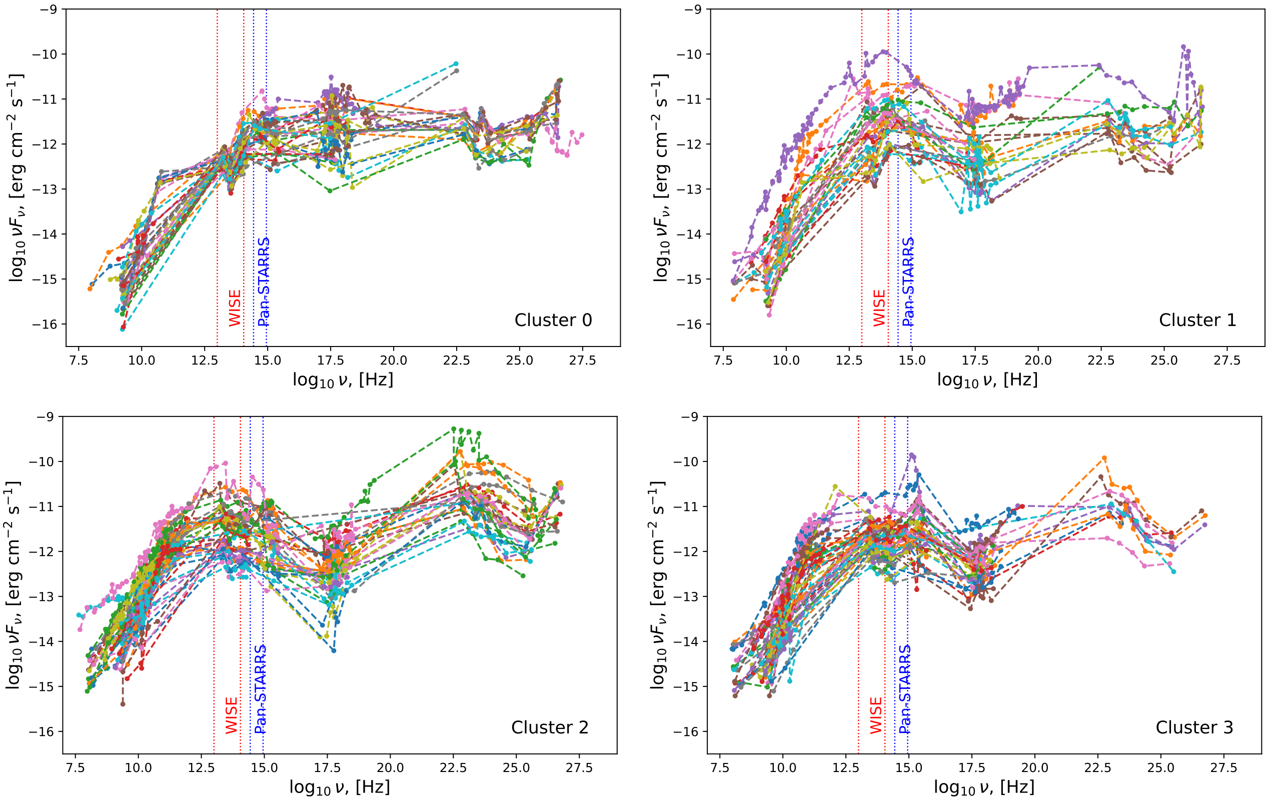


Fig. 2. Differences in typical SED shapes between clusters (rest frame frequencies). For each cluster, several tens of randomly selected spectra are plotted. We connect individual measurements with colored dashed lines to better visualize spectrum shapes. The frequency range is from radio waves (lower abscissa values) to gamma-ray radiation; the WISE (IR) and Pan-STARRS (optical) regions are designated with the vertical dotted lines. The SEDs in cluster 4 are very similar to those in cluster 3, therefore we have omitted them here to save the space.

drop out the axes with the least variances, thus reducing the dimensionality while preserving most of the information (the remaining "explained" variance). Here we chose the 90% threshold for the explained variance and reduced our 14D feature space to 6D. The final stage was k-means clustering. The number of clusters was chosen to be five, as this number gave the best results for the model with imputed missing measurements (see the next section).

In the SOM method, we have a neural network of two fully connected layers. The input layer corresponds to the dimensionality of the object vector (14D), and the second layer is organized in a grid of 200 × 320 neurons. During training, the input vectors are sequentially fed to the network and the neuron with the best-matching weights is determined. After that, the weights of this neuron and its neighboring neurons are adjusted so that to become closer to the input vector, the value of the adjustment is dependent on the learning rate and the distance from the best-matching neuron on the layer grid. As a result, the distribution of the neuron weight vectors in the feature space gradually becomes close to the data distribution, while the neurons remain structured in a 2D-grid fashion. Thus, after training the network we have an ordered 2D structure of neurons with the high-dimensional topology encoded in their 14D weights. The final step is also k-means, but in this case we make it using the 14D weights of the trained neurons and then labeling the objects according to the cluster label of their nearest neuron. The advantage of this method over the PCA dimensionality reduction is that it can restore possible nonlinearities in data distribution, while the PCA is a more straightforward and interpretable method of linear algebra.

