Exploratory data analysis of the quasar candidates catalog

Franciszek Humieja
fhumieja@oa.uj.edu.pl

May 13, 2024

Contents

1 Description of the database

I provide the analysis of the catalog of quasar[1] candidates made by [Richards et al., 2015]. In this section, I characterize the data set and the way it was built.

1.1 Origin of the data set

[Richards et al., 2015] describe in detail the procedure of obtaining their data as well as the astrophysical analysis of the results. Here, I report the main steps very concisely; the inquisitive reader is referred to the cited paper.

To build their catalog of candidates for quasars, the authors utilized optical imaging data of a sample of astronomical objects that require classification; this constitutes the *test* set. They selected such objects from the *Sloan Digital Sky Survey* (**SDSS**) optical data that had also been included in the mid-infrared (**MIR**) survey data of *Wide-field Infrared Survey Explorer* (**WISE**) or *Spitzer Space Telescope*. This specific combination of the optical+MIR data of the investigated objects is a novelty in this research field, bringing more detailed information of the objects and allows for more robust classification as potential quasars than the approach using only visible or MIR data alone. The test set contains 50,225,630 objects to be classified.

Candidates for quasars are classified by the Bayesian kernel density algorithm which is described in Section 3.2 of [Richards et al., 2015]. The classification is based on the *training* set which is built of the *SDSS* objects that are spectroscopically[2] confirmed to be quasars (based on 14 catalogs with spectroscopic data). The authors made the data publicly available in the file master.dat[3]. As in the case of the test set, only the quasars possessing MIR photometry from either *WISE* or *Spitzer* were chosen for the final quasar training set. It contains 157,701 confirmed quasars. Moreover, another training set – the "star" training set – was prepared out of the objects that had been proven not to be quasars, in order to reinforce the quasar classification; this set contained around 700,000 objects.

1.2 The final catalog

The Bayesian kernel density algorithm, using the described training sets, classified 885,503 candidates for quasars in the test set. Among them, 748,839 were classified as low-redshift[4] ($0 < z < 2.25$), 205,060 as mid-redshift ($2.15 < z < 3.55$) and 13,060 as high-redshift ($3.45 < z < 5.5$); note that the redshift ranges overlap, so the last three numbers do not sum up to the total quasar candidates number. Of the total number of candidates, 733,713 lacked spectroscopic confirmation, 305,623 had not been classified as quasar candidates and 150,453 objects had already been known to be quasars.

The obtained catalog of quasar candidates is written to the file cand.dat[5]. This is the data base I am going to investigate in this report. Feature names along with their explanation can

[1] A quasar (quasi-stellar object, QSO) is a very luminous type of an active galactic nucleus (AGN). Active galaxy is a galaxy which emits strong radiation at its central compact region, containing a supermassive black hole and matter accreting around it.

[2] Spectrometry in astronomy is an investigation of light (spectrum) that passes through a spectroscope, being a prism or a system of prisms attached to a telescope. It enables to get information that standard photometry (i.e., measuring the light as it comes, without refraction on prisms) does not give, like spectroscopic redshift, etc.

[3] Available at `https://cdsarc.cds.unistra.fr/ftp/J/ApJS/219/39/`.

[4] Redshift is a shift of the spectrum of light towards lower frequencies (red). In extragalactic astrophysics it is mostly due to the expansion of the Universe, so the further away, the bigger the redshift (roughly). Although, a cosmological spacetime is constructed in such a way that it is impossible to define *unambiguously* any general "distance", there exist specific types of distances, e.g., luminosity distance or angular diameter distance, but in general, they give different results for the same object. Redshift, in turn, is not even a type of distance, but at least gives some most basic and experimentally easy-to-obtain indication of the distance.

