Cluster Analysis of the Roma-BZCAT Blazars (Postprint)

Authors: Dmitry O. Kudryavtsev, Yulia V. Sotnikova, Vladislav A. Stolyarov, Timur V. Mufakharov, Valery V. Vlasyuk, Margarita L. Khabibullina, Alexander G. Mikhailov and Yulia V. Cherepkova

Date: 2024-05-24T00:00:00+00:00

Abstract

Based on the collected multiwavelength data, namely in the radio (NVSS, FIRST, RATAN-600), IR (WISE), optical (Pan-STARRS), UV (GALEX), and X-ray (ROSAT, Swift-XRT) ranges, we have performed a cluster analysis for the blazars of the Roma-BZCAT catalog. Using two machine learning methods, namely a combination of PCA with k-means clustering and Kohonen's self-organizing maps (SOMs), we have constructed an independent classification of the blazars (five classes) and compared the classes with the known Roma-BZCAT classification (FSRQs, BL Lacs, galaxy-dominated BL Lacs, and blazars of an uncertain type) as well as with the high synchrotron peaked (HSP) blazars from the 3HSP catalog and blazars from the TeVCat catalog. The obtained groups demonstrate concordance with the BL Lac/FSRQ classification along with a continuous character of the change in the properties. The group of HSP blazars stands out against the overall distribution. We examine the characteristics of the five groups and demonstrate distinctions in their spectral energy distribution shapes. The effectiveness of the clustering technique for objective analysis of multiparametric arrays of experimental data is demonstrated.

Full Text

Research in Astronomy and Astrophysics, 24:055011 (23pp), 2024 May

https://doi.org/10.1088/1674-4527/ad3d14

Cluster Analysis of the Roma-BZCAT Blazars

Dmitry O. Kudryavtsev[1], Yulia V. Sotnikova[1], Vladislav A. Stolyarov[1,2], Timur V. Mufakharov[1,3], Valery V. Vlasyuk[1], Margarita L. Khabibullina[1], Alexander G. Mikhailov[1], and Yulia V. Cherepkova[1]

[1]Special Astrophysical Observatory of the Russian Academy of Sciences, Nizhny Arkhyz 369167, Russia; dkudr@sao.ru
[2]Kazan Federal University, 18 Kremlyovskaya St, Kazan 420008, Russia

Received 2024 February 25; revised 2024 April 3; accepted 2024 April 8; published 2024 May 10

Abstract

Based on collected multiwavelength data in the radio (NVSS, FIRST, RATAN-600), IR (WISE), optical (Pan-STARRS), UV (GALEX), and X-ray (ROSAT, Swift-XRT) ranges, we have performed a cluster analysis for the blazars of the Roma-BZCAT catalog. Using two machine learning methods—a combination of PCA with k-means clustering and Kohonen's self-organizing maps (SOMs)—we have constructed an independent classification of the blazars (five classes) and compared the classes with the known Roma-BZCAT classification (FSRQs, BL Lacs, galaxy-dominated BL Lacs, and blazars of uncertain type) as well as with the high synchrotron peaked (HSP) blazars from the 3HSP catalog and blazars from the TeVCat catalog. The obtained groups demonstrate concordance with the BL Lac/FSRQ classification along with a continuous character of the change in properties. The group of HSP blazars stands out against the overall distribution. We examine the characteristics of the five groups and demonstrate distinctions in their spectral energy distribution shapes. The effectiveness of the clustering technique for objective analysis of multiparametric arrays of experimental data is demonstrated.

Key words: methods: data analysis – galaxies: active – (galaxies:) BL Lacertae objects: general

1. Introduction

Blazars are a rare type of active galactic nuclei (AGNs) with relativistic plasma pointing toward the Earth at a jet of relatively small angle (e.g., Urry & Padovani 1995; Blandford et al. 2019). Blazars are also among the brightest AGNs, and the Doppler-beaming effect (Madau et al. 1987; Ghisellini et al. 1993; Fan et al. 2017) makes their jet emission even more boosted and visible up to z 6 (Belladitta et al. 2020). They are characterized by complex properties such as extreme variability at all wavelengths, high luminosity, high degree of polarization, and brightness temperatures exceeding the Compton limit (Urry 1999).

Blazars are the dominant sources in the extragalactic gamma-ray sky. Because of the relativistic amplification of their emission, sources even at high redshifts are observed. The investigation of the multiwavelength properties of high redshift blazars is especially important as they are the most powerful non-explosive astrophysical sources and their study can be crucial for understanding jet formation and propagation around supermassive black holes. The recently found con-

nection between blazars and IceCube sources of high-energy neutrinos (Plavin et al. 2020) also adds to the topicality of their investigation.

The typical spectral energy distribution (SED) of a blazar is dominated by the non-thermal radiation from the jet and consists of the synchrotron (peaking between the far-infrared and the soft X-ray bands) and inverse-Compton (peaking in the hard X-ray to gamma-ray bands) humps (Abdo et al. 2010). Besides that, the SED of a blazar can also feature thermal radiation from the host galaxy (infrared (IR) hump or stellar emission) and emission from the accretion disk around the central black hole ("blue hump") and from the broad line region (Giommi et al. 2012b). Blazars exhibit large and rapid variations on a variety of timescales from years to intervals even shorter than an hour (e.g., Padovani et al. 2017 and references therein).

Blazars are subclassified as flat-spectrum radio quasars (FSRQs) and BL Lacerta-type objects (BL Lacs) based on their optical spectra: FSRQs show broad emission lines, while BL Lacs display either very weak emission lines or are completely featureless (e.g., Urry & Padovani 1995; Falomo et al. 2014). Another classification was proposed based on the broad-line region (BLR) luminosity in Eddington units (Ghisellini et al. 2011): sources with L_{BLR}/L_{Edd} higher or lower than 5×10^{-4} were classified as FSRQs or BL Lacs, respectively, according to a transition of the accretion regime from radiatively efficient to inefficient between the two classes.

Based on the peak frequency (_{peak}) of the synchrotron energy hump, blazars are usually subclassified as low (LSP, _{peak} $< 10^{14}$ Hz), intermediate (ISP, 10^{14} Hz $<$ _{peak} $< 10^{15}$ Hz), or high-synchrotron peaked (HSP, _{peak} $> 10^{15}$ Hz) blazars (Abdo et al. 2010; Fan et al. 2016). Most HSP and ISP blazars have been classified as BL Lacs, while the LSP class contains both FSRQs and LSP BL Lacs (Böttcher 2019; Prandini & Ghisellini 2022).

Inspired by observational data, alternative physical categorizations for blazars have been proposed: for instance, based on sources with intrinsically weak or strong O II and O III emission lines (Landt et al. 2004); based on different accretion rates (the luminosity of the broad line region relative to the Eddington luminosity) of the two subclasses of blazars (e.g., Ghisellini et al. 2011; Sbarrato et al. 2012); based on the ionizing radiation emitted from the accretion disk (Giommi et al. 2012a; Giommi et al. 2013; Giommi & Padovani 2015); based on the kinematic features of radio jets (e.g., Hervet et al. 2016); etc.

The above-mentioned numerous approaches to blazar classification tend to use a single categorical parameter (presence/absence of emission lines) or a single numerical parameter (HSPs, L_{BLR}/L_{Edd}), in the latter case also categorized by setting a threshold defined by the researchers. At the same time, blazars, like any objects, have numerous measurable characteristics that define their properties, and contemporary computing power and machine learning (ML) methods allow us to investigate a large number of characteristics in all their complexity.

In this paper we perform multiparametric cluster analysis for the Roma-BZCAT catalog (Massaro et al. 2015), a sample of blazars with the most complete set of characteristics observed in different ranges of the electromagnetic spectrum. The aim is to divide the blazars into groups with similar properties to further analyze the differences between the groups, check the performance of the ML (clustering) methodology, and compare it with generally accepted classification approaches.

2. General Conception of Cluster Analysis

Cluster analysis, or clustering, is a classical problem of unsupervised ML—learning with unlabelled data—when the model is not given any target variable in advance, in this case the classes of considered objects. The aim of the clustering model is to combine similar objects into groups (clusters) based on the similarity of their characteristics, or features. The principal idea is that when these characteristics are expressed numerically, objects with similar properties are located closer to each other in the feature space than those with greater differences. In the simplest case of two–three features and clearly separated clusters, this problem can be solved visually by constructing usual two-dimensional (2D) or three-dimensional (3D) scatter diagrams. In the general case of an arbitrary number of characteristics, the clustering must be performed in an n-dimensional feature space. ML algorithms are capable of solving such problems successfully even for complex distributions. Note that clustering in terms of ML should be distinguished from classification, which is a separate problem of supervised ML, when the model is trained to guess classes known a priori. The main difference between the unsupervised problem of clustering and the supervised or semi-supervised problem of classification is the approach itself: while in the latter case we exploit a known classification developed by other methods and assign known classes to new objects, in clustering we develop a new classification based solely on the data collected for the objects. This allows one to describe a sample based on experimental data, avoiding as much as possible the subjective approach to the division of objects into different types.

