

A Refined QSO Selection Method Using Diagnostics

Dae-Won Kim^{1,2,3}, Pavlos Protopapas^{1,3}, Markos Trichas¹,
Michael Rowan-Robinson⁴, Roni Khardon⁵, Charles Alcock¹
and Yong-Ik Byun^{2,6}

¹Harvard-Smithsonian Center for Astrophysics, Cambridge, MA, USA

²Department of Astronomy, Yonsei University, Seoul, South Korea

³Institute for Applied Computational Science, Harvard University, Cambridge, MA, USA

⁴Astrophysics Group, Imperial College, London, UK

⁵Department of Computer Science, Tufts University, Medford, MA 02155, USA

⁶Yonsei University Observatory, Yonsei University, Seoul, South Korea

Abstract. We present 663 QSO candidates in the Large Magellanic Cloud (LMC) that were selected using multiple diagnostics. We started with a set of 2,566 QSO candidates selected using the methodology presented in our previous work based on time variability of the MACHO LMC light curves. We then obtained additional information for the candidates by cross-matching them with the Spitzer SAGE, the 2MASS, the Chandra, the XMM, and an LMC *UBVI* catalogues. Using that information, we specified diagnostic features based on mid-IR colours, photometric redshifts using SED template fitting, and X-ray luminosities, in order to discriminate more high-confidence QSO candidates in the absence of spectral information. We then trained a one-class Support Vector Machine model using those diagnostics features. We applied the trained model to the original candidates, and finally selected 663 high-confidence QSO candidates. We cross-matched those 663 QSO candidates with 152 newly-confirmed QSOs and 275 non-QSOs in the LMC fields, and found that the false positive rate was less than 1%.

Keywords. Magellanic Clouds, methods: data analysis, quasars: general

1. Introduction

In previous work (Kim *et al.* 2011a) we developed a QSO selection method using a supervised classification model trained on a set of variability features extracted from the MACHO light curves, and including a variety of variable stars, non-variable stars and QSOs. The trained model showed a high efficiency of 80% and a low false positive rate of 25%. Using that method, we selected 2,566 QSO candidates from the light-curve database. We then developed and employed a decision procedure on the basis of diagnostics using (1) mid-IR colours, (2) photometric redshifts and (3) X-ray luminosities for those candidates in order to separate *high confidence* QSO candidates (hereinafter hc-QSOs). We thus chose in total 663 hc-QSOs out of 2,566. Those 663 candidates are very probably QSOs; if confirmed, they will increase the number of known QSOs in the MACHO LMC database by a factor of ~ 12 .

2. Selection Methods

We selected 2,566 QSO candidates from the MACHO light-curve database using the QSO selection method developed by Kim *et al.* (2011a), and applied multiple diagnostics tests on them.

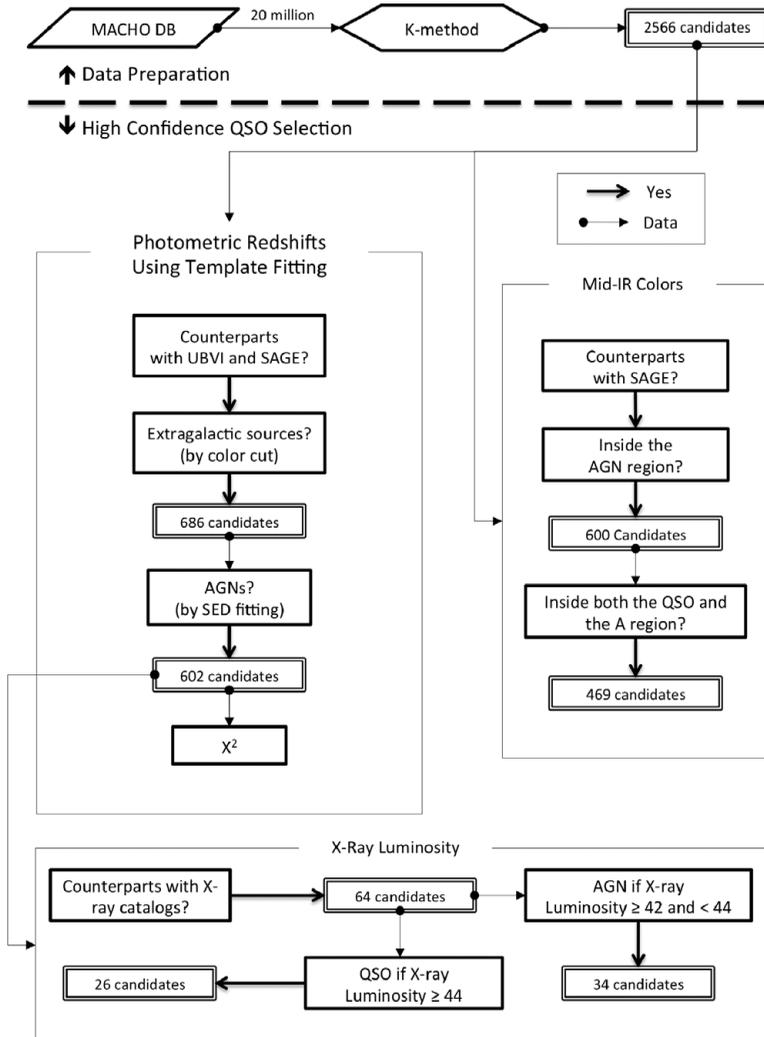


Figure 1. The rectangles with bold borderlines are the diagnostics. For most of the diagnostics, we determined whether the candidates are likely to be QSOs (solid line arrows). The thin arrows show the data flow. The double-lined rectangles show the number of candidates.