The results obtained using both methods and their comparison are shown in Fig. 1. We use two coordinate systems: the obtained self-organized map at the bottom of the figure and the conditional 2D coordinates derived using the t-SNE method (van der Maaten, Hinton, 2008), a nonlinear transformation based on matching specific distance-based probability distributions in the high-dimensional feature space and in some lower-dimensional space, here it is our 2D plane.

Both methods expectedly show that there are no localized groups of blazars, except for cluster 0, which looks outstanding to a certain degree in the t-SNE coordinates, but nevertheless have no clear separation from the overall cloud. A somewhat surprising result, though, is that both methods construct almost the same borders between the clusters, despite of the certainly continuous distribution of data in the feature space. The Rand index between the two clusterings is 0.92, which means that the results are 92% similar. This also proves that there are no nonlinearities in the data distribution, which could not be taken into account by the PCA+k-means method. Thus, PCA+k-means has been used for further analysis as a more straightforward approach.

By estimating the influence of each of the 14 features on the first two primary components (PCA biplot), we can also

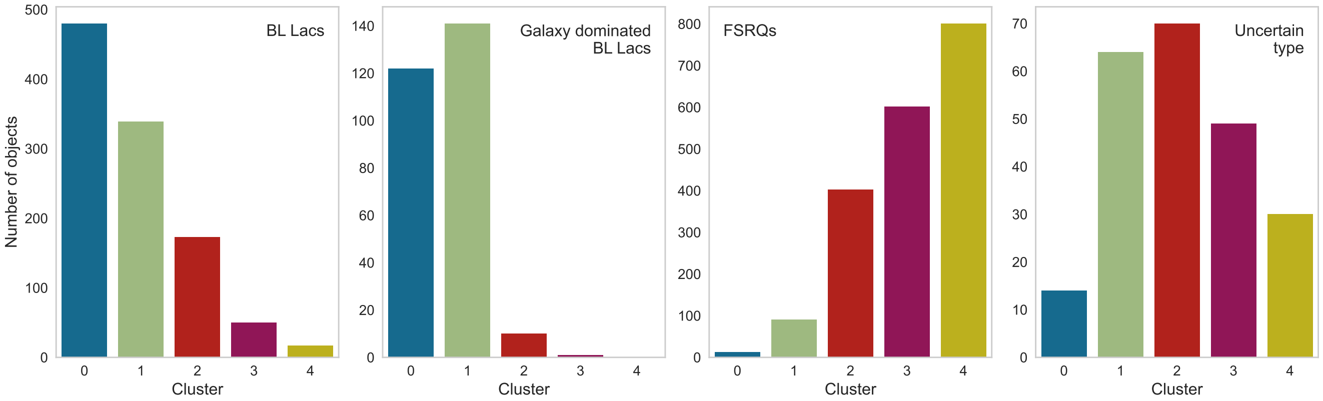


Fig. 3. Cross identification with the Roma-BZCAT classes.

evaluate their contribution to the final result. The estimates obtained show that most features contribute with similar percentage of 6–8% (taking the total as 100%), with the lowest contribution of 2.6% from the synchrotron peak frequency and the highest contribution of ~11% from the IR-to-UV hardness parameter.

5 Imputation of missing measurements

Our model dataset has a fairly great number of missing measurements. We made two clusterings: (1) with dropping all missing measurements, which shortened the catalog to 858 objects (24% of the whole sample) and (2) by imputing the missing measurement using several methods. By comparing the two clusterings, we settled on probabilistic PCA (pPCA, Tipping, Bishop, 1999) as the best choice of imputation with a Rand index of 0.89. The optimization of this index was also used as the choice for the number of clusters. The pPCA implementation[10] from Porta et al. (2005) has been adopted. A more detailed description will be given in further papers.

6 Preliminary results

The obtained blazar clusters demonstrate differing feature distributions, which is a subject of further investigation beyond the scope of this paper. Here we only demonstrate the most general result concerning the SED shapes (Fig. 2) and make a comparison with the existing blazar classification (Fig. 3).