[5] Available under the same address as the file master.dat above.

be found in Table 1. The long names from this table were being used in the original paper and the short names (in bold) are used as the feature names in the database and my code. Here, I provide additional information on the most important variables:

- `RAdeg` and `DEdeg` – the celestial coordinates (right ascension and declination) of an object in degrees.
- `Class` – the spectral classification result of an object. If the spectral analysis was available, the object was classified either as quasar (*QSO*), galaxy (*GALAXY*), star (*STAR*), compact emission line galaxy (*CELG*) or hard to interpret (*??*). Otherwise, if no spectrum is provided, the object is unclassified (*U*).
- `zsp` – spectroscopic redshift for the objects with spectrum available.
- `umag`, ..., `zmag` – *SDSS* magnitudes in optical and near-infrared *u, g, r, i* and *z* passbands.
- `3.6mag` and `4.5mag` – *Spitzer* magnitudes in mid-infrared *ch1* and *ch2* passbands (having effective wavelength midpoints at 3.6 μm and 4.5 μm, respectively).
- `Ymag`, ..., `Kmag` – *UKIDSS* and *VHS* magnitudes in near-infrared *Y, J, H* and *K* passbands. Provided to only a fraction of objects.
- `FUV` and `NUV` – *GALEX* magnitudes in ultra-violet passbands. Provided to only a fraction of objects.
- Columns starting with "`e_`" – standard errors (uncertainties) of the corresponding values.
- `gisig` – an indicator of distance (in units of confidence level) from the average curve in a regression-like problem of finding photometric redshift.
- `zphO` – the best estimate for photometric redshift calculated through regression trained with a relationship between spectroscopic redshift and *ugriz* magnitudes.
  - `b_zphO` – the minimum estimate for photometric redshift (*ugriz*).
  - `B_zphO` – the maximum estimate for photometric redshift (*ugriz*).
  - `zphOP` – a probability of the true redshift to be between minimum and maximum estimates of photometric redshift (*ugriz*).
- `zphIR` – similar to `zphO` but trained using *ugrizJHK* magnitudes. Available only for those sources with *JHK* photometry performed.
- `pm` – proper motion (speed of motion in the sky) of an object in milliarcseconds per year.

In their astrophysical analysis of the catalog, the authors define "robust" candidates as those having `ZPHOTPROB` > 0.8 and abs(`GI_SIGMA`) ≤ 0.95[6]. These criteria were met by 517,586 candidates in the catalog. Of those, only 717 were known nonquasars and 114,120 were known quasars. For the high-redshift candidates, the robustness criterium was further restricted to nondetections in *GALEX* and $i < 22$[7]. There were 10,955 such sources, of which 6779 had no previous spectroscopic or photometric quasar classification.

[6] These two parameters are related to the photometric redshift which is estimated by the authors based on the spectroscopic redshift (the one measured by the spectrometry) using an algorithm cited therein.

[7] u, g, r, i, z (and other) are the *magnitudes* in different passbands of the light frequencies. The magnitude is a measure of brightness and is proportional to the logarithm of the intensity of light of the astronomical object. The aspect ratio is usually -2.5, so the lesser the magnitude is, the brighter the object shines.

2 The plan for data exploration

I plan to clean the catalog data (from the cand.dat file) and check outliers and relations between variables. I wish to pay attention to errors given for some of the variables and include the uncertainty they provide into the analysis. Among the investigated dependencies between the variables, the standard and well-known relationships in astrophysics, like color-redshift or color-color diagrams[8], should be tested since they may provide a base for future regression and classification problems. Moreover, it would be beneficial to compare the correlation between target and feature variables for the "robust" subsample, defined by the authors. Finally, hypothesis tests ought to be provided, for example, to compare the properties of different classes of objects from the catalog.

3 Data cleaning and feature engineering

3.1 Data wrangling

The imported data is quite large with 885503 rows and 60 columns. In the first step, it needed some wrangling. Firstly, the first two columns, `RAdeg` and `DEdeg`, being celestial coordinates of the objects, were not separated in cand.dat with the given separator, so they were imported as one text (i.e., the `object` *dtype*) column. Although this issue would be trivial to fix with a text editor like Vim, I decided it would be more instructive to make it programmatically in Python – with basic and the fastest string slicing, since all coordinates have constant length. After this operation, the number of columns is 61, with the full agreement with Table 1.

