HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Discovering Extragalactic Supermassive Black Holes with Multiwavelength Data Analysis

Krish Singh

The International School Bangalore, India

ABSTRACT

Differentiating extragalactic and galactic sources of light allows for the identification of supermassive black holes for further study. Using the newest X-ray dataset from eROSITA, a machine learning approach helps classify extragalactic and galactic sources. It was found that a Gaussian Mixture Model was effective at this classification, achieving a silhouette score of 0.84. This result shows that a Gaussian Mixture Model is suitable for tasks like this, working toward discovering more supermassive black holes.

Introduction

Black holes are interesting and poorly understood objects. Correctly identifying them will give us a much larger set to study. The goal of this paper is to use machine learning techniques to distinguish galaxies from objects within the Milky Way. These extragalactic objects contain supermassive black holes, which can then be further studied. Black holes are some of the most extreme objects in the Universe. Studying them will allow us to test the most extreme physics.

This paper uses a novel approach, combining data from multiple telescopes. These operate at different wavelengths, including infrared, visible light, and X-rays. Data from multiple telescopes was combined to form a training dataset for a machine learning model. Various models were applied to this dataset to classify them as extragalactic or galactic. The results were compared with existing classifications.

Supermassive black holes emit in all these wavelengths (Padovani, 2017). As matter falls into the deep gravitational potential of these black holes, dust is ejected, leading to infrared emission. The material itself is subject to strong friction, which leads to optical, ultraviolet and X-ray emission. Finally, various mechanisms have been proposed to explain relativistic jets which are ejected from the black hole and are sources of radio emission.

Data Sources

The dataset was assembled from a range of telescopes. It covers 27369 objects in a small area of the sky. In astronomy, right ascension and declination are coordinates used to define a region of the sky. The area of the dataset ranges from 125° to 145° right ascension (RA) and -3° to 6° declination (DEC), as seen below. This represents about 0.25% of the entire sky.

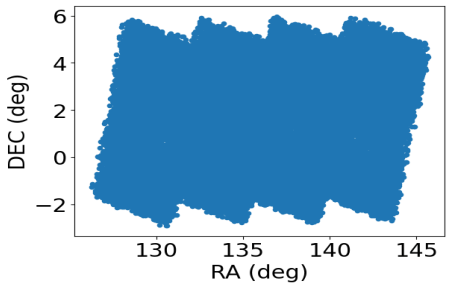

Figure 1. Graph of the region of the sky covered by the dataset

The telescopes used included the Spektr-RG (SRG), a space-based X-ray telescope. It contains the instrument eROSITA which was designed by the European Space Agency (ESA). The SRG telescope was launched in 2019. The data used is from the eFEDS survey (Brunner et al., 2022).

This dataset was combined with optical spectroscopy from various surveys, most importantly the Sloan Digital Sky Survey (York et al. 2000) This allows us to obtain redshift values, as explained in the next section. Gaia is a space observatory operated by the ESA. It can detect optical light. In addition, it provides parallax information for astronomical objects, which can be used to find their distance from Earth. Gaia was launched in 2013.

The Galaxy Evolution Explorer (GALEX) was a telescope in orbit. It was operated by the National Aeronautics and Space Administration (NASA) from 2003 until 2013. GALEX observed in ultraviolet wavelengths. It also collected data on the redshift of astronomical objects, which can indicate their age.

The Wide-field Infrared Survey Explorer (WISE) is an orbital telescope. It was launched by NASA in 2009. It observes in medium infrared wavelengths.

The Visible and Infrared Survey Telescope for Astronomy (VISTA) is a ground-based telescope that observes in near-infrared wavelengths. It began operating in 2010.

The Very Large Array (VLA) is a ground-based array of radio telescopes. It is operated by the National Science Foundation (NSF). The VLA has been operating since 1980.

The Dark Energy Camera Legacy Survey (DECaLS) was conducted in 2016 by the Blanco telescope in Chile. It observed at 3 different wavelengths of optical light.

