Draft version January 11, 2019
Typeset using LaTeX **twocolumn** style in AASTeX62

The QUEST-La Silla AGN Variability Survey: selection of AGN candidates through optical variability

P. Sánchez-Sáez,[1,2] P. Lira,[1] R. Cartier,[3] N. Miranda,[4] P. Coppi,[5] L. C. Ho,[6,7] P. Arévalo,[8] and C. Yovaniniz[1]

[1]*Departamento de Astronomia, Universidad de Chile, Casilla 36D, Santiago, Chile*
[2]*European Southern Observatory, Casilla 19001, Santiago 19, Chile*
[3]*Cerro Tololo Inter-American Observatory, National Optical Astronomy Observatory, Casilla 603, La Serena, Chile*
[4]*Institut für Informatik, Humboldt-Universität zu Berlin*
[5]*Yale Center for Astronomy and Astrophysics, 260 Whitney Avenue, New Haven, CT 06520, USA*
[6]*Kavli Institute for Astronomy and Astrophysics, PekingUniversity, Beijing 100871, China*
[7]*Department of Astronomy, School of Physics, Peking University, Beijing 100871, China*
[8]*Instituto de Fisica y Astronomia, Facultad de Ciencias, Universidad de Valparaiso, Gran Bretana No. 1111, Playa Ancha, Valparaiso, Chile*

ABSTRACT

Keywords: galaxies: active - methods: statistical - surveys

1. INTRODUCTION

Active galactic nuclei (AGN) are one of the most energetic phenomena in the universe and are characterized by their time-variable emission in every waveband in which they have been studied. Variability studies are fundamental to understanding the extreme physical conditions of accretion disks near super massive black holes (SMBH). AGN variability seems to be well described as a stochastic process (Kelly et al. 2009, 2014), with characteristic time-scales ranging from days to years.

AGN are commonly classified in the optical range by the presence or absence of broad permitted emission lines (FWHM> 2000 km s^{-1}), into broad-line (BL) AGN (or type 1) and narrow-line (NL) AGN (or type 2), respectively. The unified model is one of the most successful approaches to explain this dichotomy. It postulates that a dusty torus around the central engine is responsible for the different classes of AGN, which occur when we observe the source at different angles (Antonucci & Miller 1985). The most promising models include a clumpy torus and disk winds (see Netzer 2015 and references therein), as they would explain the torus spectral energy distribution (SED) observed in the near-infrared (NIR) and mid-infrared (MIR) bands and the existence of at least some "changing look" AGN (e.g. LaMassa et al. 2015).

The Large Synoptic Survey Telescope (LSST; Ivezic et al. 2008) will revolutionize time-domain astronomy, providing for the first time the opportunity to study variable objects for a long period of time ($\sim$ 10 years), at very low magnitudes ($r \sim 24.5$ for the single images), and with a large total covered area (25,000 deg^2). Simulations performed by the LSST AGN Science Collaboration predict a detection of over 10^7 AGN to beyond $m \sim 24$ (LSST Science Collaboration et al. 2009). This is a huge improvement in the number of sources available for variability analysis since current studies typically reach a limiting magnitude of $m \sim 21$ with between 10 and 10^5 sources. Given the importance of variability and the large number of objects expected it is critical to characterize AGN variability and define reliable selection criteria before LSST's first observations.

Traditionally, the AGN selection follows the philosophy of finding regions in an UV/optical color-color space in which AGN can be cleanly separated from stars (e.g. Richards et al. 2002, 2009; Ross et al. 2012). However, some AGN populations are known for having colors that fall outside the region occupied typically by AGN, mimicking those of stars, such as broad absorption line quasars (BAL QSO) and high redshift quasars (high-z QSO) (Butler & Bloom 2011; Palanque-Delabrouille et al. 2011), or low luminosities AGN (LLAGN), whose colors can be highly contaminated by the emission from the host galaxy. Therefore, we need alternative methods to identify AGN candidates with the upcoming time domain surveys. Promising selection methods involve the use of variability techniques.

Butler & Bloom (2011) implemented a variability–based selection algorithm to classify high redshift quasars in the Sloan Digital Sky Survey (SDSS; York et al. 2000) Stripe 82 field. They used damp random walk modelling (Kelly et al. 2009) to separate sources showing stochastic (or quasar–like) variability from those with temporally uncorrelated variability. Particularly, they targeted sources with redshifts in the range

$2.5 \leq z \leq 3$, where color–based selection of AGN is quite difficult. Palanque-Delabrouille et al. (2011) used the variability structure function (Schmidt et al. 2010) to separate quasars, variable stars, and non–variable stars, in the SDSS Stripe 82 data. They implemented a neural network algorithm that separates the classes considering their structure function parameters. A similar technique has been used by the SDSS IV *the extended Baryon Oscillation Spectroscopic Survey* (eBOSS) team to select quasar candidates with $z > 2.1$ by variability (Myers et al. 2015). More recently, Tie et al. (2017) used data from the supernova fields of the Dark Energy Survey (DES; Abbott et al. 2018) to select quasars by combining color and variability selection methods. All these previous works have shown the capability of selecting AGN candidates through variability analyses, demonstrating that variability–based techniques can increase considerably the number of AGN candidates in the redshift range where the colors of stars are similar to those of AGN.