Secondly, the column names needed to be imported from another file and put into the catalog dataframe.

3.2 Data cleaning

3.2.1 Removing unnecessary spaces

Because of the form of the source file, cand.dat, the categorical columns, of the *pandas' dtype* `object`, still had unnecessary spaces at the end. There was one such column, `Class` and the issue was resolved with the `str.strip()` method.

3.2.2 Handling duplicates

There were no whole duplicated rows in the data. However, duplicates appeared only after contracting the subset of searching to the coordinate columns (i.e., `RAdeg` and `DEdeg`). This criteria revealed the nine mismatched pairs of objects, mentioned in [Richards et al., 2015]. It was an effect of mismatching optical *SDSS* data with mid-IR *WISE* and *Spitzer* data (two different IR sources matched to one optical source). These 9 objects belong to the two biggest classes *U* and *QSO*, so they might not affect the analysis significantly. However, looking at the color-color diagrams in the original paper, we can see that some of these points are likely to be outliers. So, for purity, all these 18 examples were ruled out of the catalog.

3.2.3 Handling missing values

Some columns of the data contain missing values. Since these features can be treated as "nonobligatory" (i.e., observations of these variables for some examples are not available yet,

[8] *Color* in astrophysics is a difference in magnitudes at two different passbands, for example, $u - g$ or $r - z$. It provides useful information, for instance when a full spectrum is not available. Since magnitude is proportional to the logarithm of flux density, color is the logarithm of the ratio of flux density at the corresponding passbands.

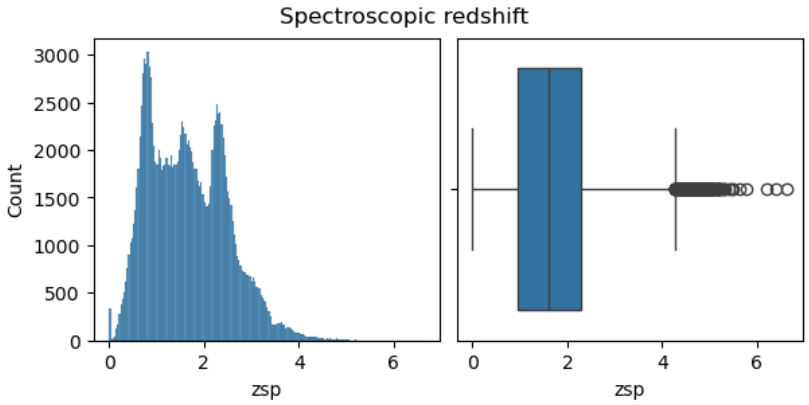

Figure 1: Distribution histogram with 200 bins (left) and box plot (right) of the spectroscopic redshift, `zsp`.

so we have to deal without them), we are not going to fill in the missing values or remove the entire columns or rows. For some future analyses, we will be restricting only to the non-null examples with respect to given features.

3.2.4 Handling outliers

Finding outliers should be supported by analyzing the distributions of the variables, as well as their interpretation. We inspect the distribution and the skewness of a given variable and then look at the boxplot to search for outliers.

Let us start with the variables for which the outlier notion does not make sense. This is the case for the coordinates, `RAdeg` and `DEdeg`, whose distribution is a matter of choice, availability by the hardware and geography, necessity to avoid the Milky Way's dense regions, etc. – so we expect neither any regular distribution nor the need for removing outliers. Similarly, no interpretation of outliers are expected in the indicator features: `gisig`, `Leg`, `Uni`, `Prim` and `DB` being indeed categorical features with numerical (`float` or `int`) values.