Most of these data sources have started operating recently. This paper is the first attempt to combine the data from all of these sources for a multiwavelength analysis. Most of these telescopes survey the entire sky, with the exception of SRG, which only observes the area in the dataset. SRG is the newest X-ray telescope. This is the reason for choosing to work with this particular region of the sky.

Data Visualization

Before beginning with modeling, a variety of data visualizations were used to identify the best way to approach this dataset. A few of the most important visualizations are included below.

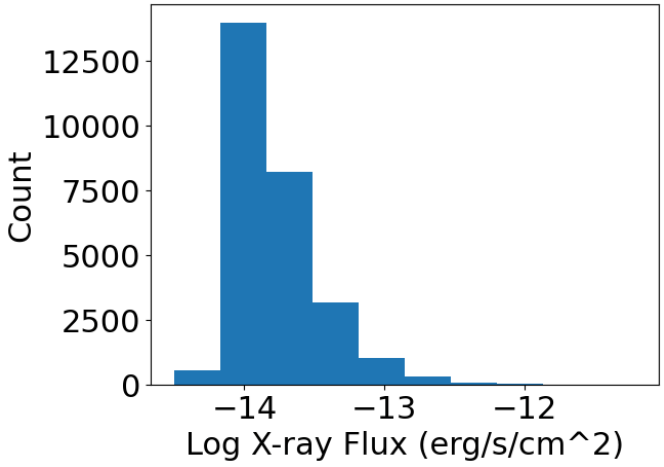

Figure 2. Histogram of X-ray flux with logarithmic scale (lower values indicate fainter objects)

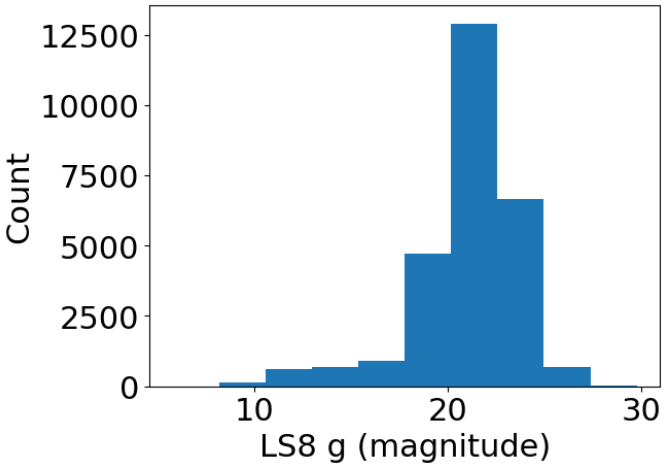

Figure 3. Histogram of magnitude of green visible light (higher values indicate fainter objects)

Figure 2 and 3 show similar features, illustrating a property of the dataset. The majority of astronomical objects in this dataset are dim, with only a few brighter ones. This leads to a tapering tail towards the high-energy part of the histograms above. The maximum value of magnitude and the minimum x-ray flux indicate the limit of the sensitivity of the telescopes used.

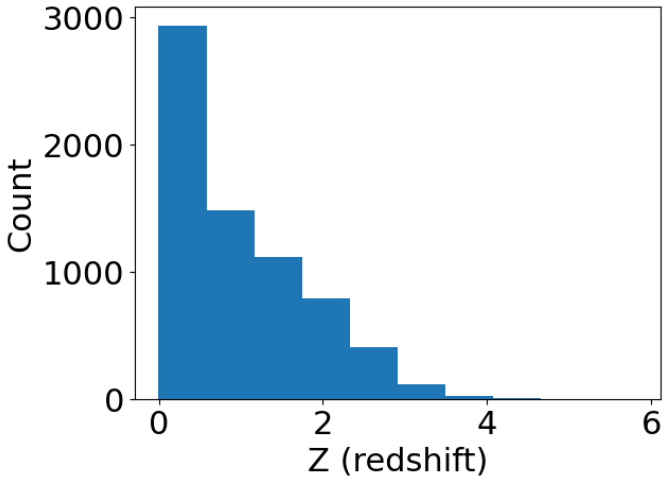

Figure 4. Histogram of redshift (a redshift of 2 indicates the peak of star formation in the history of the Universe; Williams et al., 1996)

Redshift is a measurement in astronomy that is used to measure how far away and how old an observation is. One of the greatest achievements of astronomy was the discovery of the Hubble law (1929): farther galaxies move faster away from us. This means that 1) the Universe is expanding and 2) that redshift can be used as both a velocity and age indicator.