In this paper we present our variability–based technique to select AGN candidates using data from the QUEST-La Silla AGN variability survey (Cartier et al. 2015). Variability features, like the structure function, have been used to characterize the variable sources. We then used a Random Forest algorithm to classify our objects as either AGN or non-AGN. We did not consider colors in our selection, in order to avoid a selection biased by typical type 1 AGN. For some of our candidates we have performed spectroscopic follow ups. Four of the fields observed by the QUEST-La Silla AGN variability survey correspond to the LSST Deep Drilling Fields (DDF), and the expected cadence of the DDF will be similar to the one used by the QUEST-La Silla AGN variability survey (but covering 10 years)[1]. Thus, our work is a perfect pilot study for the selection of AGN with LSST.

The paper is organized as follows. In Section 2 we describe the QUEST-La Silla AGN variability survey, and the light curve construction procedure. In Section 3 we describe the Random Forest algorithm, the variability features, and the labeled set used for the selection. We also discuss the performance of our Random Forest classifier, and we provide the list of selected candidates. In Section 4 we provide the results on confirming the nature of some of our candidates by using public data and spectroscopic follow ups. Finally, in Section 5 we discuss the implications of our findings and summarize the main results.

[1] https://www.lsst.org/scientists/survey-design/ddf

2. DATA

2.1. *The QUEST–La Silla AGN variability survey*

Between 2010 and 2015 we carried out "The QUEST–La Silla AGN variability survey" (hereafter QUEST–La Silla), using the wide-field QUEST camera on the 1m ESO-Schmidt telescope at La Silla Observatory (Cartier et al. 2015). Our survey includes the COSMOS, ECDF-S, ELAIS-S1, XMM-LSS and Stripe-82 fields. These are some the most intensively observed regions in the sky, with a huge amount of ancillary data ranging from X-rays to radio waves. Our QUEST fields are much larger than just COSMOS, ELAIS-S1, etc., even though we use the same names for them, with a surveyed area of ~ 7 deg^2 per field. One of the advantages of our survey over other surveys was the very intense monitoring, observing the fields every possible night. We obtained between 2 to 5 observations per night to remove spurious variability due to artefacts, to potentially study intra-night AGN variability, and to produce stacked images to reach deeper magnitudes. Individual images reached a limiting magnitude between $r \sim 20.5$ and $r \sim 21.5$ mag for a exposure time of 60 seconds or 180 seconds, respectively.

The aims of our survey are: 1) to test and improve variability selection methods of AGN, and find AGN populations missed by other optical selection techniques (Schmidt et al. 2010; Butler & Bloom 2011; Palanque-Delabrouille et al. 2011); 2) to obtain a large number of well–sampled light curves, covering time-scales ranging from days to years; 3) to study the link between the variability properties (e.g., characteristic time-scales and amplitudes of variation) with physical parameters of the system (e.g., black-hole mass, luminosity, and Eddington ratio).

Cartier et al. (2015) presented the technical description of the survey, the full characterisation of the QUEST camera, and a study of the relation of variability with multi-wavelength properties of X-ray selected AGN in the COSMOS field. In Sánchez-Sáez et al. (2018) we performed a statistical analysis of the connection between AGN variability and physical properties of the central SMBH. We constructed optical light curves using data from QUEST–La Silla. To model the variability, we used the structure function (among other features). For the measurement of SMBH physical properties, we used public spectra from the SDSS. We found that the amplitude of the variability (A) depends solely on the rest-frame emission wavelength (λ_{rest}) and the Eddington ratio, where A anticorrelates with both λ_{rest} and $L/L_{\rm Edd}$. This suggests that AGN variability does not evolve over cosmic time, and its amplitude is inversely related to the accretion rate.

Table 1. Number of light curves per field.

Field	total light curves	well sampled light curves
COSMOS	68,514	45,323
XMM–LSS	104,962	82,697
Elais–S1	49,504	38,106
ECDF–S	54,649	42,457
Total	277,626	208,583

2.2. *Light curve construction*

We reduced the data from the QUEST–La Silla using our own customized pipeline, following the same procedure described by Cartier et al. (2015), which includes dark subtraction, flat-fielding, and astrometric and photometric calibration. To calibrate the photometry, we used public photometric SDSS catalogs (Gunn et al. 1998; Doi et al. 2010) for the COSMOS, Stripe 82 and XMM–LSS fields, and public catalogs from the first year of DES (Abbott et al. 2018) for the ELAIS–S1 and ECDF–S fields. We performed aperture photometry using SExtractor (Bertin & Arnouts 1996), with the same optimal aperture found by Cartier et al. (2015) for the QUEST camera ($\sim 6''.18$). We then constructed light curves for all the sources from the SDSS and DES catalogs with detections in the QUEST–La Silla data, using the same methodology as in Cartier et al. (2015). From the SDSS catalog, we could obtain photometry of every source in the COSMOS, XMM–LSS, and Stripe 82 fields in the u, g, r, i, and z bands, and from the DES catalog we obtained photometry in the g, r, i, and z bands for the ELAIS–S1 and ECDF–S fields.