The mathematical formulation of clustering is as follows. Let X be a set with dimension N × M, where N is the number of objects x_n and M is the number of their features x_m. The set {X} can be represented as a matrix X = (x_{nm}) with n = 1, 2,…,N [1, N], m = [1, M]. Let Y be a set of cardinality K, where K is the number of clusters, {Y} = [1, K]. The solution of the clustering problem is finding an algorithmic function a: {X} → {Y} that assigns a singular label y_k, k = [1, K] to each object x_n, n = [1, N] in such a way that objects with similar properties (ideally forming separated groups in the feature space x_m) correspond to the same label (cluster). Each object x_n can be represented in the feature space as a vector of dimension M: x_n = (x_{n1},…,x_{nM}). The measure of object similarity is a metric of distance between the vectors; in our case this is the Euclidean distance where x_i and x_j are any two vectors x_n (sample objects). In order for the features to have equal priorities in the

clustering process, they should be normalized beforehand to the same scale.

We should note that cluster analysis is a heuristic and its exact results are always model dependent, both on the choice of features selected for the modeling and on the clustering algorithm a: X → Y. An additional degree of freedom is the number of clusters K, which in most algorithms is defined a priori and evaluated heuristically based on the data. In this sense the obtained structure of the clusters should not be considered as an established natural phenomenon, especially when the clusters are not well separated; the cluster analysis, to a greater degree, is an instrument to search for patterns in the sample rather than investigation of individual objects.

Generally, the problem solution can be divided into several stages: (1) data collection and feature engineering; (2) selection of characteristics for the model feature space; (3) clustering with different algorithms in search of a model with the best quality metrics; (4) interpretation of the result, i.e., analysis of the difference between objects in different groups. In the following sections we successively consider these stages.

3. Initial Data

In this section, we describe the databases and characteristics used to compile the dataset for this project. Some characteristics are not directly used for clustering because they are available for only a small number of blazars; nevertheless, they are useful for subsequent analysis.

The basis for our dataset is the 5th edition of the Roma–BZCAT catalog of blazars (Massaro et al. 2009, 2015). The catalog contains a list of 3561 AGNs classified by the authors as blazars based on their observed properties. The following information is available: coordinates; redshifts; optical magnitudes in the R band from USNO-B1.0, r filter from Sloan Digital Sky Survey (SDSS) DR10, or in other filters when these data are absent; 1.4 GHz (NVSS, FIRST, 21 cm) and 143 GHz (Planck, 2.1 mm) radio flux densities; X-ray (ROSAT, Swift-XRT) and gamma-ray (Fermi-LAT) fluxes as well as the radio-to-optical spectral index characterizing the ratio between radio and optical emission. Note that the R band magnitude presented in Roma-BZCAT describes the optical radiation rather loosely, being sometimes obtained from different photometric filters. For this reason we used more consistent data on optical magnitudes from other catalogs.

Based on the BLcat RATAN-600 measurements at frequencies 1–22 GHz covering the period of observations 2006–2022 (Mingaliev et al. 2014; Sotnikova et al. 2022) and the CATS database (Verkhodanov et al. 1997, 2005), we calculated the averaged spectral flux density at a frequency of 5 GHz, radio spectral indices, and radio variability. The averaged spectral indices α were calculated for the 1–2, 2–5, 5–8, 5–11, 8–11, 11–22, 8–22, and 5–22 GHz ranges. The radio variability is given at frequencies of 1, 2, 5, 8, 11, and 22 GHz. The variability index was calculated using the formula from Aller et al. (1992) where

S_{max}, S_{min} are the maximum and minimum flux densities, respectively, and σ{Smax}, σ{Smin} are their standard errors.

The IR measurements are represented by data from the Wide-field Infrared Survey Explorer (WISE) in the W1, W2, W3, W4 bands (3.4, 4.6, 12, 22 m) and by the Two Micron All-Sky Survey (2MASS) data in the JHK bands (1.25, 1.65, 2.2 m). The data were taken from the NASA/IPAC Infrared Science Archive using the pyvo Python library; identification of the blazars was carried out by coordinates using the cone search query in a 9 field of view (the WISE angular resolution is about 6).

The optical range is represented by the Pan-STARRS measurements in the grizy filters (effective wavelengths of 4810, 6170, 7520, 8660, and 9620 Å). The data were obtained by a standard request to the archive with a list of object coordinates. Additionally, for the optical range we downloaded the SDSS DR17 data in the ugriz filters (effective wavelengths 3557, 4702, 6175, 7491, and 8946 Å) via the provided web form (the requests were made automatically using a script).

The ultraviolet (UV) range is represented by the GALEX FUV and NUV channels (effective wavelengths of 1538.6 and 2315.7 Å respectively). The data were obtained from the Mikulski Archive for Space Telescopes (MAST) using the astroquery Python library. A UV counterpart was identified as the object closest to the coordinates in a 7 .2 field of view (GALEX angular resolution is 4).

In all the above cases, if there were several objects in the field of view of a search query, we chose the closest by angular distance. The possible presence of outliers was additionally controlled by histograms of angular distance deviations from the blazar coordinates. Since the identification was carried out in automatic mode, we cannot completely exclude incorrect identifications in some cases, but since we took the minimum possible search radius, comparable to the resolution of the instruments, such cases should be rare and can be discarded during subsequent data cleansing.

The extinction was determined using NED's Coordinate and Galactic Extinction Calculator. For the GALEX FUV and NUV channels, we used the extinction law from Fitzpatrick (1999); the calculations were made with the extinction Python library. For the WISE IR range, the extinction was considered zero.

To determine the peak frequency of the synchrotron component, we used SEDs obtained from the SED Builder tool of the Italian Space Agency (ASI) Space Science Data Center (SSDC). The measurements were downloaded using Selenium WebDriver, which allows interaction with websites in an automated mode.

Thus, in the initial dataset we managed to collect a fairly extensive set of observed data in various ranges of the electromagnetic spectrum: from radio to gamma emission. The dataset also includes information about redshift, spectral indices, and estimates of variability in the radio range. In the next section, we describe additional processing of the derived data in order to extract more infor-

mative features and present a complete dataset of the obtained and calculated characteristics.

4. Calculation of Blazars' Characteristics

The initial characteristics obtained from catalogs often cannot be directly used in the model because they might describe not only the properties of the objects but also other factors affecting the result. For example, the magnitudes depend on the photometric system of a particular catalog. Therefore, to solve our problem, we should obtain characteristics that in the best possible way describe the physical properties of the blazars.

Since blazars are located at different cosmological distances, these characteristics should be related to the rest frame of an object. Unfortunately, this is often impossible in practice. For instance, to estimate the luminosity at a certain frequency at cosmological distances, a good description of the SED shape is necessary, but for most of the blazars we have only point estimates of this shape at a number of frequencies; moreover, the SED can be variable. Empirical analytical dependencies (see, e.g., Chilingarian et al. 2010) work only at small redshifts $z < 0.5$.

Here we will use rest-frame characteristics where possible; the remaining features will be given in the observer's frame of reference, and the distance to a blazar will also be set as one of the parameters. In Section 5 we describe how this might affect the results in more detail.

The distance to a blazar could be described directly by the redshift, but in this scale the distance distribution of the blazars appears rather crowded and uneven. In Figure 1 the dependences of the monochromatic radio luminosity on redshift and on comoving distance are presented. The comoving distance scale is more suitable for modeling. Distances were determined from the redshifts using the astropy library (Astropy Collaboration et al. 2022) and based on the ΛCDM cosmology with Planck Collaboration parameters (Planck Collaboration et al. 2016): $H_0 = 67.74$ km s^{-1} Mpc^{-1}, $\Omega\_m = 0.3089$, $\Omega\_\Lambda = 0.6911$. Along with this, some other parameters were calculated: luminosity distance, distance modulus, lookback distance, and the Universe's age in the blazar rest frame at the time of light emission. The monochromatic (5 GHz) radio luminosity was estimated by the formula

$$\log L_5 = \log S_5 + \log(4\pi D\_L^2) + (1+\alpha)\log(1+z),$$

where D_L is the luminosity distance, S_5 is the flux density at 5 GHz, z is the redshift, and α is the averaged spectral index taken for 5–11 GHz or, where measurements were absent, for 5–8 GHz.