Since the universe is expanding, light that comes towards telescopes is stretched. This makes its wavelength increase into the red color region. This stretched light is said to be "redshifted". The extent of the redshift indicates how far away the source of the light is. Light that comes from farther away has taken longer to arrive, so redshift also indicates the age of the observation.

The redshift values indicate that these data points cover a large portion of the Universe's history. The oldest object in this dataset is from less than 1 billion years after the Big Bang. The Universe is about 13.8 billion years old. This object will later be shown to be an extragalactic supermassive black hole. This shows that supermassive black holes formed even in the early Universe.

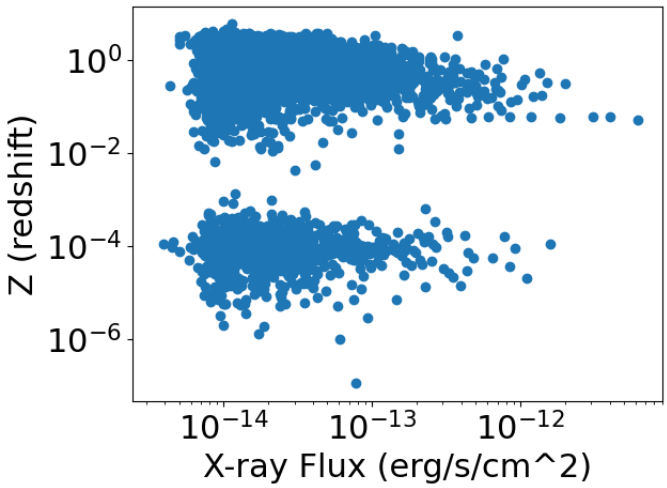

Figure 5. Plot of logarithmic X-ray flux against redshift

In Figure 5, clustering is clearly visible. On the y-axis, there is redshift, while the x-axis has X-ray flux. This plot shows clearly that there are two groups of objects. The lower group is nearer to Earth, so it is likely made up of objects within the Milky Way. The upper group is farther away, showing that these are extragalactic objects. Most of these are likely supermassive black holes due to their brightness and the type of radiation they are emitting.

Methods

The multiwavelength dataset was compiled through a crossmatch of X-ray positions with the best-match optical, infrared, UV, and radio positions. Various surveys were used including the SDSS, LS8, WISE, GALEX, and VLASS surveys.

Because of the pristine X-ray positions obtained from this new instrument (5 arcseconds), most sources contained at only one or two optical matches. In cases when there were multiple matches, an algorithm based on Bayesian statistics was used to determine the most likely match. Further description of the matching process is beyond the scope of this work, but the catalog is freely available with a description of this matching (see references). Only rows with no null values were used in the data analysis. This resulted in 6542 objects remaining from the initial dataset.

The dimensionality of the data was reduced by focusing on two dimensions: X-ray flux and redshift. After seeing the results of the data visualization, it was clear that the X-ray flux against redshift space was best for analysis.

Clustering Techniques

To cluster the data points, two algorithms were used. The first was DBSCAN, and the second was a Gaussian Mixture Model. DBSCAN is a simple clustering algorithm free of any assumptions about the number of clusters. This made it a suitable first choice to begin the analysis. A Gaussian Mixture Model is a clustering algorithm that models the data points as a mixture of multiple normal distributions (Reynolds, 2009). The data points are assigned to a cluster probabilistically based on which normal distribution they fall under. The number of clusters must be specified beforehand for the Gaussian Mixture Model.