We decided to bin our light curves every three days, in order to reduce the noise in our light curves, produced by changes in atmospheric conditions, the relatively low quality of the QUEST camera, among other factors. In this work, we excluded the Stripe-82 field, since it is a crowded field, and requires point spread function (PSF) photometry. We generated a total of $277,629$ light curves for sources located in the COSMOS, ECDF–S, ELAIS–S1, and XMM–LSS fields. In order to have statistically significant variability features of the sources, we decided to include in our analysis only those light curves with at least 40 epochs and a length greater than or equal to 200 days, after the three days binning was applied to the original light curves (hereafter well sampled light curves). There are $208,583$ well sampled light curves in the four fields. In table 1 we summarize the total number of light curves and the number of well sampled light curves in each field.

3. SELECTION OF AGN CANDIDATES

We developed a variability–based AGN selection technique to find AGN populations missed by other optical selection methods. We implemented a supervised automatic classification using a Random Forest algorithm (RF; Breiman 2001) to classify our 208,583 objects with well sampled light curves as either AGN or non-AGN according to their variability features. We tested two classifiers, one that includes only variability features, and one that includes optical colors. In the following sections we describe the selection methodology, the features used in our analysis, and the results of the classification for sources from the QUEST–La Silla survey.

3.1. *Random Forests*

A type of learning algorithm that has been particularly effective in various applications are the so-called algorithms of ensemble learning. The idea of these algorithms is to add the results in the work of learning (or classification) of a large number of very simple models on subsets of the training data. The RF algorithm performs this by using simple models of decision trees. A decision tree is a hierarchical structure that performs successive partitions on the data, each of them according to a certain criteria, such as a cut-off value in one of the descriptors or features. In this way, the data is divided into smaller and smaller subsets as the tree goes deeper, until it reaches the leaves of the tree. Each of these leaves is associated with a single class, and therefore the elements that falls on the leaves corresponding to a particular class are those that will be classified as belonging to that class.

A RF algorithm constructs a large number of decision trees from random sub-sets of a training set. Each of these classifiers then assigns a class to a certain input element. The final classification function of the algorithm ponders each of these results according to the size of the sub-set used by each tree, and generates an average score; which can then be interpreted as the probability that the input element belongs to a certain class ($P_{\rm RF}$).

For the selection of AGN candidates we used the *scikit-learn*[2] Python package implementation of RF. We performed an hyperparameter selection procedure in order to obtain the optimal values for the RF classifier, by means of a cross-validated randomized search procedure and using the "accuracy" (see its definition in section 3.4) as the target score to optimize. The parameters considered in this randomized search includes the num-

[2] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

ber of trees in the forest, and the number of features to consider when looking for the best split. In order to take into account the class imbalance in the classification process, we initialized the class weight hyperparameter as "*balanced_subsample*"

The variability features used by the RF classifier are described in the following section (3.2), and are listed in Table 2. We trained the RF classifier using a labeled set of AGN and stars with spectroscopic classification from SDSS and with detections in the QUEST–La Silla survey (see section 3.3). During the RF classifier training, we used the 30% of the labeled set as a test set, and a 70% of the labeled set as a training set. We then applied the obtained RF classifier to our unlabeled set, composed by our 208,583 sources with QUEST–La Silla light curves, to classify them as either AGN or non-AGN. As a result, we obtain a predicted class and the predicted class probability ($P_{\rm RF}$) associated to each source of the unlabeled set.

3.2. *Variability features*

In order to have a complete description of the variability of our sources, we used several variability features. Following the same approach of Sánchez et al. (2017) and Sánchez-Sáez et al. (2018), we used two parameters related to the amplitude of the variability, P_{var} and the excess variance ($\sigma_{\rm rms}$), and one method related to the structure of the variability, the structure function (SF).

In particular, P_{var} (see Sánchez et al. 2017 and references therein) corresponds to the probability that the source is intrinsically variable, it considers the χ^2 of the light curve, and calculates the probability $P_{var} = P(\chi^2)$ that a χ^2 lower or equal to the observed value could occur by chance for an intrinsically non-variable source.

$\sigma_{\rm rms}$ is a measure of the intrinsic variability amplitude (see Sánchez et al. 2017 and references therein), and it is calculated as $\sigma^2_{rms} = (\sigma^2_{LC} - \overline{\sigma}^2_m)/\overline{m}^2$, where σ_{LC} is the standard deviation of the light curve, $\overline{\sigma}_m$ is the mean photometric error, and $\overline{m}$ is the mean magnitude.

The SF (e.g. Schmidt et al. 2010) is a measure of the amplitude of the variability as a function of the time lapse between compared observations (τ), and it can be modelled as a power law: $\mathrm{SF}(\tau) = A_{\rm SF}\left(\frac{\tau}{1\mathrm{yr}}\right)^{\gamma_{\rm SF}}$, where $A_{\rm SF}$ corresponds to the amplitude of the variability at 1 year, and $\gamma_{\rm SF}$ is the logarithmic gradient of this change in magnitude.

We also used some variability features from the Feature Analysis for Time Series (FATS; Nun et al. 2015) Python package, related with the amplitude of the variability (e.g. the mean variance and the percent amplitude) and the structure of the light curve (e.g. the linear trend and the auto-correlation function length), as well as the period of the Lomb-Scargle periodogram (VanderPlas 2018), derived by using the AstroML module for Python (VanderPlas et al. 2012). A list of all the variability features used in this work is shown in Table 2, together with a brief description of each feature and its reference.