The distribution and the box plot of the spectroscopic redshift, `zsp`, is shown in Figure 1. We can see peaks in the histogram which are caused by the selection effects in the *SDSS* samples coming from various data releases, see [Richards et al., 2015]. The *Data Release 7* (*DR7*) quasar sample has a peak at $z \approx 1.5$, while the *DR10* quasar selection was optimized for $z \approx 2.5$, with contamination coming at $z \approx 0.8$. At low redshift, there are, moreover, losses coming from cuts imposed on the *WISE* infrared data. Thus, since the distribution shape is subjected to the selection effect in the domain of redshift, we're not going to remove any outliers in `zsp`.

The AB magnitudes, `umag`, ..., `zmag`, `3.6mag`, `4.5mag`, have roughly bell-shaped distributions, centered near 20 mag (Figure 2), with the *SDSS ugriz*-filter magnitudes having low skewness (between -0.5 and 0.5) and the *Spitzer* 3.6 μm- and 4.5 μm-filter magnitudes having moderate skewness (between -1 and -0.5 or 0.5 and 1). In order to get the full picture of outliers, we should classify them on the basis of the magnitude values, along with the information of the uncertainties provided, `e_umag`, ..., `e_4.5mag`. Magnitudes seem not to have significant outliers; their errors, in turn, show the presence of strongly outstanding points, as shown in the boxplots in Figure 3.

In this case, we allow, however, for a more systematic cut of outliers, basing on the physical interpretation of magnitude and their uncertainties. Namely, standard errors greater than 15–20 mag seem to be absurd for magnitudes with the mean around 20 mag. It means that, within a 0.95 confidence level, a given star can be as well so faint that it cannot be detected by any of our giant telescopes, as so bright that it becomes one of the brightest stars

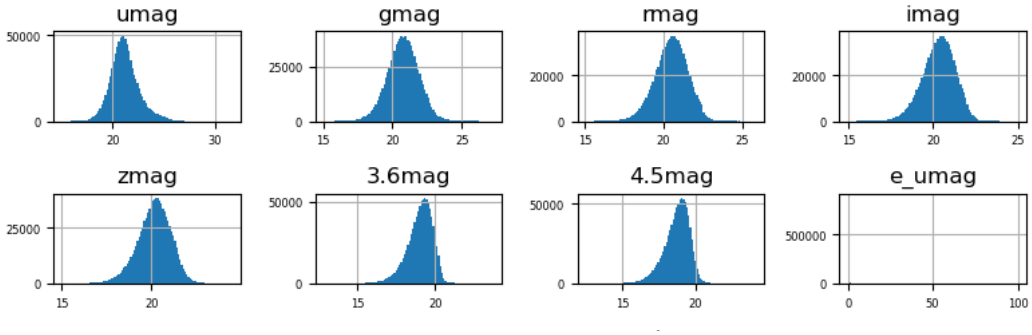

Figure 2: Histograms of the *SDSS* and *Spitzer* magnitudes in optical+near IR *ugriz* filters and mid-IR 3.6 μm and 4.5 μm filters. Each histogram contains 100 bins.