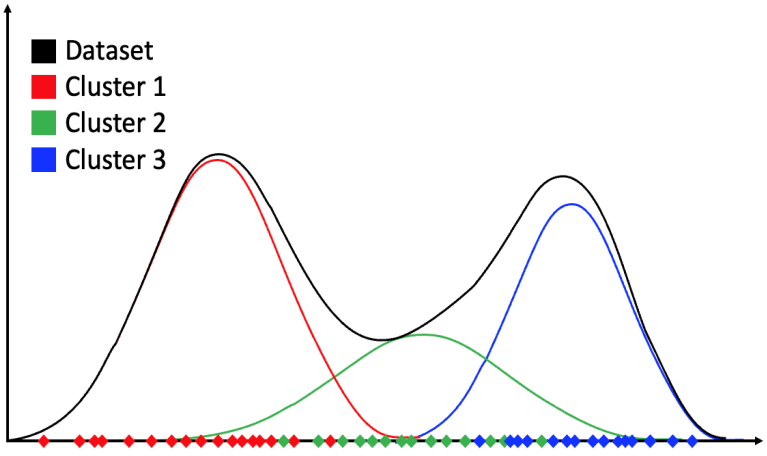

Figure 6. Visualization of 1-dimensional Gaussian Mixture Model (own work)

Results from the Gaussian Mixture Model were favored over DBSCAN. This was because normal distributions are common in astronomy, so the Gaussian Mixture Model has a solid mathematical foundation.

Silhouette Score

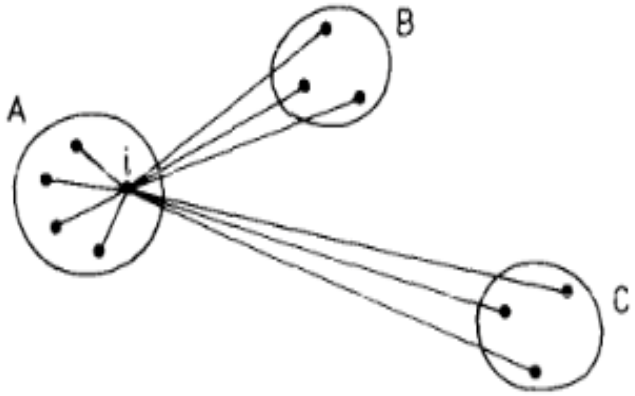

Figure 7. Visualization of silhouette score for a point i (Rousseeuw, 1986)

The silhouette score is a metric to evaluate the quality of clusters that will be used to evaluate the performance of the model. The average distance to all points in the same cluster is called a. The average distance to the points of the next nearest cluster is called b. The silhouette score is found as

Equation 1: Calculation of silhouette score

$$\frac{b-a}{\max(a,b)}$$

Values of the silhouette score close to 1 indicate the best clustering. Values near 0 indicate weak clustering. If the silhouette score is negative, it indicates a certain point should have been assigned to a different cluster.

Results

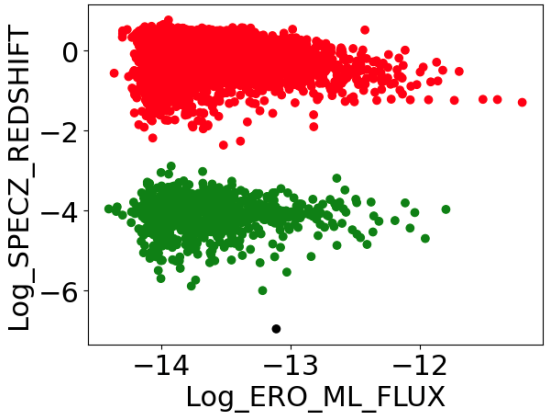

Figure 8. Results of DBSCAN algorithm

The DBSCAN algorithm was run on the data points. It determined that there were 3 clusters. They are represented in Figure 8 by red, green, and black (only one point at the bottom is black).