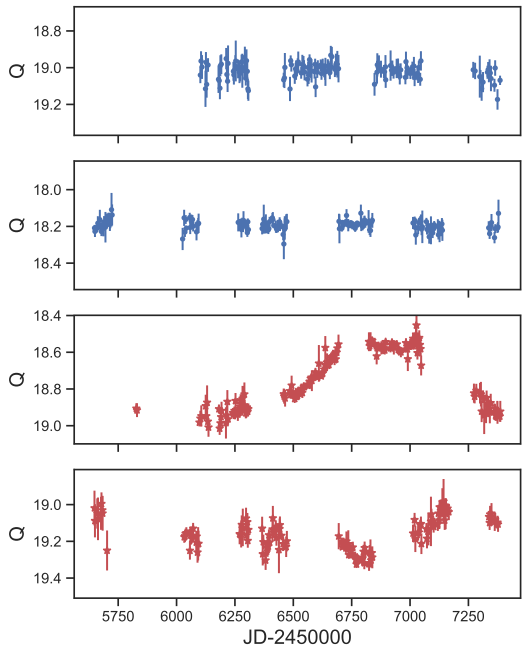

Figure 1. Example of four light curves from the QUEST–La Silla labeled set: two stars (blue dots) and two AGN (red stars).

3.3. *Labeled set*

Three of our fields (COSMOS, Stripe-82, and XMM-LSS) have spectroscopic information from SDSS. We selected a sample of 2405 AGN and 2608 stars with both SDSS spectral classification from the the SDSS-DR14 database (Abolfathi et al. 2018), and well sampled light curves from QUEST–La Silla, as our labeled set for the RF classifier training. As mentioned in section 3.1, 30% of the labeled set was used as a test set and 70% as training set for the RF modelling. Figure 1 provides examples of QUEST–La Silla light curves for 4 sources of the labeled set.

In Figure 2 we show three color–color diagrams of the labeled set: $u-g$ versus $g-r$, $g-r$ vs $r-i$, and $r-i$ vs $i-z$. It can be seen that stars and AGN in general occupy well defined regions in the color–color spaces, this is expected, because the techniques used to select

Table 2. List of features.

Feature	Description	Reference
P_{var}	Probability that the source is intrinsically variable	McLaughlin et al. (1996)
$\sigma_{\rm rms}$	Measure of the intrinsic variability amplitude.	Allevato et al. (2013)
$A_{\rm SF}$	Amplitude of the variability at 1 year, derived from the SF	Schmidt et al. (2010)
$\gamma_{\rm SF}$	Logarithmic gradient of the change in magnitude, derived from the SF	Schmidt et al. (2010)
Std*	Standard deviation of the light curve (σ_{LC})	Nun et al. (2015)
Meanvariance*	Ratio of the standard deviation to the mean magnitude ($\sigma_{LC}\overline{m}$)	Nun et al. (2015)
MedianBRP*	Fraction of photometric points within amplitude/10 of the median magnitude	Richards et al. (2011)
Autocor-length*	Lag value where the autocorrelation becomes smaller than e^{-1}	Kim et al. (2011)
StetsonK*	A robust kurtosis measure	Kim et al. (2011)
η^e*	Ratio of the mean of the square of successive differences to the variance of data points	Kim et al. (2014)
PercentAmp*	Largest percentage difference between either the max or min magnitude and the median	Richards et al. (2011)
Con*	number of three consecutive data points that are brighter or fainter than $2\sigma_{LC}$	Kim et al. (2011)
LinearTrend*	Slope of a linear fit to the light curve	Richards et al. (2011)
Beyond1Std*	Percentage of points beyond one σ_{LC} from the mean	Richards et al. (2011)
Q31*	Difference between the third quartile and the first quartile of a light curve	Kim et al. (2014)
PeriodLS	Period from the Lomb-Scargle periodogram	VanderPlas (2018)

Note. (*) Features from FATS

stars and AGN (e.g. Richards et al. 2002). In general, most of the stars are located in a region of the color–color space called stellar locus (e.g. Covey et al. 2007; Sesar et al. 2007). Since stellar colors become monotonically redder as the effective temperature decreases (Covey et al. 2007), we normally observe a high concentration of cold stars in a region around $g-r \sim 1.5$, with $r-i \gtrsim 0.8$. Moreover, extragalactic sources are normally located in regions of the color–colors space with $r-i \lesssim 1.5$ and $i-z \lesssim 1.5$ (e.g Rahman et al. 2016), since their integrated emission have a very low contribution from cold stars. In Figure 2 we can see that AGN in the labeled set occupy different positions of the $u-g$ vs $g-r$ spece, however, in the $r-i$ vs $i-z$ space, they are concentrated in a particular area around $r-i \sim 0$ and $i-z \sim 0$. Therefore we can use the $r-i$ and $i-z$ colors to separate cold stars and extragalactic sources.

3.4. *Performance of the Random Forest classifier*

We tested two different RF classifiers, the first one includes only variability features (hereafter RF1), and the second one includes variability features and the $r-i$ and $i-z$ colors (hereafter RF2). We only include the $r-i$ and $i-z$ colors, because they can easily separate cold stars and extragalactic sources. Besides, AGN in the labeled set occupy different positions in the $u-g$ vs $g-r$ diagram, thus, in order to avoid a selection biased by typical type 1 AGN, we exclude these colors.