• **Spitzer mid-IR properties.** It is known that the mid-IR colour is an efficient discriminator between AGNs and stars/galaxies because their spectral energy distributions are substantially different from one another (Laurent *et al.* 2000; Lacy *et al.* 2004). Lacy *et al.* (2004) introduced a mid-IR colour cut to separate AGNs in the Spitzer SAGE catalogue (Surveying the Agents of a Galaxy’s Evolution; Meixner *et al.* 2006). We used those mid-IR colour selections as the first diagnostic.

• **Photometric redshifts using SED fitting.** We cross-matched the 2,566 QSO candidates with the *UBVI* catalogue for the LMC (Zaritsky *et al.* 2004) and the 2MASS catalogue (Skrutskie *et al.* 2006), to extract *UBVI* and *JHK* magnitudes. We next separated stars from Galaxies plus AGNs (i.e., extragalactic sources) using a criterion proposed by Eisenhardt *et al.* (2004) and Rowan-Robinson *et al.* (2005). The 686 extragalactic sources thus identified were then fitted with galaxy templates in order to derive photometric redshifts (Rowan-Robinson *et al.* 2008). The template objects included three

QSO, one starburst and 10 galaxies. The candidates which matched the QSO templates were considered to be QSOs.

- **X-ray luminosities.** We cross-matched the 2,566 QSO candidates with two X-ray point source catalogues, the Chandra X-ray source catalogue (Evans *et al.* 2010) and the XMM-Newton 2nd Incremental Source catalogue (Watson *et al.* 2009). We found 88 counterparts. 64 of them fitted the SED templates and therefore had estimated photometric redshifts; we used those photometric redshifts and the X-ray fluxes from the catalogues to calculate their X-ray luminosities. The candidates showing high X-ray luminosity were deemed likely to be QSOs (Elvis *et al.* 1994; Persic *et al.* 2004).

We then employed the one-class Support Vector Machine (SVM) classification method to select just the high-confidence QSO candidates because we do not have negative examples (i.e., a non-QSO training set). To train a model, we define the diagnostics results as feature vectors. Fig. 1 outlines the calculation of the diagnostics, and the number of candidates for which the diagnostics are available. Kim *et al.* (2011b) give details about the selection method.

3. Crossmatching with Newly Discovered QSOs by Kozłowski

To estimate the efficiency and the false positive rate of our selection method, we cross-matched the 663 candidates with the 152 newly-discovered QSOs (Kozłowski *et al.* 2011) and 275 confirmed non-QSOs (i.e., false positives). We found that the yield was higher than 43%, and the false positive rate was less than 1%.

4. Summary

From 2,566 QSO candidates that were selected by the time variability of their MACHO light curves in the MACHO light-curve database, we used the methods described above to identify 663 high-confidence QSO candidates in the LMC fields. This set can be used as a target set for spectroscopic surveys as they should maximize the yield; that is important because spectroscopic observations for relatively faint objects such as the QSO candidates in dense- and wide-field areas around the LMC are extremely expensive in telescope time.

References

- Eisenhardt, P. R., *et al.* 2004, *ApJS*, 154, 48
 Elvis, M., *et al.* 1994, *ApJS*, 95, 1
 Evans, I. N., *et al.* 2010, *ApJS*, 189, 37
 Kim, D.-W., *et al.* 2011a, *ApJ*, 735, 68
 Kim, D.-W., *et al.* 2011b, *ArXiv e-prints*
 Kozłowski, S., *et al.* 2011, *ArXiv e-prints*
 Lacy, M., *et al.* 2004, *ApJS*, 154, 166
 Laurent, O., *et al.* 2000, *A&A*, 359, 887
 Meixner, M., *et al.* 2006, *AJ*, 132, 2268
 Persic M., *et al.* 2004, *A&A*, 419, 849
 Rowan-Robinson, M., *et al.* 2008, *MNRAS*, 386, 697
 Rowan-Robinson, M., *et al.* 2005, *AJ*, 129, 1183
 Skrutskie, M. F., *et al.* 2006, *AJ*, 131, 1163
 Watson, M. G., *et al.* 2009, *A&A*, 493, 339
 Zaritsky, D., Harris J., Thompson, I. B., & Grebel, E. K. 2004, *AJ*, 128, 1606