The cross identification with the Roma-BZCAT blazars (Fig. 3) shows that in general the algorithms have divided BL Lac-type objects and flat-spectrum radio quasars (FSRQs) into two subclasses: clusters 0 and 1 and clusters 3 and 4, respectively. The "intermediate" cluster 3 contains both BL Lacs and FSRQs. The clustering has not revealed difference between the BL Lacs and galaxy-dominated BL Lacs (the sources usually reported as BL Lacs in the literature, but having SEDs with significant dominance of the galaxian emission over the nuclear one, Massaro et al., 2015), unlike in the Roma-BZCAT classification. It should be noticed that we did not aim anyhow to recreate the original classification of Roma-BZCAT, neither the algorithms "knew" about the existing classes (it would have given a trivial division).

Samples of SEDs in Fig. 2 demonstrate notable differences between the clusters. In the most distinct cluster 0, populated by BL Lac-type blazars, the synchrotron "hump" is almost unobservable. It is worth noting that all high synchrotron peakers are the members of this cluster. Cluster 1, also populated with BL Lacs, demonstrates distinctly different shapes with clear synchrotron maxima in the IR–optical range. Cluster 2, a mix of BL Lacs and FSRQs, has SEDs with synchrotron maxima shifted to lower frequencies compared to cluster 1. Blazars of this cluster have also a powerful second hump in the gamma-ray range. Interestingly, gamma-ray fluxes have not been represented in the clustering feature space due to their sparse measurements in Roma-BZCAT; the algorithms separated this cluster using other features. The SEDs in clusters 3 and 4, populated with FSRQs, are pretty similar with each other (we show only cluster 3 in the figure). It is not clearly seen on the broad frequency range but these SEDs have a specific peculiarity in their shape: an additional growth of flux densities in the optical–near-UV range, probably caused by the light from the host galaxies.

The presented effects are not subjected to the above mentioned dependence of dataset characteristics on the redshift, as the frequencies in Fig. 2 have been recalculated to the rest frame.

7 Conclusions

- The paper is devoted to the application of the cluster analysis technique to multiparametric astrophysical data: compiled characteristics of the Roma-BZCAT blazars. Similar methods can be applied to an arbitrary tabular dataset.
- A comparison of two clustering algorithms has been carried out. We have reached a coincidence of ~90%. We should notice, nevertheless, that because no localized groups are revealed in the feature space and due to

[10] https://github.com/el-hult/pyppca

the incompletness of the Roma-BZCAT catalog, the borders between the clusters are only conditional and might change for a more complete sample.
- An independent classification of the Roma-BZCAT blazars has been developed based on the analysis of their multifrequency (radio-to-X-ray) observations. Noticable differences in the SEDs of the derived classes are observed.
- There exists certain correlation with the original Roma-BZCAT classification.

Acknowledgements. The reported study was funded by the Ministry of Science and Higher Education of the Russian Federation under contract 075-15-2022-1227.

References

Arthur D., Vassilvitskii S., 2007. In SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, pp. 1027–1035.

Caliński T., Harabasz J., 1974. Communications in Statistics-Simulation and Computation, vol. 3, no. 1, pp. 1–27.

Davies D.L., Bouldin D.W., 1979. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-1, no. 2, pp. 224–227.

Kohonen T., 2001. Self-Organizing Maps.

Massaro E., Maselli A., Leto C., et al., 2015. Astrophys. Space Sci., vol. 357, no. 1, 75.

Mingaliev M.G., Sotnikova Y.V., Udovitskiy R.Y., et al., 2014. Astron. Astrophys., vol. 572, A59.

Pedregosa F., Varoquaux G., Gramfort A., et al., 2011. Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825–2830.

Planck Collaboration, Aghanim N., Akrami Y., et al., 2020. Astron. Astrophys., vol. 641, A6.

Porta J., Verbeek J., Krose B., 2005. Autonomous Robots, vol. 18, no. 1, pp. 59–80.

Rousseeuw P., 1987. Journal of Computational and Applied Mathematics, vol. 20, no. 1, pp. 53–65.

Sotnikova Y.V., Mufakharov T.V., Mikhailov A.G., et al., 2022. Astrophysical Bulletin, vol. 77, no. 3, pp. 246–263.

Tipping M.E., Bishop C.M., 1999. Neural Comput., vol. 11, no. 2, pp. 443–482.

van der Maaten L., Hinton G., 2008. Journal of Machine Learning Research, vol. 9, pp. 2579–2605.

Verkhodanov O.V., Trushkin S.A., Chernenkov V.N., 1997. Baltic Astronomy, vol. 6, pp. 275–278.

Verkhodanov O.V., Trushkin S.A., Andernach H., Chernenkov V.N., 2005. Bulletin of the Special Astrophysics Observatory, vol. 58, pp. 118–129.

Wittek P., Gao S.C., Lim I.S., Zhao L., 2017. Journal of Statistical Software, vol. 78, no. 9, p. 1–21.