3.4.1. *Classification considering variability features*

Our first RF classifier (RF1) includes only variability features. After the training of RF1 we tested its performance by using the confusion matrix, which is shown in Figure 3. It can be seen that AGN (true positives) are in general well classified, and also that the fraction of stars classified as AGN (false positives) is very low.

We also compute the accuracy (A), precision (P), recall (R), and F1 scores, which are defined by means of the True Positives (TP: known AGN classified as AGN by the RF classifie)r, the False Positives (FP: known stars classified as AGN), the True Negatives (TN: known stars classified as stars), and the False Negatives (FN: known AGN classified as stars):

$$\begin{aligned} A &= \frac{TP+TN}{Total\ Sample} \\ P &= \frac{TP}{TP+FP} \\ R &= \frac{TP}{TP+FN} \\ \mathrm{F1} &= 2 \times \frac{P \times R}{P+R} \end{aligned} \qquad (1)$$

Table 3 shows the calculated scores. for the RF1 classifier. From these scores, and from the confusion matrix, we can say that RF1 presents a low fraction of False Positives, thus, the sample of predicted AGN has a low contamination of stars. However, we tend to miss

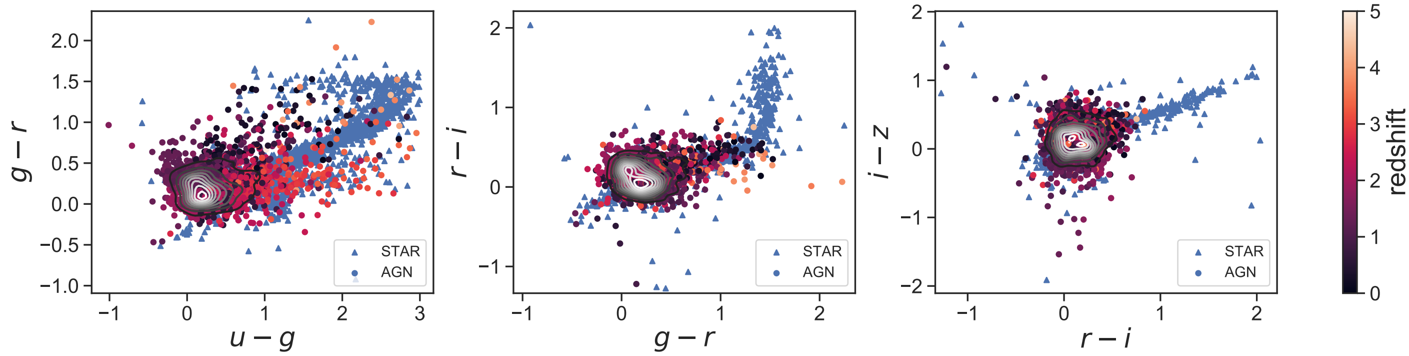

Figure 2. Color–color diagrams of the labeled set. In the left panel we show $u-g$ versus $g-r$, in the middle panel $g-r$ vs $r-i$, and in the right panel $r-i$ vs $i-z$. The stars are represented by blue triangles, and the AGN are represented by circles whose colors depend on the redshift of every source. The contour plots show the distribution of AGN.

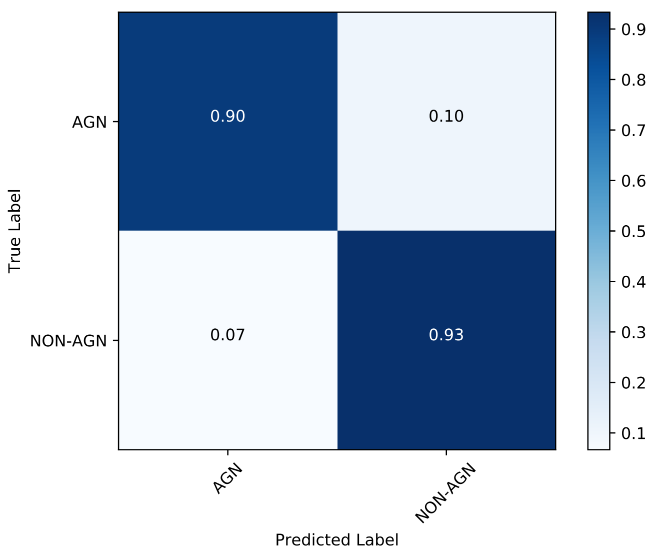

Figure 3. Confusion matrix from testing the RF1 classifier in the test set. True Label represent the classification done from SDSS spectra, and Predicted Label is the outcome of the RF1 classifier.

a fraction of real AGN ($\sim 10\%$). This is produced by the difficulty of detect a variable signal from AGN with low variability, and since we are only considering variability properties for the classification, they can be confused by stars.

It is important to consider that we are testing the RF1 classifier in a sample of AGN selected mostly by means of their optical colors, and since we are only considering variability features in our selection, the confusion matrix and the different scores, obtained from our labeled sample, will not be a good prediction of the performance of our method in the unlabeled sample.