The dependence of radio luminosity on distance in Figure 1 is completely or partially caused by selection effects. One of them is the Malmquist bias (e.g., Butkevich et al. 2005): roughly speaking, if we assume a normal luminosity

distribution, the same at all distances d, then with increasing distance the detection limit shifts to higher luminosities, which reduces the number of objects in the left wing of the distribution, and simultaneously the expected number of high-luminosity objects grows with increasing area of the sphere of radius d, which increases the probability of detecting bright objects in the right wing of the distribution. Since both effects are proportional to d^2, a linear increase in average luminosity with distance should be observed. It must be noted that the observed dependence is not linear at distances 2000 Mpc; however, selection cannot be excluded in this case either, since the measurements were obtained in different surveys, and the interest in the nearest but not necessarily luminous objects is natural.

The optical flux densities in the observer's frame of reference were calculated based on the AB magnitude system using the formula

F_ = F_ 0 × 10^(-0.4(m_ - E_)),

where is the effective frequency of a photometric band, F_ 0 is the flux density from zero magnitude, and m_ and E_ are the magnitude and extinction in the photometric band (extinction according to the NED data), respectively. The and F_ 0 values were taken from the descriptions of corresponding photometric systems (Cohen et al. 2003; Morrissey et al. 2007; Jarrett et al. 2011; Tonry et al. 2012). For the SDSS we implemented the corrections u = u - 0.04, z = z + 0.02, according to accepted practice. Summary data are given in Table 1.

Based on the available stellar magnitudes and extinctions, we calculated optical colors in various photometric systems. However, their use in clustering seems impractical. In Figure 2, SEDs for five random blazars are presented; points mark flux densities in the WISE, 2MASS, Pan-STARRS, and GALEX passbands. For better visualization of individual SEDs the points are connected by lines. Flux densities are related to stellar magnitudes, and the difference between pairs of points generally represents the optical colors. It can be seen that in the frequency range of each individual instrument—for example WISE and Pan-STARRS, whose data we further use in the clustering—a predominant slope of the spectrum can be assigned for a particular blazar, while the ratios between flux densities for pairs of points ("colors") can sometimes differ significantly. Thus, the use of colors in clustering can create additional noise that would not allow the algorithm to estimate the predominant slope of the spectrum. For this reason, we calculated special features: tangents of the spectrum slope in the WISE and Pan-STARRS ranges (the slopes for 2MASS and GALEX were not used due to lack of data). The spectrum slopes were approximated by a linear dependence using the method of least squares.

To determine the frequency of the synchrotron peak, we used flux densities downloaded from the ASI SSDC SED Builder. The position of the peak was determined by approximating a SED with polynomials of the third or, in some cases, second degree. The flux density measurements were represented by data from the SSDC resident catalogs and other catalogs. To calculate the parameters

of the polynomial, the program used only the resident catalogs, while we visually controlled the obtained result using all the measurements (see Figure 3).

The frequencies were transformed to the source's rest frame using _{peak},rest = _{peak},obs(1+z). A graph of the obtained values versus comoving distance is shown in Figure 4.

The estimates of optical variability were calculated from the minimum and maximum point-spread function (PSF) magnitudes presented in the Pan-STARRS data for each of the five filters (grizy). Our variability estimates are simple differences between these values for all blazars with two or more observing epochs (according to the number of measurements included in the mean PSF magnitude from detections in a corresponding filter). The estimates are rough, as they depend on the time when the observing epochs were carried out and on their number. As an example, the distribution of the number of observations in the i filter is shown in Figure 5.

Roma-BZCAT presents X-ray fluxes in the range 0.1–2.4 keV (5–124 Å); we recalculated them into the logarithmic radiation scale $\log_{10}$[erg cm^{-2} s^{-1}]. Gamma-ray fluxes represented in Roma-BZCAT in photons cm^{-2} s^{-1} for the range 1–100 GeV were also recalculated into the scale $\log_{10}$[erg cm^{-2} s^{-1}], taking the middle of the range as the photon energy: 50 GeV.

In addition to the Roma-BZCAT radio-to-optical spectral index, which characterizes the ratio between radio and optical fluxes, we calculated other parameters that describe flux ratios at different frequencies of the electromagnetic spectrum (let us call these parameters the "hardnesses"). They are calculated as decimal logarithms of the ratios between flux densities at the frequencies F$_{1\cdot4}$ GHz (radio), F_{W2} (IR), F_i (optical), F_X (X-rays), and F_γ (gamma-rays); e.g., the IR/optical hardness is $\log_{10}$(F_{W2}/ F_i). The clustering model uses six such ratios because not enough data are available for all frequencies.

The complete list of parameters available in the final dataset includes more than 100 items. The dataset is schematically presented in Figure 6 and is published in the VizieR database.

5. Feature Space and Model Data Set

The immediate use of an entire dataset by ML algorithms is impossible. First, the data must be appropriately preprocessed to obtain meaningful results. Moreover, some features may have a large amount of missing data, some strongly correlate with each other, others are auxiliary and are related to the actual properties of the objects (e.g., extinction), and some should be discarded since they do not affect the final result but increase dimensionality. Best practice also suggests providing the ML model not only with direct characteristics of the objects but also with ML-specific features, which are combinations or transformations of actual characteristics that help the model to "understand" data better. Therefore, the next step after collecting the data is constructing the

model dataset that will be directly used by the clustering algorithms. This includes selection of characteristics relevant to the problem, feature engineering and transformations, data cleansing, imputation of missing values, scaling, etc.

Note that while this model dataset is constructed to be used by ML algorithms and undergoes certain transformations during this process, the predictions that a trained model produces in the end—such as the cluster label in our case—are object-specific. In other words, having cluster labels (membership) for the blazars as a result of our clustering, we can further analyze any other characteristics of the blazars from the original or model dataset, along with even new ones.

The choice of features to form the feature space of the clustering can be made using various approaches. In our case we used as many available blazar characteristics as possible. In general, if there is not a predetermined scope of investigation—that is, the study is not aimed at revealing relationships between some specific characteristics selected beforehand—this approach allows us to describe the objects under investigation most completely, form groups of similar objects without a priori assumptions, and increase the reproducibility of the clustering results.

Here we consider in detail the preparation of the model dataset to be directly used by the clustering algorithms.

5.1. Dropping Unnecessary Characteristics

Different ways of describing cosmological distances (the "Cosmology" cell in Figure 6) are a priori related by analytical dependencies and do not give new information to the model; therefore only one of them, the comoving distance, was chosen, as the distribution of blazars in this scale is most uniform (see Figure 1 above).

Stellar magnitudes and flux densities are also analytically related. For the model dataset the latter were selected since they do not depend on the photometric system and are more directly connected to physical properties: the luminosity of an object and the distance to it.

The R.A., decl. coordinates were excluded as we do not expect heterogeneity here and also because the spherical coordinate system (full circle in R.A., semicircle in decl.) leads to an artificial global structure in the data.

The blazar types according to the Roma-BZCAT classification (BL Lacs, FSRQs, etc.) are a categorical feature, which in combination with other characteristics having continuous distributions leads to a trivial solution: division into the known types. Therefore, this information was removed from the model dataset.

As noted above, instead of colors we used more smoothed parameters: tangents of the spectrum slope in the WISE and Pan-STARRS passbands. For this

reason, colors were not considered in modeling. Other photometric passbands were not included due to lack of data (see below).

Unfortunately, we had to exclude from the model dataset the data on radio and optical variability: these values significantly correlate with the number of observations (Tornikoski et al. 2000; Nieppola et al. 2007; Khabibullina et al. 2023), which in our case generated an artificial cluster of radio variability: the differences were clearly visible at the most observed frequency of 5 GHz, while at other frequencies with fewer measurements the cluster was not as distinguished.

We also excluded the radio spectral indices. The modeling showed that the groups found in clustering did not have significant differences; therefore the final model was built without them.

Extinctions were dropped as these are auxiliary data used in flux density calculations.

5.2. Dropping Characteristics with Many Missing Values

A separate problem is missing values: almost all characteristics, to a greater or lesser extent, suffer from lack of data. The processing of missing values is covered further in more detail, but characteristics with a very large number of absent measurements cannot be used in modeling. According to accepted empirical practice, we excluded features with more than 40% missing data: the GALEX FUV values and associated characteristics, SDSS and 2MASS data, Roma-BZCAT (Fermi) data on gamma-ray fluxes, and data on radio flux densities at 143 GHz (1.4 and 5 GHz have remained).

We once again note that characteristics excluded from modeling can be used for analysis after clustering.

5.3. Outliers

Measurements outstanding significantly from the distribution of a characteristic can distort results of most clustering algorithms. We considered the distributions of the features selected for modeling and visually evaluated their boundaries. Blazars outside designated distribution boundaries were excluded from the model dataset (not more than several objects for a feature, a total of 34 objects), and their classification into groups was carried out after clustering using a separately trained k-nearest neighbors (KNN) model (Section 7.2.2).