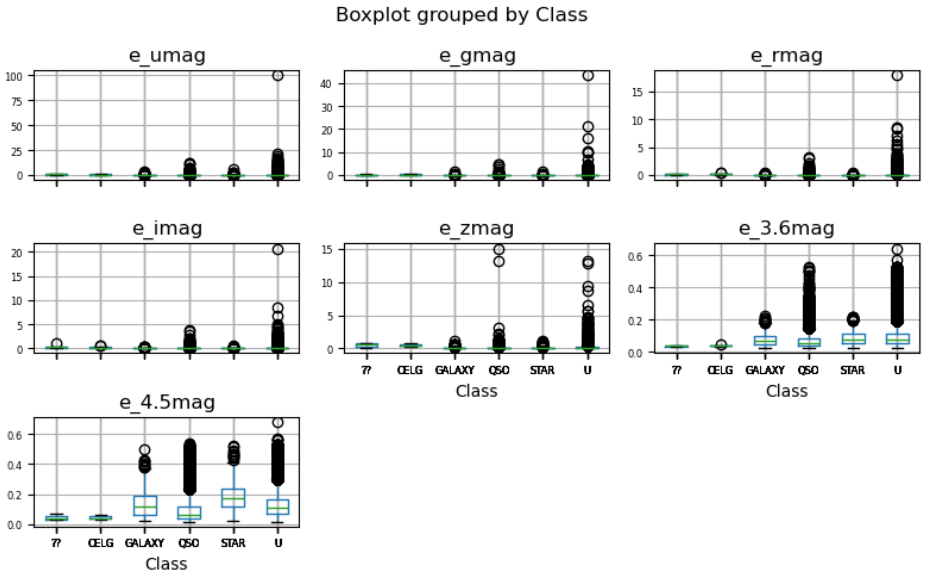

Figure 3: Box plots of the errors of the *SDSS* and *Spitzer* magnitudes in optical+near IR *ugriz* filters and mid-IR 3.6 μm and 4.5 μm filters.

in our sky. Therefore, it seems not to be radical to cut off all the objects whose errors are greater than 10–15 mag. We then decided to remove all the objects whose *SDSS* magnitude errors exceed 10 mag (the *Spitzer* magnitude errors do not exceed 0.7 mag). Such a threshold turned out not to be drastic – it cast aside only 31 most uncertain magnitude measurements.

Next, we move on to the near-IR Vega magnitudes from *UKIDSS* and *VHS* surveys, that is `Ymag`, ..., `Kmag`, and UV magnitudes from *GALEX*, that is the `FUV` and `NUV` columns. These magnitudes had been available only for a fraction of sources and were included in the catalog, mainly to improve the photometric redshift estimate. There was only one outlier point in this set, in a value of `e_Ymag` making a corresponding magnitude highly uncertain.

Fluxes, `Fu`, ..., `F4.5`, will not be in the interest of our analysis, because all the information they provide (along with the Galactic absorption correction) is stored in the *SDSS* and *Spitzer* magnitudes we already discussed. No outlier investigation was made in this set of variables.

Finally, the photometric redshifts, where there were four outliers. Namely, four sources had an erroneous value of 100.0 assigned to each of the columns: `b_zph0`, `zph0`, `B_zph0` and `zph0P`. These objects were removed. Figure 4 shows distributions of photometric redshifts after the outlier removal. As we can see, these distributions are highly irregular and have many peaks which is a result of the work of the algorithm used by [Richards et al., 2015] for finding photometric redshift trained on the correlation between spectroscopic redshift and

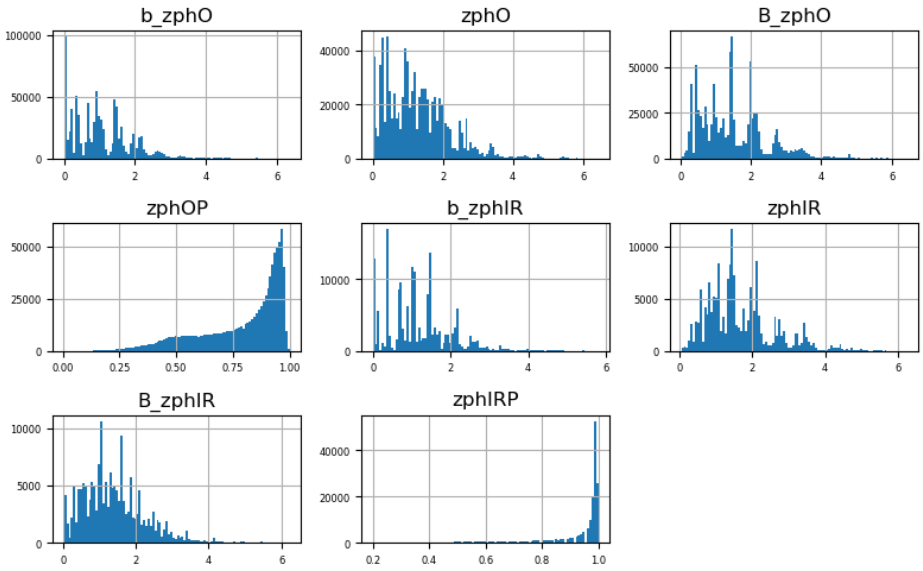

Figure 4: Distributions of the best estimates for photometric redshift `zphO` and `zphIR` along with their minimum estimates `b_...`, maximum estimates `B_...` and probabilities for the real redshift to be between the minimum and maximum estimate, `...P`. Each histogram was made with 100 bins.

colors of quasars.