One of the advantages of the RF classification is that we can easily know the feature importance, since it provides a ranking score for each feature, which tell us how well every feature separate the two classes. In the firsts columns of Table 4 we provide the list of features, ordered by importance (according to the rank value), for the RF1 classifier. It can be seen that the four most important features are the amplitude of the structure function, the excess variance, and the Meanvariance and Q31 parameter from FATS. As an example, we show in Figure 4 the distribution of the $A_{\rm SF}$ and Q31 features for the labeled set. We show with black dots those AGN classified as variable, according to the definiton proposed by Sánchez et al. (2017), where a source is classified as variable when its light curve satisfies $P_{var} \geq 0.95$ and $(\sigma^2_{rms} - err(\sigma^2_{rms})) > 0$. From the figure, it can be seen that AGN and stars are clearly separated by these two features, and also that the majority of the AGN with low amplitude of the variability are classified as non-variable.

Table 3. Scores measured in the test set for each classifier

Score	RF1	RF2
Accuracy	0.916	0.923
Precision	0.909	0.909
Recall	0.933	0.950
F1	0.921	0.930

3.4.2. *Classification considering variability features and optical colors*

Our second RF classifier (RF2) includes variability features and the $r-1$ and $i-z$ colors. Figure 5 shows the confusion matrix for RF2. In this case, the confusion matrix is similar than the confusion matrix for RF1, however, in the case of RF2 we have a slightly purer population of AGN candidates. The accuracy, precision, re-

Table 4. Feature importance for each classifier.

RF1		RF2	
Feature	Rank	Feature	Rank
$A_{\rm SF}$	0.197	$A_{\rm SF}$	0.209
$\sigma_{\rm rms}$	0.139	$\sigma_{\rm rms}$	0.149
Meanvariance*	0.127	Q31*	0.102
Q31*	0.111	P_{var}	0.093
P_{var}	0.095	Std*	0.088
Std*	0.090	Meanvariance*	0.086
PercentAmp*	0.040	PercentAmp*	0.045
$\gamma_{\rm SF}$	0.036	Autocor-length*	0.035
Autocor-length*	0.033	$\gamma_{\rm SF}$	0.031
MedianBRP*	0.025	$r-i$	0.028
LinearTrend*	0.023	MedianBRP*	0.021
PeriodLS	0.023	PeriodLS	0.020
η^e*	0.023	η^e*	0.019
Beyond1Std*	0.019	Beyond1Std*	0.019
StetsonK*	0.018	LinearTrend*	0.019
Con*	0.002	$i-z$	0.019
		StetsonK*	0.014
		Con*	0.002

Note. (*) Features from FATS

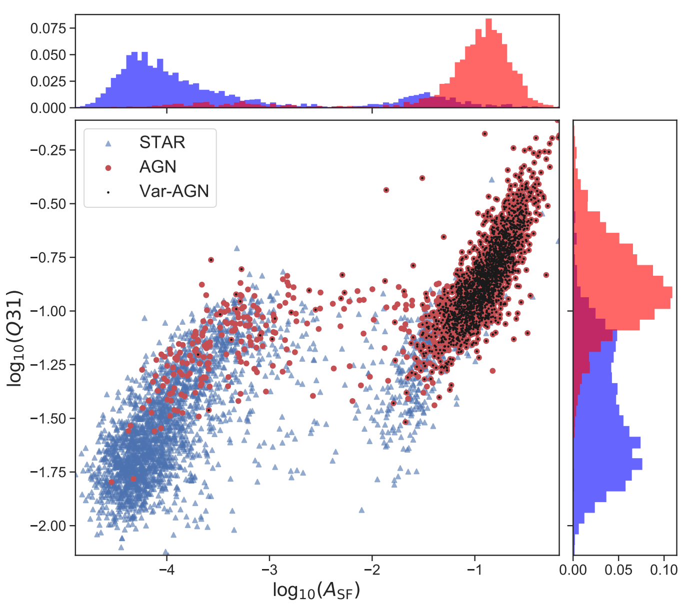

Figure 4. Distribution of the $A_{\rm SF}$ and Q31 features for the labeled set. Blue triangles correspond to stars, and red circles correspond to AGN. We demarc with black dots those AGN classified as variable, according to the definition used in Sánchez et al. (2017).

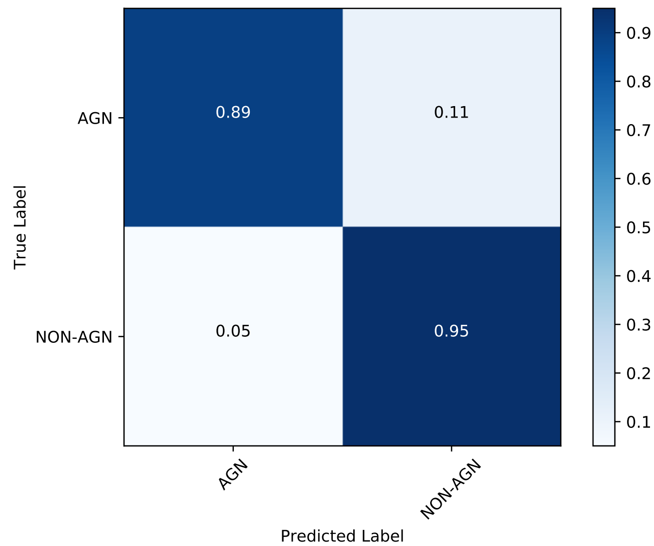

Figure 5. Confusion matrix from testing the RF2 classifier in the test set. True Label represent the classification done from SDSS spectra, and Predicted Label is the outcome of the RF2 classifier.

call, and F1 scores are shown in Table 3. There are not significant differences in the scores comparing with RF1.

