LSST AGN Data Challenge database: clustering and variability analysis of quasar light curves

M.S. Pavlović (6), A.B. Kovačević (1,2), D. Ilić (1,3), L.Č. Popović (1,2,4), N. Andrić Mitrović (5), M. Nikolić (1), I. Čvorović-Hajdinjak (1), M. Knežević (1), Đ.V. Savić (4,7)

1 Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia; dragana.ilic@matf.bg.ac.rs (D.I.); lpopovic@aob.rs (L.Cˇ .P.); mladen.nikolic@matf.bg.ac.rs (M.N.)
2 PIFI Research Fellow, Key Laboratory for Particle Astrophysics, Institute of High Energy Physics, Chinese Academy of Sciences, 19B Yuquan Road, Beijing 100049, China
3 Humboldt Research Fellow, Hamburger Sternwarte, Universitat Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany
4 Astronomical Observatory, Volgina 7, 11000 Belgrade, Serbia
5 Department of Mathematics "Tullio Levi Civita", University of Padova, Via Trieste, 35121 Padova, Italy; andricmitrovicnikola@yahoo.com
6 Mathematical Institute of the Serbian Academy of Sciences and Arts, Kneza Mihaila 36, 11000 Belgrade, Serbia
7 Institut d'Astrophysique et de Géophysique, Université de Liège, Allée du 6 Août 19c, 4000 Liège, Belgium

Introduction

• LSST will capture diverse time domain events with frequent observations and broad sky coverage [1] (Figure 1).

• Gaps in LSST quasar light curves pose a challenge for evaluating variable features.

• Data-driven machine learning algorithms, such as Neural Process-NP, are desirable for analyzing LSST AGN multiband light curves [2].

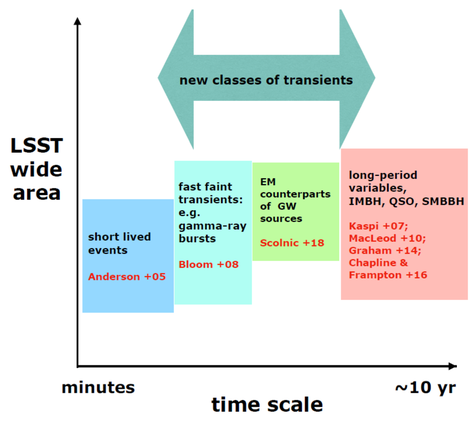


Figure 1. Schematic view of the LSST third scientific pillar.

Logical flow of analysis:

LSST AGN Data Challenge → Clustering according to local topology → 36 clusters of quasar light curves → AGN light curve modeling utilizing Neural Processes

Data

• LSST AGN data challenge 2021*:
  ○ The LSST_AGN_DC mimics the future LSST data release catalogs [3]
  ○ The total number of objects is 440,000
  ○ The total number of epochs for all objects is 5x10^6

• We used a dataset of 1006 quasars with >100 epochs in u-band light curves
• The subsample exhibits a stratification into three non-intersecting branches (Figure 2).
• The Self-Organizing Maps (SOM) algorithm [4] was employed to cluster light curves with similar patterns, resulting in 36 clusters.

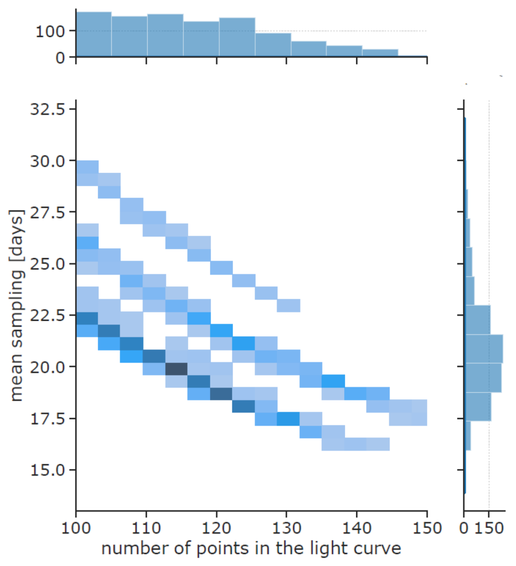


Figure 2. Statistical description of 1006 quasar light curves in the u band having >100 points. The non-Gaussianity of the mean sample and the number of points in the light curves are brought to light by the marginal distributions that are presented on the side plots. Upper marginal: y-axis counts a number of light curves having a number of points given in x-axis in the main plot. Right marginal: y-axis counts a number of light curves having mean sampling in y-axis in the main plot.