5.4. Multicollinearity: Combining Similar Characteristics into Metafeatures

The initial data contain sets of characteristics that naturally correlate with each other: the flux densities at different frequencies and part of the flux density ratios (hardnesses). Standard techniques for processing multicollinear features in ML are either their exclusion or Principal Component Analysis (PCA): transformation of the features by linear algebra methods into new mutually orthogonal

(zero correlation) characteristics oriented in the feature space along the axes of greatest variance.

In this paper we used PCA to combine a number of flux densities at different frequencies into one meta-feature, using the first principal component. The sets of input characteristics and their corresponding meta-features are shown in Table 2. The choice of these two meta-features is based on the simple core–jet model of AGNs, where 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.

Note that for an individual blazar, measurements for some input characteristics may be missing. In such cases we imputed the missing values using probabilistic PCA (PPCA; Tipping & Bishop 1999). The PPCA implementation from Porta et al. (2005) was adopted. The method is considered in more detail in Section 7.2.1. We applied PPCA separately for each set of characteristics from Table 2. If all corresponding values for an object were missing, we left them empty (at this first stage).

In contrast to the above, for the hardnesses—some of which may also correlate with each other—it is important to preserve information about differences in flux densities at various ranges of the electromagnetic spectrum. In this case, multicollinearity was removed later during dimensionality reduction of the entire model dataset, also using PCA but taking more principal components (see Section 7.1.1).

5.5. Scaling

To ensure equal priority of features in PCA and clustering, all of them must be expressed in a unified numerical scale. At all stages where this was necessary, we used the scikit-learn (Pedregosa et al. 2011) standard scaler, which produces zero mean and unit variance for a feature.

5.6. Model Data Set

The clustering model dataset after data cleansing and feature transformations includes 14 features. The heatmap of the dataset is shown in Figure 7. The columns of the table are designated along the x-axis, the y-axis corresponds to its rows, in which individual object vectors are located. The heatmap shows missing data (less than 40% in each column).

6. Selection and other Hindering Effects

We should note that our feature space is subjected to some effects that are negative for interpretation of results that could be obtained from clustering. In the first place, all selection effects are preserved, and almost all characteristics are dependent on the distance to the blazars (or redshift z). For example, as

already mentioned in Section 4, the redshift-corrected radio luminosity nevertheless shows strong dependence on z due to the Malmquist effect. Even the flux density ratios are dependent on distance because of the cosmological rest-frame drift, which could not be corrected due to the absence of an accurate SED model for each blazar.

At the same time, these effects can be considered useful for clustering because they could potentially help separate classes that naturally demonstrate different distance distributions (because of selection in the data or not); they also contribute to more accurate probabilistic imputation of missing values (see Section 7.2.1). It is for these reasons that we leave the comoving distance and raw flux densities as model characteristics. The 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 sources. In our dataset we took the average characteristics of the objects in the way they are presented in most catalogs. In some cases different characteristics may be measured in different states of blazar activity (active/quiescent) (see, e.g., Raiteri et al. 2014). This restricts our results to only the groups' statistical properties, and any conclusion for an individual source must be treated with great caution.

Finally, the BZCAT catalog is not a complete flux-limited list of blazars. Although the incompleteness of the sample still allows us to perform clustering and analyze the observed differences, it could influence the distribution of blazars within the clusters, i.e., population of certain groups (boundaries of the clusters in the feature space) may change for a more complete sample. We evaluate the effect of data incompleteness in more detail in Section 7.3.

7. Clustering

The clustering was carried out first with a subsample of blazars that had no missing data in the model dataset (858 blazars, 24% of the BZCAT catalog) and then with the full sample and imputed missing data. We also tested various clustering algorithms: several from the scikit-learn library (Pedregosa et al. 2011) and Kohonen's self-organizing maps (SOMs; Kohonen 2001; Wittek et al. 2017) based on a competitive neural network.

7.1. The Subsample Without Missing Values

7.1.1. Clustering with K-means and PCA Dimensionality Reduction We experimented with several clustering methods offered by the scikit-learn library (Pedregosa et al. 2011): k-means, Gaussian mixture, agglomerative clustering, and spectral clustering. The results were compared by internal clustering validation metrics: the silhouette (Rousseeuw 1987), Calinski–Harabasz (Caliński & Harabasz 1974), and Davies–Bouldin (Davies & Bouldin 1979) scores. As the final option, the combination of PCA dimensionality reduction with k-means clustering was chosen. For dimensionality reduction, we took an explained variance of 90% as a criterion, thus converting the 14 model dataset

features into six (uncorrelated) principal metafeatures: mutually orthogonal components. After that, k-means clustering (Arthur & Vassilvitskii 2007) was performed in this 6D space.

A comparison of clustering validation metrics calculated for various numbers of clusters shows that the data distribution is, not surprisingly, a continuous cloud without localized groups. Thus, the number of clusters may be determined rather loosely. We selected the number of clusters based on the best (90%) match of clustering results for the subsample without missing values and for the full sample (see Section 7.2.1). With this approach, the optimal number of clusters turns out to be five. For comparison, the popular "elbow" method, taken as a first approximation and based on analysis of decreasing distortion (average squared Euclidean distance from the centroids of respective clusters), gives the number of clusters equal to four.

These results are further used as a baseline model for clustering the entire Roma-BZCAT catalog and are also compared with the approach based on SOMs. The visualization of the obtained clusters is presented in Section 7.1.3 along with the SOMs.

The PCA also allows us to estimate the significance of features for clustering. To this end, we constructed a PCA biplot (Figure 8) using the Yellowbrick library (Bengfort & Bilbro 2019). The figure shows the projection of the dataset onto the plane of the two primary components, and the lengths of vectors correspond to the importance of each feature, reflected by the magnitude of corresponding values in the eigenvectors of the primary components. The directions of vectors also demonstrate the degree of correlation (same direction) or anticorrelation (opposite direction) between features. Numerically, the contributions are given in Table 3.

Note that here we do not use characteristics related to gamma-ray measurements because of data scarcity. The gamma-ray range, nevertheless, may be of great importance for blazar classification. We consider this problem in more detail in Section 7.3 and give there a similar table (Table 4), where the importance of features is recalculated taking into account gamma-ray emission.

7.1.2. Self-organizing Maps Kohonen's SOMs (Kohonen 2001) are a neural network with competitive learning, used for clustering and visualization of multiparametric data that can contain non-obvious patterns. In particular, SOMs solve the problem of projecting a multidimensional space into lower dimensions—onto a plane or into 3D space—where data vectors are grouped according to the degree of similarity of their parameters, which allows one to perform clustering, e.g., to separate different populations of sources.

There are many software packages for data analysis by the SOM method, written in different programming languages. In this study we chose the Python package somoclu (Wittek et al. 2017).

In the SOM clustering, a grid of 200 × 320 output neurons was built with the number of weights for each neuron equal to the dimension of the input vector (our object). The SOM algorithm finds the Euclidean distance between vectors in a multiparametric space (the parameters are scaled to the interval [0, 1]) and adjusts the weights of neurons so that they become structurally similar to the distribution of input vectors in the feature space. In this way, input vectors (objects) become arranged in certain areas on the output 2D SOM map such that objects with similar parameters are located close to each other. At the same time, the distribution of neuron weight vectors in the feature space becomes close to the data distribution. In other words, after training the network we have an ordered 2D structure of neurons with high-dimensional data topology encoded in their multidimensional weights.

The final step is also k-means, but in this case we apply it using the weights of the trained neurons and then label objects according to the cluster label of their nearest neuron.

The advantage of this method over PCA dimensionality reduction is that it can restore possible nonlinearities in data distribution, while PCA is a more straightforward and interpretable method of linear algebra.

7.1.3. Cluster Visualization and Comparison of the Two Methods
To visualize clustering results we can use the t-distributed stochastic neighbor embedding (t-SNE) algorithm (van der Maaten & Hinton 2008) or the 2D coordinates derived in SOM clustering. The t-SNE algorithm converts similarities between data points to joint probabilities and tries to minimize the Kullback–Leibler divergence (Kullback & Leibler 1951) between joint probabilities in the low-dimensional embedding and the high-dimensional data. In the SOM approach, coordinates are obtained as a result of neural network mapping.

The result of our PCA + k-means six-dimensional clustering embedded in 2D space is shown in the left part of Figure 9 in t-SNE (top) and SOM (bottom) coordinates, respectively. The right part of the figure shows, accordingly, the SOM clustering on the SOM plane.

We should note that t-SNE (top panels in Figure 9) is a nonlinear algorithm focusing on local similarity of points, and results also depend on selection of hyperparameters (mainly perplexity); therefore t-SNE visualization cannot be interpreted as a precise description of object positions in feature space. For example, formation of apparently localized groups or some displacement of points belonging to different clusters can be of artificial nature. Nevertheless, the figure shows that results of both PCA + k-means (upper left) and SOM (upper right) clusterings are well described by t-SNE visualization. Some separation of cluster 0 from the general cloud is clearly visible.