In Table 4 we also show the ranking of features for the RF2 classifier. There are not important differences comparing with RF1. It can be seen that the variability features are in general more important than the $r-i$ and $i-z$ colors for the classification. In this case, the most important features are $A_{\rm SF}$, $\sigma_{\rm rms}$, Q31*, and P_{var}. We can also see that $r-i$ is more important than $i-z$ to classify our sources.

3.5. *AGN candidates from QUEST–La Silla*

We applied the trained RF1 and RF2 classifiers to our unlabeled well sampled set of 208,583 light curves. In order to improve the purity of our selection, we considered the predicted class probability ($P_{\rm RF}$) to select the final set of AGN candidates. We defined two samples of AGN candidates: a) a sample with all those sources classified as AGN by the RF classifier (the full–AGN sample), and b) a sample with all those sources classified as AGN by the RF classifier, with a high probability, i.e. with $P_{\rm RF} \geq 0.8$ (the hp–AGN sample). In table 5 we provide a summary with the number of sources classified as AGN in both samples, for the classifiers RF1 and RF2. For the case of the RF1 classifier, there are 17,120 sources in the full–AGN sample, and 5,940 sources in the hp–AGN sample. For the case of the RF2 classifier there are 15,100 sources in the full–AGN sample, and 5,252 sources in the hp–AGN sample.

Table 5. Number of AGN candidates per field, for each classifier

Field	RF1 full–AGN	RF1 hp–AGN	RF2 full–AGN	RF2 hp–AGN
COSMOS	3,968	1,503	3,562	1,201
XMM–LSS	6,441	2,373	5,774	2,106
Elais–S1	3,374	988	2,936	942
ECDF–S	3,337	1,076	2,828	1,003
Total	17,120	5,940	15,100	5,252

Figure 6 shows the $r - i$ vs $i - z$ distribution of the unlabeled set, and the hp–AGN samples for the RF1 and RF2 classifiers. Comparing with Figure 2, we can see that several of our AGN candidates are located in regions of the color–color space where AGN are not normally found, particularly for the case of RF1. As expected, the main difference between the candidates of RF1 and RF2, is the exclusion of sources in the color–color region where we normally find cold stars, for the case of RF2.

We did a visual inspection of the RF1 hp–AGN candidates, located in different areas of the color–color diagram, particularly in those areas where AGN are not normally found, like in the stellar locus. In Figure 6 we show the position in the $r - i$ vs $i - z$ diagram of some of these candidates. Two of these candidates are also present in the RF2 hp–AGN sample. In Figure 7 we show the light curves of these four candidates. From the figure, it can be seen that their are highly variable.

4. CONFIRMATION OF AGN CANDIDATES

4.1. *Confirmation by ancillary data*

4.2. *Spectroscopic follow up of AGN candidates*

4.3. *Main candidates contaminators*

5. DISCUSSION AND CONCLUSIONS

PS was supported by CONICYT through Beca Doctorado Nacional, Año 2013 grant #21130441. PS received partial support from Center of Excellence in Astrophysics and Associated Technologies (PFB 06). PL acknowledges Fondecyt Grant #1161184. LCH was supported by the National Key R&D Program of China (2016YFA0400702) and the National Science Foundation of China (11473002, 11721303).

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

REFERENCES

Abbott, T. M. C., Abdalla, F. B., Allam, S., et al. 2018, ArXiv e-prints, arXiv:1801.03181

