Astronomical Data Analysis Software and Systems XXXI
ASP Conference Series, Vol. 535
B. V. Hugo, R. Van Rooyen, and O. M. Smirnov, eds.

Advanced Data Analysis for Observational Cosmology: Applications to the Study of the Intergalactic Medium

Guido Cupani,[1,2] Giorgio Calderone,[1] Stefano Cristiani,[1,2] and Francesco Guarneri[1]

[1] *INAF–Astronomical Observatory of Trieste, I-34151 Trieste, Italy; guido.cupani@inaf.it*

[2] *IFPU–Institute for Fundamental Physics of the Universe, via Beirut 2, I-34151 Trieste, Italy*

Abstract. The analysis of absorption features along the line of sight to distant sources is an invaluable tool for observational cosmology, giving a direct insight into the physical and chemical state of the inter/circumgalactic medium. Such endeavour entails the accessibility of bright QSOs as background beacons, and the availability of software tools to extract the information in a reproducible way. In this article, we will present the latest results we obtained in both directions within the QUBRICS project: we will describe how machine learning techniques were applied to detect hundreds of previously unknown QSOs in the southern hemisphere, and how state-of-the art software like QSFit and Astrocook was integrated in the analysis of the targets, opening up new possibilities for the next era of observations.

1. Introduction: the quest for bright beacons

Several science cases in cosmology and fundamental physics rely on the availability of luminous background beacons (typically quasars, or QSOs) at high redshift, to shed light on the intervening matter. Here is a partial list (see Calderone et al. 2019 for details and references): (i) the determination the matter power spectrum at small scales; (ii) the measure of the abundances of primordial elements; (iii) the measure of a possible variation of fundamental constants; (iv) the Sandage Test, i.e. the direct measurement of the cosmic expansion rate from the redshift drifts of distant objects.

Historically, there has been a dearth of detected luminous QSOs in the Southern hemisphere with respect to the North, due to the lack of appropriate surveys like the SDSS (Blanton et al. 2017). The scenario is now changing, thanks to several recent photometric databases (see Section 2). It is therefore crucial to (i) mine the databases to identify quasar candidates; (ii) acquire the spectra of selected candidates to refine the mining techniques; (iii) consolidate the confirmed QSOs into a database; (iv) develop and test the analysis procedures required to pursue the science cases described above. The QUBRICS project (QUasars as BRIght beacons for Cosmology in the Southern hemisphere, Calderone et al. 2019; Boutsia et al. 2020, 2021) is addressing all these tasks: it has identified so far several hundreds of new QSOs at $\delta < 0$, which are now suitable to be observed with ground-based facilities like VLT UVES and VLT ESPRESSO, and in the future with the ESO ELT.

2. The QUBRICS database

The problem of detecting high-z QSOs in the sky entails two different tasks: (i) distinguishing QSOs from stars, galaxies, and other sources; (ii) estimating the redshift of the source to discard low-z candidates. To tackle the problem, we collected photometric data from various sources into a database, and used it to train both classification and regression models, in order to identify QSO candidates.

Photometric data from the Skymapper (Wolf et al. 2018) and the PanSTARRS (Chambers & Pan-STARRS Team 2018) surveys were cross-matched against the WISE (Wright et al. 2010), GAIA (Gaia Collaboration 2016), and 2MASS (Skrutskie et al. 2006) catalogs. Matching was restricted to targets with $\delta < 15°$, galactic latitude $|b| > 25°$, and a limiting magnitude $i < 18$ ($Y < 18.5$ since 2021). Ambiguous matches were flagged and rejected. The resulting entries are "standardized" (e.g. translating their magnitudes to the same photometric system) and fed into a single database, where they are assigned a unique "QUBRICS ID" (*qid* for short). The database is hosted on a 4TB (RAID01) machine and managed with MariaDB[1] using Julia (Bezanson et al. 2017) scripts for maintenance. The overall architecture of the database is summarized in Figure 1. In the database, the pieces of information information (including the correspondence between the *qid* and the source catalogs with their original identifiers) are joined together produce a "live view" with all columns from the different catalogs, and a row for each match (*AllMatch*). The history of the modifications is also preserved.