The bottom panels show the same PCA + k-means and SOM clustering on the SOM 2D plane. Comparing left and right panels in Figure 9, we can conclude

that the two methods give similar results and, despite the absence of localized groups, the boundaries of clusters are pretty much the same for both methods.

To compare results numerically, we used the Rand index (Rand 1971) calculated in the standard way:

R = (a + b) / C(N, 2),

where a is the number of object pairs that remained in the same cluster after new clustering, b is the number of pairs that remained in different clusters, and N is the total number of compared pairs. The Rand index shows what percentage of objects does not change cluster membership between two clusterings. For comparison of PCA + k-means and SOM clusterings on the subsample without missing values, we achieved a Rand index of 0.95.

Additionally, we also tested the approach where t-SNE is used directly to reduce dimensionality of data, followed by clustering in the 2D or 3D space of t-SNE coordinates. This method gave good results for the sample without missing values; however, for imputed data and the full sample the results were unsatisfactory.

7.2. Full Sample

7.2.1. Missing Data Imputation Clustering of the complete BZCAT catalog was carried out according to the same scheme as for the subsample without missing values: dimensionality reduction using PCA followed by k-means clustering. The main difference is the need to fill in missing values in the model dataset (Figure 7). We tested several approaches: imputation with median values, ML regression models (namely XGBoost (Chen & Guestrin 2016) and scikit-learn implementations of Random Forest and Histogram-based Gradient Boosting), and finally imputation of missing values using PPCA (Tipping & Bishop 1999), which showed the best results.

The method introduces a latent variable z corresponding to an M-dimensional PCA subspace and probability distributions p(x|z) and p(z) such that p(x) = p(x|z)p(z)dz, where x is the observed variable (D-dimensional object vector in feature space) and p(z) is the standard normal distribution. The matrix W (equal to zero after scaling), vector , and constant σ^2 determine the PCA transformation. PPCA reduces to classical PCA at $\sigma^2 = 0$. Such mathematical formalism allows determination of W, , and σ^2 via the expectation-maximization (EM) algorithm and imputation of missing values by sampling from latent p(z|x). In our case we used a PPCA implementation that calculates imputed values along with W and σ^2 (Porta et al. 2005), which is considered by the authors as a more efficient approach.

This final variant (PPCA) for imputation of missing values was chosen based on maximum similarity of feature distributions in obtained clusters for two samples: (1) when all missing values had been dropped and (2) with imputed values. The clustering results of the sample without missing values were used as reference cluster labels.

The similarity of distributions was estimated by the Kullback–Leibler divergence (Kullback & Leibler 1951) calculated in our case as:

D_{KL}(P||Q) = Σ_{k K} Σ_{x X} P_k(x) log(P_k(x)/Q_k(x)),

where X are features in the model dataset, K are clusters, and P(x) and Q(x) are probability distributions on sample space X, with P(x) for the full sample and Q(x) for the reference subsample without missing values.

Additionally we evaluated the Rand index between the two clusterings: R 90%, which is quite good given the increase in number of objects by about four times and the fact that cluster boundaries are drawn in a continuous "cloud" without clear localization of groups.

The visual comparison of model feature distributions in the subsample without missing values and in the full sample is shown in Figure 10. We can see that distributions retain their shape well after the substantial increase of the sample. Note that observed discrepancies for some parameters (e.g., substantial difference in median log peak_n values for cluster 0) should not be treated as errors, as membership of new objects in clusters is defined based on the entire number of features, and some changes in distribution shapes are expected after increasing the sample by several times. For instance, the strongest observed difference in log peak_n distributions is in good agreement with the lowest importance of this feature for clustering (see Table 3).

We also performed clustering of the full dataset by the SOM method, which resulted in a Rand index of 0.92 with respect to the PCA + k-means technique. Thus, the two clustering methods showed 92% similarity for the complete dataset, along with even better concordance for the smaller dataset without missing values (95%). The similar results also affirm that there are no nonlinearities in the data distribution that could not be taken into account by the PCA + k-means method. Therefore, PCA + k-means has been used for further analysis as a more straightforward approach.

The totality of results obtained allows us to perform cluster analysis not only for the subsample without missing values but for all blazars in the Roma-BZCAT catalog. The flowchart of clustering stages is shown in Figure 11, with the final result for the full sample highlighted in the lower right corner.

7.2.2. Classification of Outliers To include all Roma-BZCAT blazars in clustering results, the membership of 34 outliers filtered out at the data cleansing stage (Section 5.3) must be determined. We used the KNN classifier from scikit-learn (Pedregosa et al. 2011) to complete this task. The dataset with obtained cluster labels, which acted as a target variable for training the classifier, was divided into training and test samples in a ratio of 0.1. Hyperparameters were optimized by 5-fold cross-validation on the training sample using grid search: the number of nearest neighbors in the [5, 100] range without weighting and with distance-based weights. The quality metric was the F1-score. The final

hyperparameters—70 nearest neighbors with weights inversely proportional to distance—gave an F1-score of 0.94 on the test sample: the harmonic mean of precision and recall for the trained classifier reaches 94%.

For the trained classifier to work correctly, all stages of data preprocessing for outliers must be performed in the same way as for the training dataset, but here we could not use PPCA to replace missing values since the PPCA implementation we used computed them along with PCA transformation parameters, which must be fixed during inference. For that reason, we used the following approach: (1) instead of the first PPCA step (for multicollinear flux densities transformed to metafeatures), we took the mean value over corresponding flux densities for each object; (2) for the second PPCA step, missing values in the model dataset were imputed as mean values over a column (which is actually zero after scaling); (3) other transformations (scaling, traditional PCA, etc.) corresponded to the main clustering model.

7.3. Robustness of Clustering to Dataset Incompleteness and Feature Selection

Two conditions that can influence obtained results are the incompleteness of the Roma-BZCAT sample and the features we selected for clustering. The former can change cluster boundaries after taking a sufficiently large amount of new blazars, and the latter could form a new feature space with additional information about objects. Particularly, in our clustering dataset we did not take into account characteristics connected with gamma-ray emission, but blazars emit a large amount of their radiation in gamma-rays, which means the gamma-ray band should carry important information about sources.

Gamma-ray measurements, though, are too scarce to be used in clustering of the whole Roma-BZCAT blazars: the Fermi-LAT gamma-ray flux in the catalog is given for only 28% of sources. New Fermi-LAT measurements (Ajello et al. 2022) do not fundamentally change the situation: we evaluated that now 44% of objects would have gamma-ray fluxes, which is still insufficient (over 60%–70% available data are needed). Nevertheless, using present gamma-ray data, we can still evaluate the degree to which gamma-ray measurements could change obtained clustering results as well as evaluate the influence of dataset incompleteness. To this end, we took only objects with available gamma-ray measurements and calculated for this subsample additional gamma-ray features analogously to our previously described data preparation. The added features are gamma-ray flux, luminosity, and hardness ratios relative to other spectral ranges—a total of seven new features to complement the 14 already available.

After dropping missing values for all 21 features, we end up with a small dataset of 396 sources. Thus we first shorten the list of objects to as few as 11% of the complete sample and second add 50% new features with sufficiently different information concerning the gamma-ray range. To separate the influence of these two effects, we (1) compared results of clustering performed with 14 original

features on the small dataset and on the complete sample; (2) compared results of clustering performed with 14 original features on the small dataset and results obtained on the same dataset with 21 features. The Rand indices for these two comparisons are 0.85 and 0.80, respectively; i.e., 90% incompleteness of the sample could change the result by about 1–0.85 = 15%, while addition of 50% new features preserves it at a level of about 80%.

From these evaluations, we can state that clustering labels for sources should stay the same within 80% of current clustering results if new sufficient data on gamma-ray fluxes become available in the future. We also evaluated the importance of features when taken along with the gamma-ray range; the result is presented in Table 4. As one can see, gamma-ray features occupy the upper rows of the table, thus demonstrating that the gamma-ray range is important for blazar classification, and new more abundant measurements would lead to better, more accurate results.

8.1. Comparison with Known Classes

First, it is interesting to compare our clusters with known types of blazars. Here we make such a comparison for Roma-BZCAT blazar types, HSP blazars from the 3HSP catalog, and blazars detected in the TeV energy range from the TeVCat catalog.

In the Roma-BZCAT catalog, blazars are divided into the following subtypes:

- **BZB**: BL Lac objects and BL Lac candidates, which are AGNs with featureless optical spectra or having only absorption lines of host galaxy origin and weak narrow emission lines;
- **BZG**: sources usually reported in literature as BL Lac objects but having SEDs with significant dominance of host galaxy emission;
- **BZQ**: FSRQs with optical spectra showing broad emission lines and dominant blazar characteristics;
- **BZU**: blazars of uncertain type, a small number of sources having peculiar characteristics but also exhibiting blazar activity: occasional presence/absence of broad emission lines or other features, transition between a radio galaxy and a BL Lac, galaxies hosting a low-luminosity blazar nucleus, etc.