Abolfathi, B., Aguado, D. S., Aguilar, G., et al. 2018, ApJS, 235, 42

Allevato, V., Paolillo, M., Papadakis, I., & Pinto, C. 2013, ApJ, 771, 9

Antonucci, R. R. J., & Miller, J. S. 1985, ApJ, 297, 621

Bertin, E., & Arnouts, S. 1996, A&AS, 117, 393

Breiman, L. 2001, Machine Learning, 45, 5

Butler, N. R., & Bloom, J. S. 2011, AJ, 141, 93

Cartier, R., Lira, P., Coppi, P., et al. 2015, ApJ, 810, 164

Covey, K. R., Ivezić, Ž., Schlegel, D., et al. 2007, AJ, 134, 2398

Doi, M., Tanaka, M., Fukugita, M., et al. 2010, AJ, 139, 1628

Gunn, J. E., Carr, M., Rockosi, C., et al. 1998, AJ, 116, 3040

Ivezic, Z., Tyson, J. A., Abel, B., et al. 2008, ArXiv e-prints, arXiv:0805.2366

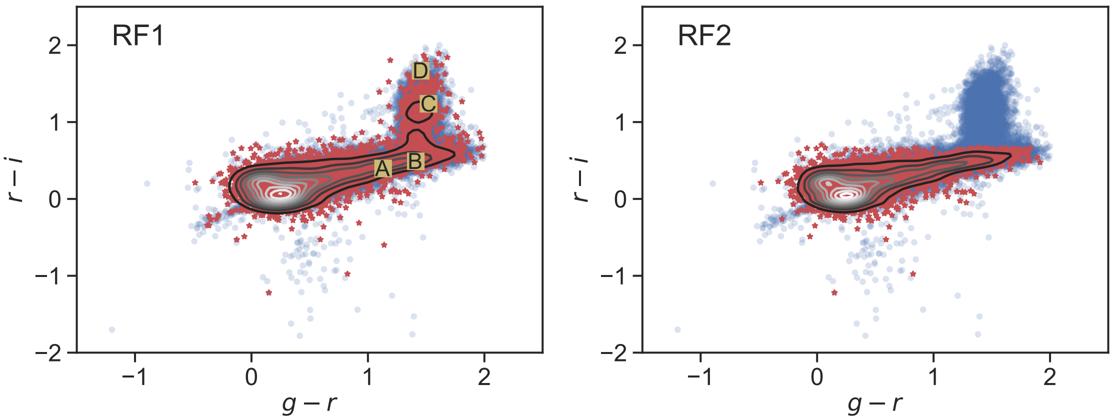

Figure 6. $r - i$ vs $i - z$ diagrams of the unlabeled set (blue circles), and the hp–AGN sample (red stars). In the left panel we show the candidates for the RF 1 classifier, and in the right panel for the RF2 classifier. The contour plots show the distribution of the hp–AGN samples for RF1 and RF2. In the left pane, we mark with yellow squares and letters the position of some RF1 candidates located in the stellar locus.

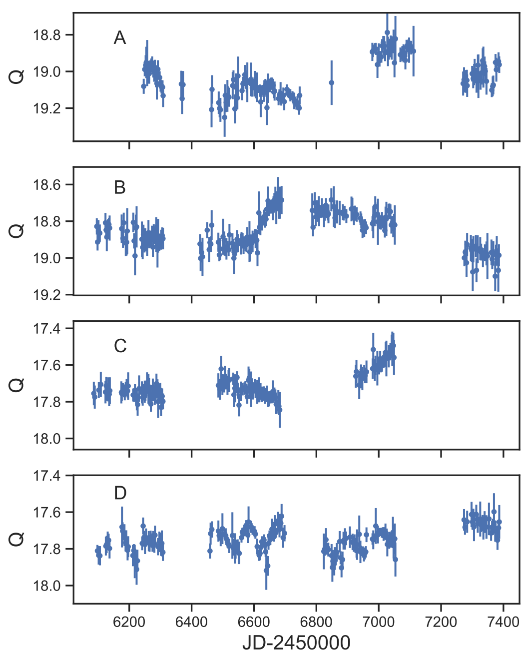

Figure 7. Light curves of some RF1 candidates located in the stellar locus, shown in Fig. 6.

Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, ApJ, 698, 895

Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A., & Uttley, P. 2014, ApJ, 788, 33

Kim, D.-W., Protopapas, P., Bailer-Jones, C. A. L., et al. 2014, A&A, 566, A43

Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68

LaMassa, S. M., Cales, S., Moran, E. C., et al. 2015, ApJ, 800, 144

LSST Science Collaboration, Abell, P. A., Allison, J., et al. 2009, arXiv e-prints, arXiv:0912.0201

McLaughlin, M. A., Mattox, J. R., Cordes, J. M., & Thompson, D. J. 1996, ApJ, 473, 763

Myers, A. D., Palanque-Delabrouille, N., Prakash, A., et al. 2015, ApJS, 221, 27

Netzer, H. 2015, ARA&A, 53, 365

Nun, I., Protopapas, P., Sim, B., et al. 2015, ArXiv e-prints, arXiv:1506.00010

Palanque-Delabrouille, N., Yeche, C., Myers, A. D., et al. 2011, A&A, 530, A122

Rahman, M., Mendez, A. J., Ménard, B., et al. 2016, MNRAS, 460, 163

Richards, G. T., Fan, X., Newberg, H. J., et al. 2002, AJ, 123, 2945

Richards, G. T., Myers, A. D., Gray, A. G., et al. 2009, ApJS, 180, 67

Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10

Ross, N. P., Myers, A. D., Sheldon, E. S., et al. 2012, ApJS, 199, 3

Sánchez, P., Lira, P., Cartier, R., et al. 2017, ApJ, 849, 110

Sánchez-Sáez, P., Lira, P., Mejía-Restrepo, J., et al. 2018, ApJ, 864, 87

Schmidt, K. B., Marshall, P. J., Rix, H.-W., et al. 2010, ApJ, 714, 1194

Sesar, B., Ivezić, Ž., Lupton, R. H., et al. 2007, AJ, 134, 2236

Tie, S. S., Martini, P., Mudd, D., et al. 2017, AJ, 153, 107

VanderPlas, J., Connolly, A. J., Ivezic, Z., & Gray, A. 2012, in Proceedings of Conference on Intelligent Data Understanding (CIDU), pp. 47-54, 2012., 47–54

VanderPlas, J. T. 2018, ApJS, 236, 16

York, D. G., Adelman, J., Anderson, Jr., J. E., et al. 2000, AJ, 120, 1579