In total, thirty-six examples were removed from the data in the entire process of data cleaning.

3.3 Feature engineering

3.3.1 Colors

We add colors to our analysis which are differences of pairs of magnitudes. We choose only neighboring pairs of the *SDSS* and *Spitzer* magnitudes and add new features: `u-g`, `g-r`, `r-i`, `i-z`, `z-3.6` and `3.6-4.5`. Since we provide full error analysis, we also add features with values of errors for each color: `e_u-g`, `e_g-r`, `e_r-i`, `e_i-z`, `e_z-3.6` and `e_3.6-4.5`. The value of uncertainty (error) u for a given color, being a difference of two magnitudes m_1 and m_2, with uncertainties $\mathrm{u}(m_1)$ and $\mathrm{u}(m_2)$, is calculated with the basic formula for the propagation of uncertainty for a complex measurand,

$$\mathrm{u}(m_1 - m_2) = \sqrt{\mathrm{u}(m_1)^2 + \mathrm{u}(m_2)^2}.$$

3.3.2 Logarithm of colors

It turned out the first three of the above colors have high skewness factors, greater than 1.5. To decrease these numbers, new features with a logarithm of colors were made, `log_u-g` and so on. In the case of the $u - g$ color, this reduced the skewness from 2.73 to 2.32, while for the $g - r$ color from 2.88 to 1.79, and for $r - i$ from 1.63 to 0.89. So, only in the $r - i$ case, the log transformation dropped the skewness to a moderate value.

It is worth noting that we can slightly fine-tune the skewness by manipulating the constant shift s for each color inside the logarithm, $\log(m_1 - m_2 + s)$. However, we should not minimize the skewness as much as we wish, because in that case, huge outliers emerge. Those outliers "balance" the skewness parameter, but it is not about fixing the real skewness. In short words: assessing the shape of the distribution qualitatively is more important than depending entirely on a mere skewness parameter.

The uncertainties u of logarithms of colors were also calculated and added with names `e_log_u-g`, and so on. Again, the formula for error propagation,

$$\mathrm{u}[\log(m_1 - m_2)] = \frac{\sqrt{\mathrm{u}(m_1)^2 + \mathrm{u}(m_2)^2}}{m_1 - m_2},$$

was utilized.

3.3.3 Polynomial of photometric redshift

As we will see in Sections 4.2 and 4.3.2, there is evidence in our data for a higher-order relationship between colors and redshift. This is why polynomial features of photometric redshift `zphO` should be prepared for future linear regression calculation. We use the *sklearn* package, containing the `PolynomialFeatures()` transformer which calculates monomials out of a given feature, up to a desired degree. Using this, we created features with `zphO` raised to the power from 2 through 10, such that we could fit a polynomial of degree 10 in the future analysis.

3.3.4 Feature scaling

For the purpose of future classification and regression problems, the features containing all magnitudes (*SDSS, Spitzer, UKIDSS/VHS* and *GALEX*; see Table 1), colors (*SDSS* and *Spitzer*, as in Section 3.3.1) and photometric redshifts, `zphO` and `zphIR`, should be scaled to improve algorithms performance. Actually, it turns out all the features in each of the considered problem have quite similar value ranges, so feature scaling seems not to be that necessary. Nevertheless, for this project, it is very instructive to perform feature scaling using the *scikit-learn* library.