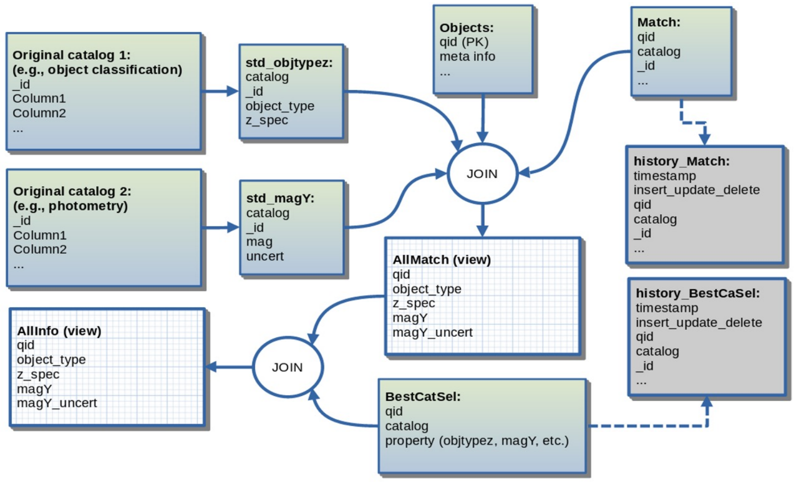

Figure 1. Simplified diagram of the database architecture.

The selection of QSO candidates is performed on a filtered version of the "live view". For each *qid* and property, the *BestCatSel* table reports the catalog that provides the most reliable value, resulting in a reduced view with a single single row for each *qid* (*AllInfo*). All entries in this view can always be traced back to the original catalog using the identifiers. Note that if a more reliable catalog is added, the new values can be uploaded by simply updating the filtering table.

[1] https://mariadb.org

3. Candidate selection and confirmation

The database described in the previous section was used for training different selection techniques to detect QSO candidates. The first QUBRICS selection was based on the Canonical Correlation Analysis (CCA), a higher dimensional selection process based on optimized linear combinations of photometric colors (Calderone et al. 2019). Since then, more advanced techniques have been applied, such as the Probabilistic Random Forest (PRF) and, recently, the XGBoost algorithm (Chen & Guestrin 2016).

The PRF (Guarneri et al. 2021) is a modification of the original Random Forest algorithm, designed to properly handle uncertainties by representing the features of the input data as probability distribution functions. XGB, on the other hand, uses gradient boosting to accommodate model misclassification while keeping overfitting under control. Both algorithms are trained to perform the selection in two stages: (i) discriminating QSOs from stars and galaxies; (ii) rejecting sources at with $z < 2.5$. This second task is not trivial, as the training dataset is heavily skewed toward low-z objects (a feature which was accounted for through oversampling techniques). Despite the difficulties, both algorithms proved effective in terms of precision and recall. XGB was also used to estimate the redshift of the candidates throgut regression.

A spectroscopic follow-up campaign was carried out to confirm the detected candidates. Around 350 new high-z QSOs have been confirmed so far (see Figure 2). Synthetic data are currently being tested to improve the performance of the algorithms.

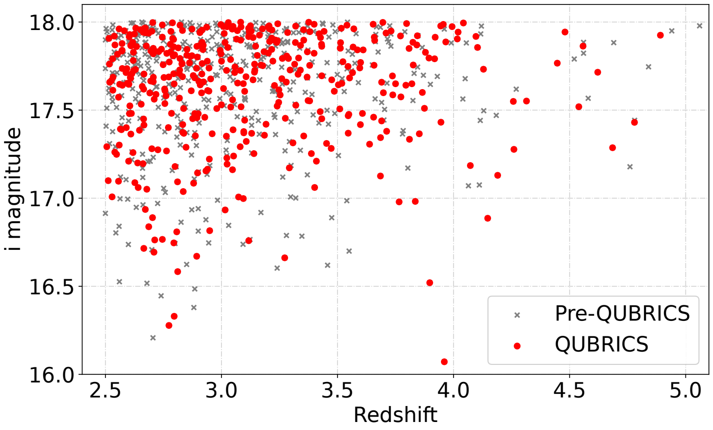

Figure 2. New high-z quasars confirmed by QUBRICS (red dots), compared with those already known from literature (black dots).