In Table 5 and Figure 12 we compare the population of obtained clusters with subtypes of blazars in Roma-BZCAT. The vast majority of BL Lacs and BZGs fall into clusters 0 and 1. Clusters 3 and 4, conversely, are dominated by FSRQs. Cluster 2 is a mixture of BL Lacs and FSRQs. Blazars of uncertain type avoid cluster 0 and are less present in clusters 3 and 4. It is noteworthy that the largest number of them are in cluster 2, a mixture of BL Lacs and FSRQs, although a comparable number is found in cluster 1.

Judging by quality metrics obtained earlier, we assign blazars to particular clusters with an accuracy of about 90%; therefore a small number of “opposite”

types in individual clusters, with the exception of cluster 2, can be considered expected. Taking into consideration the correlated continuous decrease/increase of BL Lacs and FSRQs among clusters, it could also be a real effect to some degree. In total, we can state that clustering results largely correlate with classification of blazars in the Roma-BZCAT catalog. At the same time, our clustering additionally distinguishes between two subclasses of BL Lacs (clusters 0 and 1) and two subclasses of FSRQs (clusters 3 and 4). There is also no division into BL Lacs and galaxy-dominated BL Lacs, although the almost complete absence of the latter in "mixed" cluster 2 could be noted.

Blazars are classified as a separate type of AGN since they have a distinct orientation of the jet, pointing toward the observer at a small angle. As with other AGNs, they have similar structure (e.g., supermassive black hole, accretion disk, jet) and similar processes (collimation of the jet, acceleration of electrons in a magnetic field, accretion of matter onto the central object), but these processes occur under different physical conditions, which causes division of blazars into different subclasses according to observed parameters. Thus, FSRQs have strong emission lines and higher luminosity compared to BL Lacs in almost all frequency ranges, thus demonstrating more abundant fueling matter and consequently different accretion modes. The fact that different blazar types are not isolated in our clusters but demonstrate a continuous per-cluster distribution validates the commonly accepted uniformity of blazar nature. Note also that although we intentionally avoided any predetermined categorical separation such as presence or absence of emission lines, the clustering correlates with the BL Lac/FSRQ classification; this difference in physical conditions can be obtained from other characteristics.

In Figures 13 and 14 we demonstrate how HSP blazars from the 3HSP catalog (Chang et al. 2019) and blazars detected in the TeV energy range from the TeVCat catalog (Wakely & Horan 2008) are distributed within our clusters. Figure 14 clearly shows that almost all HSP blazars are members of cluster 0, and the rest, located in cluster 1, lie closer to the boundary between the two clusters. The TeV blazars are not as concentrated in a particular cluster but have a tendency to be more abundant in BL Lac-populated clusters 0–1 than in FSRQ-populated clusters 3–4. The overall number of TeV blazars is small, and there are only a few found in the latter clusters, so their presence in FSRQ-populated clusters is questionable and may be caused by clustering inaccuracy; at the same time, the descent in number of TeV blazars from cluster 0 to cluster 4 in Figure 13 looks quite smooth.

8.2. Description of the Clusters

To investigate properties of the selected groups, we analyzed differences in distributions of blazar characteristics. These differences are partially demonstrated in Figure 10. Some statistics of the distributions are given in Table 6. Additionally, Figure 15 shows luminosities or absolute magnitudes for different spectral ranges. These values should be used with caution: while radio luminosities are

corrected for different redshifts using radio spectral indices, other values are not. Therefore, for instance, for high-redshift blazars the optical absolute magnitude in the i filter actually corresponds to the UV range in the source's frame of reference. This effect is strong for IR and optical ranges but less significant for X-ray and gamma-ray luminosities as they are measured in broad bands and anyway stay within corresponding electromagnetic ranges at any redshift, though drifting to higher frequencies. A better understanding of z-dependent bias may be obtained from Figure 16, which shows rest-frame average SEDs for the resulting clusters.

To construct average SEDs, we normalized SEDs of individual blazars by measured synchrotron peak flux density (Section 4, Figure 3) and averaged flux densities in 50 bins. Error bars in Figure 16 correspond to standard deviation within bins. To demonstrate differences in luminosities, normalized spectra are additionally adjusted to radio luminosity at 5 GHz ($\log_{10}$ L_5). The SEDs are recalculated to the rest frame.

To better visualize differences between cluster statistics demonstrated in Figures 10 and 15 and Table 6, we also constructed polar diagrams shown in Figure 17. The figure reflects differences between median values of various characteristics: polar diagrams are scaled such that maximum observed medians over all clusters correspond to values of 1 (outer edges of circles), while minimum medians correspond to zero values (centers of circles). Each "azimuth" corresponds to a particular characteristic.

Before describing average SEDs, we should mention the following. The SED of a blazar typically has a shape of two humps and constitutes a complex mix of emission from different parts of an AGN. The first hump, extending from radio waves to X-rays, is believed to be formed by synchrotron radiation in the jet. Part of photons emitted in this process may experience synchrotron self-Compton scattering (e.g., Bloom & Marscher 1996), contributing to the second, gamma-ray, hump. Photons from other AGN components such as the accretion disk, dust torus, and broad emission line clouds also contribute to the gamma-ray hump via inverse Compton scattering. The location of this gamma-ray emission is still a subject of research (e.g., Rani et al. 2016). Additionally, the accretion disk emits its own thermal radiation peaked in optical–UV ranges, while the dust torus adds to IR emission. The corona of the accretion disk can also scatter photons up to X-ray energies. Finally, the SED may be affected by the host galaxy.

All this complexity means that detailed description can only be made for a particular SED via complex modeling and/or analysis of its variability time series for different ranges of the electromagnetic spectrum. Nevertheless, in the case of blazars we have an advantage: their jets are inclined with respect to the line of sight at a small angle, therefore we can expect that differences in average SEDs are caused to a greater degree not by geometric effects but by different physical conditions in the AGN, whatever these conditions are.

8.2.1. Clusters 0 and 1: BL Lac Subclasses

These two clusters consist of BL Lacs and galaxy-dominated BL Lacs located at relatively small distances (up to 3 Gpc, z 0.9). The percentage of FSRQs is only 2% and 14% respectively (see Table 5). Cluster 0 blazars are distinguished by relatively reduced luminosity across the entire electromagnetic spectrum (Table 6, Figures 15 and 16). Also, they have significantly reduced radio hardness parameters (Table 6, Figure 10). Additionally, these objects have lower gamma-ray luminosities.

Cluster 0 has the most prominent characteristics. From Figure 16 we can see that both synchrotron and gamma-ray humps are factually not visible in the average SED. By reviewing some individual SEDs in the cluster, we found that they actually have a standard shape with two humps. Therefore, this effect in the average SED is caused by the broad distribution of synchrotron peaks seen in Figure 10.

The average SED of cluster 1 has the classical shape of two humps, though both humps are not particularly prominent. The difference in shape from cluster 0 is likely caused by a more compact distribution of synchrotron frequencies (Figure 10). This broad variation in synchrotron peak frequency, especially noticeable in cluster 0 (see Figure 10) across the entire range of $\log_{10}$ _{peak} 11.7–18.4, is a characteristic peculiarity of these clusters; in other clusters the distributions are more compact, and high synchrotron peak frequencies are practically not found. One can notice from Figure 10 that almost all HSP blazars should belong to cluster 0. Earlier in Section 8.1 we considered distribution of sources from the 3HSP catalog across our clusters and confirmed that 529 of 657 3HSP blazars present in Roma-BZCAT belong to cluster 0, and 118 to cluster 1, while only 10 are found in clusters 2, 3, and 4 (see Figure 13). Note also that 3HSP blazars from cluster 1 are located close to the boundary between clusters 0 and 1.

Within z 0.9, where blazars of clusters 0 and 1 are located, the actual frequency emitted by a blazar becomes higher for large z, but all frequencies used in the clustering feature space remain within their electromagnetic bands: 3.35 m 1.76 m (IR), 7520 Å 3960 Å (optical), 2316 Å 1220 Å (UV) for the most distant objects. Radio, UV, X-ray, and gamma-ray frequencies remain within their bands for all z considered in this paper. For average SEDs, flux densities were transformed to the rest frame beforehand.

8.2.2. Cluster 2: BL Lac–FSRQ Mix

Cluster 2 is represented by a mixture of BL Lac-type objects (29% including BL Lac candidates) and FSRQs (60%), with the remaining 11% being of uncertain type (Table 5). It practically does not contain galaxy-dominated BL Lacs; a small number (10 objects) can be attributed to clustering errors.