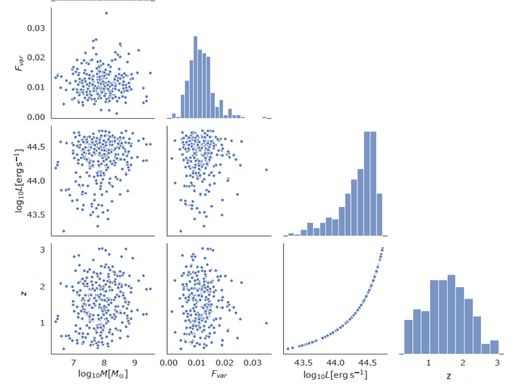


Figure 3: Corner plot of the probability distribution of quasars parameters (redshift z, optical Luminosity Lop, black hole masses M and magnitude with variance Fvar) in selected Cluster 36

Methodology

CONDITIONAL NEURAL PROCESS:
The scheme of a member of NP called the conditional neural process (CNP) [5] is shown in Figure 4. Each pair in the context set is encoded using a multilayer perceptron (eq 1).

$$R^c = \mathrm{Enc}_\theta(\mathcal{C}^{\sim}) = \mathrm{MLP}\left([x^{(c)}; y^{(c)}]\right) \quad \textit{(eq .1)}$$

the local encodings R^(c) are then aggregated by a mean pooling to a global representation R (eq. 2):

$$R = \mathrm{Enc}_\theta(\mathcal{C}^{\sim}) = \frac{1}{C}\sum_{c=1}^{C} \mathrm{MLP}\left([x^{(c)}; y^{(c)}]\right) \quad \textit{(eq .2)}$$

The global representation R is fed along with the target input x^(t) into a decoder MLP to yield the mean and variance of the predictive distribution of the target output (eq. 3):

$$(\mu^{(t)}, \sigma^{2(t)}) = \mathrm{Dec}_\theta(R, x^{(t)}) = \mathrm{MLP}\left([R, x^{(t)}]\right) \quad \textit{(eq .3)}$$

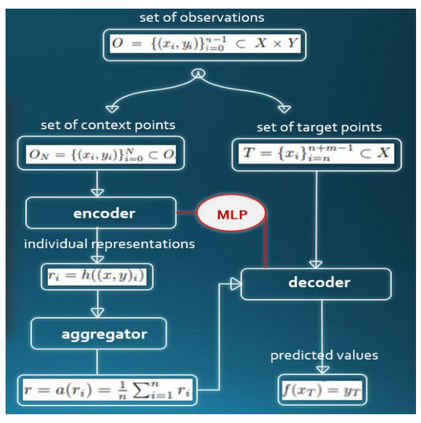


Figure 4. Schematic representation of the Conditional Neural Process, representing data flow. Context points are processed by encoder and aggregator. The output of this process is sent to the decoder along with target points to calculate Predictions [6].

CNP_AGN_LSST python package:

Requirements:

- Python >= 3.7
- pytorch==1.13.0
- cycler==0.11.0
- dill==0.3.4
- matplotlib==3.5.2
- numpy==1.21.5
- pandas==1.4.4
- scikit_learn==1.0.2
- scipy==1.9.1
- tqdm==4.64.1

Base:

- CNP_DATASETCLASS
- CNP_METRICS
- CNP_ARCHITECTURE

Modules:

- Preprocess – for data transformation
- Splitting_and_training
- Prediction
- 4 plotting modules

Results

• The dataset was split into training, testing, and validation subsets with u-band light curves transformed to [-2,2].
• Results from 100 runs from Cluster36 datasets show lower loss but higher mean squared error (MSE) for training due to sharp peaks in the light curves.
• Mean absolute error (MAE) provided more typical behavior, with lower MAE for training than validation (Figure 5).

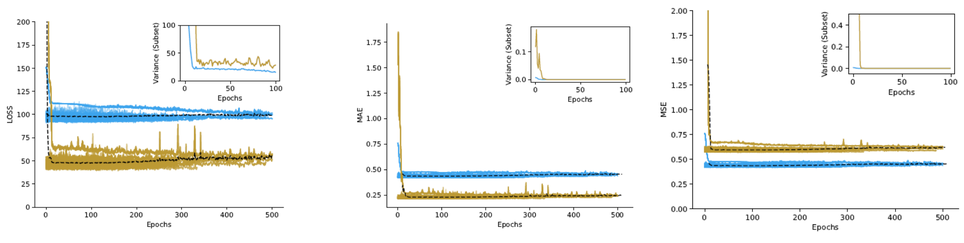


Figure 5: This graphs show the loss, mean absolute error, and mean squared error for the training and validation data sets during 100 rounds of training on Cluster36. The black dashed lines represent the average values of these metrics at each epoch over the 100 rounds.The inset plots show the variances for the training and validation data sets.