The individual clusters were further examined. The mean value of x-ray flux in each cluster was about $10^{-13.6}$, which closely matched the overall mean of the dataset. On the other hand, the mean redshift for the top cluster was $10^{-0.1}$, while mean redshift was $10^{-4.1}$ for the bottom cluster. The lone point in the third cluster had $10^{-7.0}$ redshift.

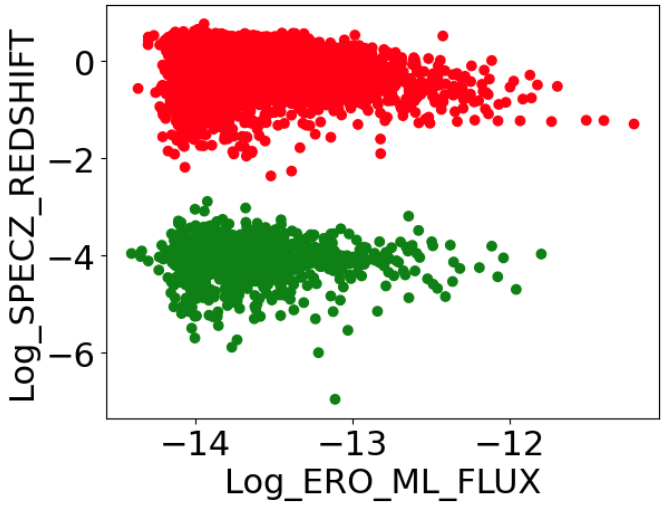

Figure 9. Results of Gaussian Mixture Model

The Gaussian Mixture Model gave similar results to the DBSCAN algorithm. The Gaussian Mixture Model in Figure 9 was told to create 2 clusters. One cluster is red, the other green.

The individual clusters also both had mean x-ray flux of about $10^{-13.6}$. The mean redshift for the mean redshift for the top cluster was $10^{-0.1}$, while mean redshift was $10^{-4.1}$ for the bottom cluster. This closely matched the results from the DBSCAN model, indicating that these results were robust.

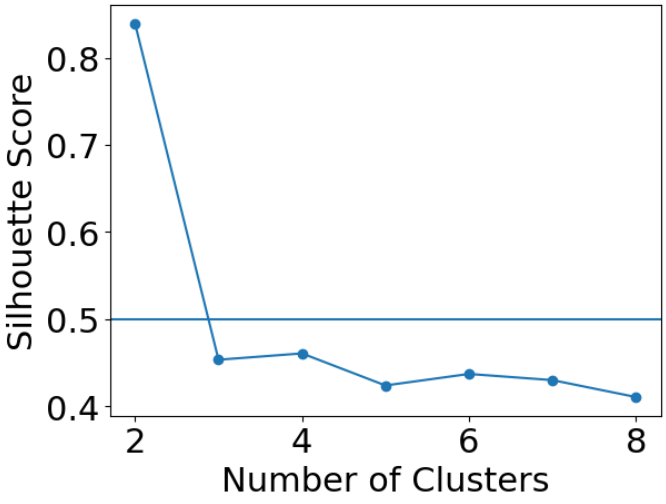

Figure 10. Hyperparameter Tuning of the Gaussian Mixture Model

The Gaussian Mixture Model could take various values, or hyperparameters, for the number of clusters. Although 2 clusters seemed to be qualitatively correct, this was tested with hyperparameter tuning. The results of the Gaussian Mixture Model for various different hyperparameter values were evaluated with the silhouette score, as seen in Figure 10.

Generally, any value of the silhouette score above 0.5 can be said to cluster well (Rousseeuw, 1986). Only 2 clusters satisfied this requirement. The results of the hyperparameter tuning confirmed that 2 was the optimal number of clusters for the Gaussian Mixture Model. The silhouette score for 2 clusters was 0.84, indicating very good clustering.

Discussion

Comparing the number of objects in the Gaussian Mixture Model clusters, the upper extragalactic cluster had 5644 objects, while the lower galactic cluster had 898 objects. More objects were extragalactic, which indicates that the telescopes were looking away from the center of the Milky Way.