In contrast to clusters 0 and 1, in cluster 2 we observe high radio, X-ray, and gamma-ray luminosities as well as bright absolute magnitudes comparable to

other FSRQs (Figures 15 and 16). Absolute magnitudes in optical and UV ranges are somewhat weakened compared to other FSRQs. In this connection, it is of interest to evaluate whether BL Lacs in cluster 2 are different from those in clusters 0 and 1, or whether statistically higher luminosities are only related to presence of FSRQs. In Figure 18 we compare radio luminosity distributions for BL Lacs, and the observed difference clearly testifies that BL Lacs from cluster 2 form a special BL Lac subclass with high luminosity.

The average SED of this cluster in Figure 16 shows the most smooth shape of classical two humps. The distributions and statistics in Figures 10 and 15 and Table 6 naturally follow from this shape.

Within cluster 2, blazars span z 0.05–2.5; IR radiation, when converted to the source rest frame, remains generally within the IR range, approaching optical wavelengths for the most distant objects: 3.35 m 0.96 m; optical radiation moves into the UV range with growing distance: 7520 Å 2150 Å; UV radiation becomes harder: 2316 Å 660 Å. Again, average SEDs in Figure 16 are calculated in the rest frame and do not suffer from differing redshifts.

8.2.3. Clusters 3 and 4: FSRQ Subclasses

Clusters 3 and 4 are populated by FSRQs: 85% and 94% respectively, or 91% and 97% if we exclude blazars of uncertain type (see Table 5). Blazars from these clusters have high luminosities across the entire frequency range. The main difference between clusters 3 and 4 and the above-described cluster 2, as can be seen from average spectra in Figure 16, is the degree of irregularities in the two-hump SED shape. While in cluster 2 we observe a smooth spectrum, in cluster 3 the synchrotron hump becomes somewhat flattened due to enhanced emission at frequencies $\log_{10}$ > 14.5 in the rest frame, and in cluster 4 the average SED obtains a step-like shape. Statistical characteristics in Figures 10 and 15 and Table 6 reflect the shape of average SEDs. The source of observed irregularities may be excessive emission from central parts of AGNs.

Blazars from cluster 3 have sufficiently lower radio hardness parameters (Figure 10). Cluster 4 demonstrates noticeably higher median luminosities in radio, X-ray, and gamma-ray ranges; statistically, this cluster contains the most luminous objects at higher redshifts, though we do not consider a few individual objects in cluster 2 with the greatest redshift and gamma-ray luminosity.

Distances range from 500 to 6500 Mpc (z 0.05–2.5) for cluster 3 and from 2000 to 8000 Mpc (z 0.4–4) for cluster 4. For $z = 4$ the frequency shifts in the rest frame are: IR radiation goes into optical range 3.35 m 6700 Å; optical radiation goes into UV range: 7520 Å 1500 Å; UV radiation becomes harder: 2316 Å 460 Å.

9. Summary

In this paper we discuss applications of cluster analysis techniques to multi-wavelength properties of blazars from the Roma-BZCAT catalog. We divided blazars into five groups and compared them with Roma-BZCAT classification, HSP blazars from the 3HSP catalog, and TeV blazars from TeVCat. We found similar trends in blazar grouping, which confirms the effectiveness of clustering techniques. The obtained groups (clusters) are derived based on multiparametric distributions of blazar characteristics.

To perform the project, we collected data from radio to gamma-ray ranges both from the Roma-BZCAT catalog itself and from various other point-source catalogs, mostly those containing sufficient data for the sample. During clustering, blazars were treated uniformly regardless of our degree of knowledge about them—e.g., we did not add additional measurements from other catalogs for some well-known objects, thus preserving a homogeneous approach to the sample as a whole.

In general, clustering algorithms build an independent unsupervised classification based almost solely on multiple properties of objects under consideration. In this sense, clustering is a more uniform and homogeneous approach to classifying cosmic objects based on experimental data, avoiding subjective selection bias. The method nevertheless has its own hyperparameters (parameters set by the researcher rather than learned by the model from data): the feature space, which determines characteristics relevant for the scientific scope; the algorithm implemented to find clusters; and the number of clusters—a trade-off between uniformity and individuality.

For the feature space we used the maximum available number of characteristics that could be related to properties of objects. Such an approach helps describe blazars as completely as possible. The process of feature selection is described in detail in Section 5. In total, the model feature space comprises 14 continuously distributed characteristics.

Although adding new parameters (for instance, due to growing numbers of observations) will inevitably change clustering results for particular objects, especially on cluster boundaries, overall patterns observed in the sample will be preserved, unless new characteristics are comparable in number to those in the original feature space. This allows us to discuss sample properties with some robustness, which is confirmed by comparison with known blazar classifications. Cluster membership of any particular objects, though, must be considered with great caution and additional analysis. Moreover, as no localized groups are revealed in feature space and due to incompleteness of the Roma-BZCAT catalog, boundaries between clusters are only conditional and might change for a more complete sample. We evaluated the influence of sample incompleteness and feature selection on clustering results in Section 7.3 and expect that our cluster labels will be preserved with a Rand index of 80% after adding a sufficiently large amount of new information.

We tested several clustering algorithms and finally settled on two: PCA + k-means and SOMs. The advantage of the latter is that SOMs can restore possible nonlinearities in data distribution, while PCA dimensionality reduction is a more straightforward and interpretable linear algebra method. By showing 90% similarity of their results, we demonstrate the absence of nonlinearities in our data as well as some robustness of feature space division into clusters: although final stages use the k-means algorithm in both cases, they work in sufficiently different spaces—a 14D space of neuron weights for SOMs and a 6D space of PCA components for the other method. As methods in our case showed similar results, we followed PCA + k-means for interpretation clarity.

The number of clusters is a hyperparameter that can be set rather loosely in clustering problems. Generally, addition of a new cluster leads to division of an existing one into two subclasses. In the case of continuous data distribution, cluster boundaries may also vary. We chose the number of clusters to be five based on best match between data distributions within clusters obtained for the subsample without missing values and for the whole sample where missing values were imputed via PPCA.

We found PPCA to be most effective for data imputation among other methods: imputation with medians or various ML regressions. Note that these imputed values were used only to perform clustering for the complete sample; we did not use them for statistical analysis of derived clusters.

The following notable characteristics of clusters have been derived:

Cluster 0: Consists of BL Lac-type blazars with low luminosities in all ranges from radio to gamma-rays except X-ray emission. Almost all known HSP blazars fall into this cluster. The synchrotron hump is not visible in the average SED, nor is the second hump in the gamma-ray range; this effect is caused by broad distribution of synchrotron peak frequencies in the cluster. The cluster is characterized by low radio luminosity, radio hardness parameters, and redshifts z 0.9.

Cluster 1: Consists of BL Lac-type blazars with low luminosities in all ranges from radio to gamma-rays. The average SED has the usual shape with two humps, but the gamma-ray hump is weaker than usual and comparable in flux density to the synchrotron hump. These blazars have low radio luminosities and radio hardness parameters, and redshifts z 0.9.

Cluster 2: A mix of BL Lac-type objects (29% including BL Lac candidates) and FSRQs (60%). The remaining objects (11%) are of uncertain type. These blazars have high luminosities, strong radio hardness parameters, and a clear smooth average SED with two humps. Redshifts span the entire range. We show that BL Lacs from this cluster form a special subclass of high-luminosity BL Lacs compared to the low-luminosity population in clusters 0 and 1.

Clusters 3 and 4: FSRQs. These blazars have high luminosities. The clusters are primarily distinguished by degree of irregularities in SED shape that may be

caused by influence of emission from AGN central parts. Blazars from cluster 3 demonstrate statistically lower radio hardness parameters compared to FSRQs from clusters 2 and 4. Cluster 4 shows noticeably higher median luminosities in radio, X-ray, and gamma-ray ranges; this cluster contains, on average, the most luminous objects at higher redshifts, although individual record holders in redshift and gamma-ray luminosity fall into cluster 2.

Our results are consistent with the term "blazar sequence" that originated in Fossati et al. (1998) to describe properties of blazar SEDs. The most well-known feature of this phenomenological sequence is the negative correlation between synchrotron peak frequencies and synchrotron peak luminosities of the blazar population—i.e., HSP blazars have the lowest luminosities and highest synchrotron peak frequencies, while LSP blazars show opposite characteristics. However, that anti-correlation was not found in intrinsic blazar properties after correcting observed data for Doppler beaming effects, and another study by Giommi et al. (2012a) also showed that the originally reported anti-correlation was due to a selection effect. Following works based on various multiband data confirmed that emission in blazars is strongly beamed and affects the observational phenomenon known as the "blazar sequence" (e.g., Fan et al. 2017; Ouyang et al. 2023; Wan et al. 2024). In our study we did not apply the Doppler factor as a parameter because it is estimated for a limited number of blazars—e.g., for 979 Fermi blazars in one recent study (Chen et al. 2024). Therefore, we cannot test the intrinsic nature of the blazar sequence.