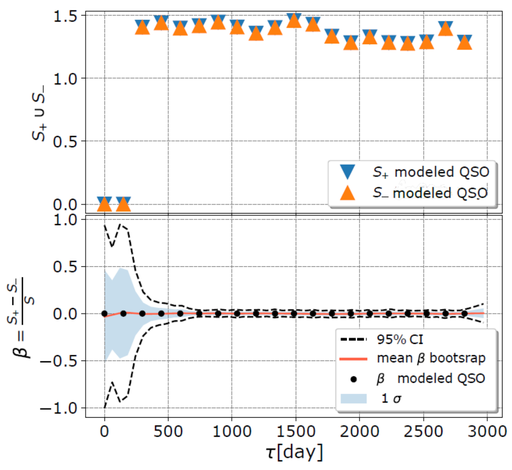


Figure 6: Time symmetry in the light curves of cluster36. Top: modified SFs of quasar sample are indicated by upward and downward pointing triangles, respectively. Bottom: The normalized relative difference β for observed quasar sample (black dots), mean β (red line), 1σ(blue band), and 95% (dashed lines) confidence intervals as inferred from the bootstrapping quasar sample.

• The results from the neural network clustering suggest that there is no significant variability in the quasar light curves from cluster 36.

• In order to test this finding, we used the structure functions (Figure 6).

•The modeling accuracy of the CNP is assessed through MSE and loss values along with 95% confidence intervals, which are wider in regions of light curves with larger errors and flare-like occurrences (Figure 8).

•The flare-like patterns remained present in g and r curves (Figure 7). We estimated that the number of microlensing quasars in the LSST data releases may be N ~ (10^4 -10^5)

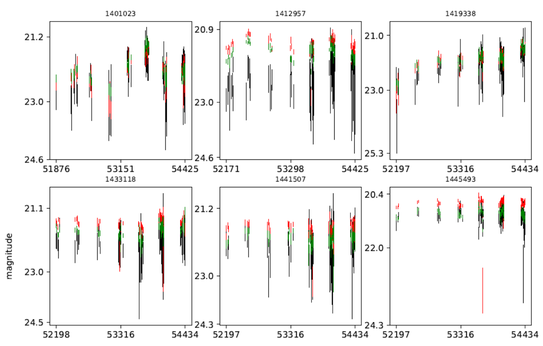
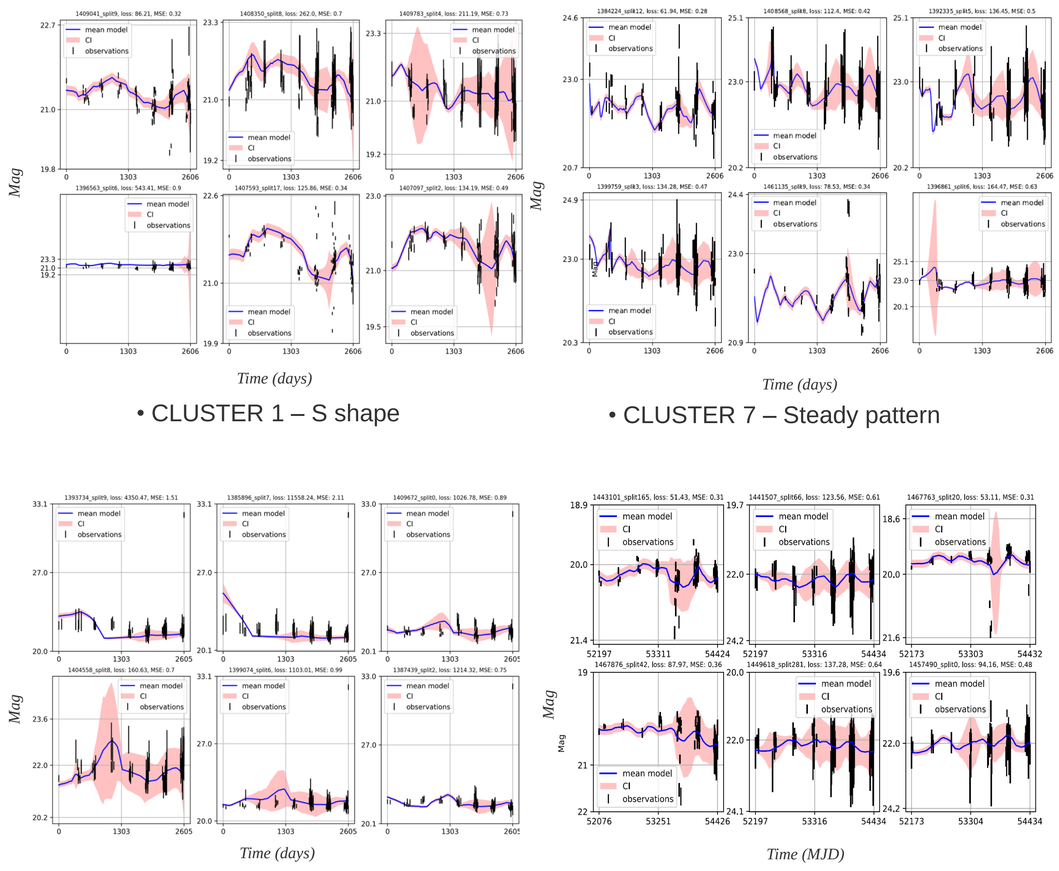