4. Data analysis: a test case for Astrocook

A natural complement to the results obtained by QUBRICS is the development of workflows to analyze the collected data. These workflows must not only be scientifically reliable, but also durable (i.e. reproducible and easily maintained). Stability of the data analysis through some decades is essential, for example, to perform the Sandage test (Liske et al. 2008); more generally, the reproducibility paradigm is getting traction in several fields, as a way to cope with increasingly complex science cases.

These requirements laid the foundation for Astrocook (Cupani et al. 2020), a Python package to design, run, and share analysis workflows for QSO spectra. Within the "cooking" metaphor, Astrocook provides both a set of "recipes" (procedures to manipulate spectral data and model emission and absorption features) and a "kitchen" (including a graphical user interface and a scripting-logging tool) to let the user prepare their own "dishes" in a controlled, repeatable way. The scripting/logging tool, in particular, can be used at run-time to transform workflows into readable JSON files, serving both as a documentation and as a way to reproduce the analysis at a later time.

Within the QUBRICS project, Astrocook was used in combination with another tool, QSFit (Calderone & Nicastro 2019), to analyze a peculiar class of objects that emerged from the survey. These objects were initially selected as high-z candidates, but could not be confirmed as bona-fide QSOs from optical spectroscopy alone. Most of them were identified through near-infrared spectroscopy as broad absorption line (BAL) QSOs, with a notable fraction showing strong Feɪɪ absorption bluewards from the emission redshift (FeLoBAL QSOs). The analysis of these objects is currently in press; notably, we are going to publish it together with the complete set of procedures to re-run it, to enforce the reproducibility paradigm and to foster further investigation.

5. Conclusions

We have described the QUBRICS project in the context of the rapid evolution of cosmology into a precision science. We have shown how the most fundamental scientific questions in this fields can be addressed only with a synergistic approach, combining several state-of-the-art software solutions to cover all the steps from data mining to data interpretation. This is our effort and our challenge as we look forward to the next generation of extremely large telescopes.

References

Bezanson, J., et al. 2017, SIAM Review, 59, 65. https://doi.org/10.1137/141000671, URL https://doi.org/10.1137/141000671
Blanton, M. R., et al. 2017, AJ, 154, 28. 1703.00052
Boutsia, K., et al. 2020, ApJS, 250, 26. 2008.03865
— 2021, ApJ, 912, 111. 2103.10446
Calderone, G., & Nicastro, L. 2019, in ADASS XXVI, edited by M. Molinaro, K. Shortridge, & F. Pasian, vol. 521 of ASPCS, 339
Calderone, G., et al. 2019, ApJ, 887, 268. 1909.06391
Chambers, K., & Pan-STARRS Team 2018, in American Astron. Soc. Meeting Abs. #231, vol. 231, 102.01
Chen, T., & Guestrin, C. 2016, arXiv e-prints, arXiv:1603.02754. 1603.02754
Cupani, G., et al. 2020, in SPIE Conference Series, vol. 11452, 114521U
Gaia Collaboration 2016, A&A, 595, A2. 1609.04172
Guarneri, F., et al. 2021, MNRAS, 506, 2471. 2106.12990
Liske, J., et al. 2008, MNRAS, 386, 1192. 0802.1532
Skrutskie, M. F., et al. 2006, AJ, 131, 1163
Wolf, C., et al. 2018, Publ. of the Astron. Soc. of Australia, 35, e010. 1801.07834
Wright, E. L., et al. 2010, AJ, 140, 1868. 1008.0031