Keenan et al. (2021) studied a dichotomy in jets, dividing more than 2000 blazars into two samples: one with inefficient accretion (weak/type I jets) and the second with efficient accretion (strong/type II jets). The first group contained blazars with synchrotron peak frequencies above 10^{15} Hz (HSPs, nearly all BL Lacs), and the second comprised mostly FSRQs and some LSP BL Lacs. This quite accurately coincides with our findings both in synchrotron peak frequency values and blazar distribution—i.e., clusters 0 and 1 contain blazars with type I jets, and clusters 3 and 4 are blazars with type II jets.

We believe that these groups of Roma-BZCAT blazars, derived from multiparametric analysis, can be used as additional information for further research, for example in searches for correlation with neutrino events or other statistical investigations.

The dataset with various characteristics of blazars and cluster labels is available in the VizieR database.

Acknowledgments

We thank an anonymous referee, whose valuable comments helped improve the paper. This study was funded by the Ministry of Science and Higher Education of the Russian Federation under contract 075-15-2022-1227. The research has made use of the Roma-BZCAT blazar catalog and references therein, Astrophysical CATalogs support System (CATS), NASA/IPAC Infrared Science Archive,

NASA/IPAC Extragalactic Database (NED), Barbara A. Mikulski Archive for Space Telescopes (MAST), SED Builder, Sloan Digital Sky Survey (SDSS), and Two Micron All Sky Survey (2MASS). The publication has made use of data products from the Wide-field Infrared Survey Explorer (WISE), Panoramic Survey Telescope and Rapid Response System (Pan-STARRS), Galaxy Evolution Explorer (GALEX), RATAN-600, FIRST, Planck, ROSAT, Swift-XRT, Fermi, and numerous other telescopes whose observed data have been compiled in Roma-BZCAT, CATS, and SED Builder. We used a set of Python libraries mentioned in the text.

Dmitry O. Kudryavtsev ORCID iDs: https://orcid.org/0000-0003-

References

Abdo, A. A., Ackermann, M., Agudo, I., et al. 2010, ApJ, 716, 30
Ajello, M., Baldini, L., Ballet, J., et al. 2022, ApJS, 263, 24
Aller, M. F., Aller, H. D., & Hughes, P. A. 1992, ApJ, 399, 16
Arthur, D., & Vassilvitskii, S. 2007, in Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms, 1027 (Philadelphia, PA: Society for Industrial and Applied Mathematics)
Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Belladitta, S., Moretti, A., Caccianiga, A., et al. 2020, A&A, 635, L7
Bengfort, B., & Bilbro, R. 2019, JOSS, 4, 1075
Blandford, R., Meier, D., & Readhead, A. 2019, ARA&A, 57, 467
Bloom, S. D., & Marscher, A. P. 1996, ApJ, 461, 657
Böttcher, M. 2019, Galaxies, 7, 20
Butkevich, A. G., Berdyugin, A. V., & Teerikorpi, P. 2005, MNRAS, 362, 321
Caliński, T., & Harabasz, J. 1974, Communications in Statistics-Simulation and Computation, 3, 1
Chang, Y. L., Arsioli, B., Giommi, P., Padovani, P., & Brandt, C. H. 2019, A&A, 632, A77
Chen, G., Zheng, Z., Zeng, X., et al. 2024, ApJS, 271, 20
Chen, T., & Guestrin, C. 2016, Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD '16 (New York: ACM), 785
Chilingarian, I. V., Melchior, A.-L., & Zolotukhin, I. Y. 2010, MNRAS, 405, 1409
Cohen, M., Wheaton, W. A., & Megeath, S. T. 2003, AJ, 126, 1090
Davies, D. L., & Bouldin, D. W. 1979, IEEE Trans. Pattern Anal. Mach. Intell., 1, 224
Falomo, R., Pian, E., & Treves, A. 2014, A&ARv, 22, 73
Fan, J. H., Yang, J. H., Liu, Y., et al. 2016, ApJS, 226, 20
Fan, J. H., Yang, J. H., Xiao, H. B., et al. 2017, ApJL, 835, L38
Fitzpatrick, E. L. 1999, PASP, 111, 63
Fossati, G., Maraschi, L., Celotti, A., Comastri, A., & Ghisellini, G. 1998, MNRAS, 299, 433

Ghisellini, G., Padovani, P., Celotti, A., & Maraschi, L. 1993, ApJ, 407, 65
Ghisellini, G., Tavecchio, F., Foschini, L., & Ghirlanda, G. 2011, MNRAS, 414, 2674
Giommi, P., & Padovani, P. 2015, MNRAS, 450, 2404
Giommi, P., Padovani, P., Polenta, G., et al. 2012a, MNRAS, 420, 2899
Giommi, P., Polenta, G., Lähteenmäki, A., et al. 2012b, A&A, 541, A160
Giommi, P., Padovani, P., & Polenta, G. 2013, MNRAS, 431, 1914
Hervet, O., Boisson, C., & Sol, H. 2016, A&A, 592, A22
Jarrett, T. H., Cohen, M., Masci, F., et al. 2011, ApJ, 735, 112
Keenan, M., Meyer, E. T., Georganopoulos, M., Reddy, K., & French, O. J. 2021, MNRAS, 505, 4726
Khabibullina, M., Mikhailov, A., Sotnikova, Y., et al. 2023, AstBu, 78, 443
Kohonen, T. 2001, Self-Organizing Maps (Berlin: Springer)
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Statist., 22, 79
Landt, H., Padovani, P., Perlman, E. S., & Giommi, P. 2004, MNRAS, 351, 83
Madau, P., Ghisellini, G., & Persic, M. 1987, MNRAS, 224, 257
Massaro, E., Giommi, P., Leto, C., et al. 2009, A&A, 495, 691
Massaro, E., Maselli, A., Leto, C., et al. 2015, Ap&SS, 357, 75
Mingaliev, M. G., Sotnikova, Y. V., Udovitskiy, R. Y., et al. 2014, A&A, 572, A59
Morrissey, P., Conrow, T., Barlow, T. A., et al. 2007, ApJS, 173, 682
Nieppola, E., Tornikoski, M., Lähteenmäki, A., et al. 2007, AJ, 133, 1947
Nieppola, E., Valtaoja, E., Tornikoski, M., Hovatta, T., & Kotiranta, M. 2008, A&A, 488, 867
Ouyang, Z., Xiao, H., Chen, J., et al. 2023, ApJ, 949, 52
Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, A&ARv, 25, 2
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Planck Collaboration, Ade, P. A. R., Aghanim, N., et al. 2016, A&A, 594, A13
Plavin, A., Kovalev, Y. Y., Kovalev, Y. A., & Troitsky, S. 2020, ApJ, 894, 101
Porta, J., Verbeek, J., & Krose, B. 2005, AuRob, 18, 59
Prandini, E., & Ghisellini, G. 2022, Galaxies, 10, 35
Raiteri, C. M., Villata, M., Carnerero, M. I., et al. 2014, MNRAS, 442, 629
Rand, W. 1971, J. Am. Stat. Assoc., 66, 846
Rani, B., Krichbaum, T. P., Hodgson, J. A., & Zensus, J. A. 2016, JPhCS, 718, 052009
Rousseeuw, P. 1987, JCoAM, 20, 53
Sbarrato, T., Ghisellini, G., Maraschi, L., & Colpi, M. 2012, MNRAS, 421, 1764
Sotnikova, Y. V., Mufakharov, T. V., Mikhailov, A. G., et al. 2022, AstBu, 77, 246
Tipping, M. E., & Bishop, C. M. 1999, Neural Comput., 11, 443
Tonry, J. L., Stubbs, C. W., Lykke, K. R., et al. 2012, ApJ, 750, 99
Tornikoski, M., Lainela, M., & Valtaoja, E. 2000, AJ, 120, 2278
Urry, C. M. 1999, APh, 11, 159
Urry, C. M., & Padovani, P. 1995, PASP, 107, 803
van der Maaten, L., & Hinton, G. 2008, JMLR, 9, 2579
Verkhodanov, O. V., Trushkin, S. A., Andernach, H., & Chernenkov, V. N.

2005, BSAO, 58, 118
Verkhodanov, O. V., Trushkin, S. A., & Chernenkov, V. N. 1997, BaltA, 6, 275
Wakely, S. P., & Horan, D. 2008, ICRC, 3, 1341
Wan, Z.-J., Xue, R., Wang, Z.-R., Xiao, H.-B., & Fan, J.-H. 2024, MNRAS, 528, 7529
Wittek, P., Gao, S. C., Lim, I. S., & Zhao, L. 2017, J. Stat. Soft., 78, 1

Note: Figure translations are in progress. See original paper for figures.