Figure 7: Comparison of presence of flare-like patterns in u (black), g (green) and r (red) bands for Cluster36.



• CLUSTER 1 – S shape
• CLUSTER 7 – Steady pattern
• CLUSTER 8 – Steady pattern
• CLUSTER 9 – Increased brightness
• CLUSTER 36 – Flare like

Figure 8: CNP modeling of light curves for 5 different clusters. The subtitle of each plot shows the object ID from the LSST AGN DC database with a tag indicating the iteration when it is selected for training, testing, or validation during the training process, the logarithmic value of probability loss, and the MSE value. Black error bars indicate observations with measurement uncertainty. Solid blue lines are model fits to the data. The pink band represents the 1σ confidence interval.

Conclusion

•The light curves of over 1000 quasars in the LSST database show different patterns of behavior, which can be grouped into 36 clusters using a neural network.

•This study examines a group of u-band light curves for 283 quasars with very low variability (Cluster 36), as well as 4 smaller clusters.
•The CNP model used in the analysis has a mean square error of 5% (0.5 mag).
• All of the light curves in cluster36 exhibit flare-like features. An initial analysis of the structure function suggests that these features may be associated with microlensing events that happen over a span of five to ten years.

•This scientific case highlights the significance of using CNP and other deep learning methods for data-driven modeling, especially when dealing with objects that have different variability patterns than the standard.

References:

*LSST AGN Data Challenge 2021 can be found on: https://zenodo.org/record/6878414
Our findings on cluster 36 were published and can be found on: https://doi.org/10.3390/universe9060287

[1]. Z.Ivezic et al., LSST: From Science Drivers to Reference Design and Anticipated Data Products. Astrophys. J. 2019, 873, 111.
[2] Zhang, S.Q.;Wang, F.; Fan, F.L. Neural Network Gaussian Processes by Increasing Depth. IEEE Trans. Neural Netw. Learn. Syst. 2022, 1–6.
[3] Yu, W.; Richards, G.; Buat, V.; Brandt, W.N.; Banerji, M.; Ni, Q.; Shirley, R.; Temple, M.; Wang, F.; Yang, J. LSSTC AGN Data Challenge 2021; Zenodo: Geneve, Switzerland, 2022
[4] Vettigli, G. MiniSom: Minimalistic and NumPy-Based Implementation of the Self Organizing Map. 2018. Available online: https://github.com/JustGlowing/minisom (accessed on 8 June 2023).
[5] Garnelo, M.; Schwarz, J.; Rosenbaum, D.; Viola, F.; Rezende, D.; Eslami, S.; Teh, Y. Neural Processes. In Proceedings of the Theoretical Foundations and Applications of Deep Generative Models Workshop, International Conference on Machine Learning (ICML), Stockholm, Sweden, 14–15 July 2018; pp. 1704, 1713
[6] Cvorovic-Hajdinjak, I.; Kovacevic, A.B.; Ilic, D.; Popovic, L.C.; Dai, X.; Jankov, I.; Radovic, V.; Sánchez-Sáez, P.; Nikutta, R. Conditional Neural Process for nonparametric modeling of active galactic nuclei light curves. Astron. Nachr. **2022, 343, e210103.**