The mean redshift for the upper cluster was $10^{-0.1}$, which corresponds to a mean age of 7 billion years. The mean redshift for the lower cluster corresponds to an age of 1 million years. This shows that these observations were considerably closer and more recent.

A redshift of 2 is a key milestone in the history of the Universe, since it represents the time period during the peak of star formation (Williams et al., 1996). The majority of objects have a redshift less than 2. However, the highest value of redshift in the dataset was 5.75. In fact, 987 objects had a redshift above 2. This indicates that black holes formed even during and before the period of large-scale star formation.

In the X-ray flux against redshift space, it is visually clear that there are 2 clusters. This is matched by the grid search conducted during hyperparameter tuning. Along with the high silhouette score, this confirms the accuracy of the clustering.

The extragalactic objects identified in this clustering were likely galaxies. These galaxies should contain a supermassive black hole at their center. Many of these galaxies may have been previously misidentified as stars within the Milky Way. The results of this paper provide a greater sample size to study the behavior of supermassive black holes.

The Gaussian Mixture Model shows promise for doing analysis with more columns and more features. This validates the assumption made earlier, that astronomical data is mostly normally distributed. This paper is the first part of a 2-part series. In the second paper, the XMM-Newton dataset will be used (Jansen et al., 2001). That dataset is about 20 times bigger than the eFEDS survey used in this paper.

Conclusion

Overall, the dataset proved amenable to analysis in the X-ray flux against redshift space. A Gaussian Mixture Model was used to cluster the data points into an extragalactic and galactic cluster. A grid search was used to confirm that 2 clusters was the optimal number. The results were that 5644 objects in the dataset were extragalactic, and 898 were galactic. The mean age of extragalactic observations was 6 billion years while it was 1 million years for galactic observations. The silhouette score of the clustering was 0.84. The Gaussian Mixture Model proved an effective tool to analyze astronomical data.

Acknowledgments

The author would like to acknowledge Antonio Rodriguez for his mentorship during the research process. His help in collecting data and understanding the theory behind black holes was invaluable.

In addition, the author would like to acknowledge the Inspirit AI Research Scholars program for making this project possible.

References

1. Brunner, H., et al. "The eROSITA Final Equatorial Depth Survey (eFEDS)-X-ray catalogue." *Astronomy & Astrophysics* 661 (2022): A1, https://doi.org/10.1051/0004-6361/202142460.
2. "eROSITA-DE: Early Data Release site." *eROSITA*, https://erosita.mpe.mpg.de/edr/eROSITAObservations/Catalogues/. Accessed 8 November 2023.
3. Hubble, Edwin. "A relation between distance and radial velocity among extra-galactic nebulae." *Proceedings of the national academy of sciences* 15.3 (1929): 168-173, https://doi.org/10.1073/pnas.15.3.168.
4. Jansen, F., et al. "XMM-Newton observatory-I. The spacecraft and operations." *Astronomy & Astrophysics* 365.1 (2001): L1-L6, http://dx.doi.org/10.1051/0004-6361:20000036.
5. Padovani, Paolo. "Active galactic nuclei at all wavelengths and from all angles." *Frontiers in Astronomy and Space Sciences* 4 (2017): pp. 35, https://doi.org/10.3389/fspas.2017.00035.
6. Reynolds, Douglas A. "Gaussian mixture models." *Encyclopedia of biometrics* 741.659-663 (2009).
7. Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): pp. 53-65, https://doi.org/10.1016/0377-0427(87)90125-7.
8. Williams, Robert E., et al. "The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry." *Astronomical Journal v. 112, p. 1335* 112 (1996): 1335, https://doi.org/10.48550/arXiv.astro-ph/9607174.
9. York, D.~G., Adelman, J., Anderson, J.~E., et al.\ 2000, *The Astronomical Journal*, 120, 1579. doi:10.1086/301513.