The above features were transformed with `StandardScaler()` which subtracts the mean of each feature from each value and divides by the feature's standard deviation. The error features were scaled only by division by the standard deviation of the corresponding features of values, because, for any value subject to uncertainty $x \pm \mathrm{u}(x)$ with the value's mean μ and standard deviation σ, the following identity holds

$$\frac{x \pm \mathrm{u}(x) - \mu}{\sigma} = \frac{x - \mu}{\sigma} \pm \frac{\mathrm{u}(x)}{\sigma}.$$

In the case of the photometric redshift, the minimal values, `b_zphO`, `b_zphIR`, and maximal values, `B_zphO`, `B_zphIR`, were scaled as errors (namely, the difference between maximum and minimum has an interpretation of an error) and the probabilities, `zphOP`, `zphIRP`, were not scaled, by their nature.

3.3.5 One-hot encoding

The data has one categorical variable of our interest, which is `Class` with possible values *U*, *QSO*, *STAR*, *GALAXY*, *CELG*, and *??*. We execute one-hot encoding of this feature using the `pandas.get_dummies()` method.

4 Key findings and insights

4.1 Quasar candidates and contaminants

The `Class` column indicates the spectral classification of the objects. The total number of objects that have been gone through by the quasar-searching algorithm in [Richards et al., 2015] and, next, were filtered by the above duplicates and outliers removal, is 885449. Among them, the following numbers of examples belong to each class according to their spectroscopic classification:

Unclassified (no spectrum available) (U):	733669
Quasars (QSO):	150443
Stars ($STAR$):	743
Galaxies ($GALAXY$):	557
Compact emission line galaxies ($CELG$):	32
Difficult to classify ($??$):	5

So, as we can see, a small fraction (1337) of the objects classified as quasar candidates, turned out to be nonquasar contaminants (stars, galaxies and 5 unrecognized objects) thanks to the availability of their spectra. The objects with spectra not available are marked as U, and this class will be mainly of our interest in the future quasar classification tasks. Most likely, this class has also a small fraction of contaminants, though. The future algorithms could be trained on the subsets of QSO and the contaminants to reclassify the U subset, improving the original classification.

4.2 Main correlations

In Figure 5, we showed examples of the main important relationships between pairs of variables of the data, which include spectroscopic-photometric redshift, magnitude-redshift, color-redshift or color-color correlations. In addition to the data points in grey, there are contour plots made with a kernel density estimation algorithm. White contours were prepared without taking weights into account and blue contours were calculated with weights. For magnitudes and colors, the weights were inverses of the corresponding errors squared (or inverses of the sum of the corresponding errors squared, for the case of uncertainties for both variables in the plot) – according to the typical convention for weighted formulas and algorithms. In the case of `zph0`, the weight is $\texttt{zph0P}/(\texttt{B\_zph0} - \texttt{b\_zph0})^2$. Discrepancies between the weighted and unweighted contours are visible. Since the error values can differ even by orders of magnitude, the weighted methods are more reliable, though, being robust with respect to highly uncertain points.

In the spectroscopic redshift vs. photometric redshift plot, we can see the outliers from the expected linear dependence, as reported by [Richards et al., 2015]. These outliers constitute a small fraction of all data and are a result of a mismatch between the value of redshift and the given color values of an example by the algorithm used by the authors of the cited paper.

In the color-redshift plots, we can see a strict nonlinear dependence. They are more informative than the magnitude-redshift plots which appear to be more blurred, so we will use the former in the following analysis.

The color-color diagrams may be helpful in further classification analysis as they separate physical features of astronomical objects. Here, we can see that the weighted contours of densities are concentrated within relatively small regions, compared to the unweighted contours. These weighted contours indicate the regions where small errors are most prevalent.

Furthermore, the proper motion variable, `pm`, is very weakly correlated with other variables and provides no interesting information.

In Figure 5, we plotted the densities contours for the examples at all redshifts altogether. Due to the overrepresentation of the low- and mid-redshift objects relative to the high-redshift ones, we need to split the data according to redshift value to show the densities in redshift groups separately. We will do this in the next subsection, taking moreover only the "robust" subsample into account.

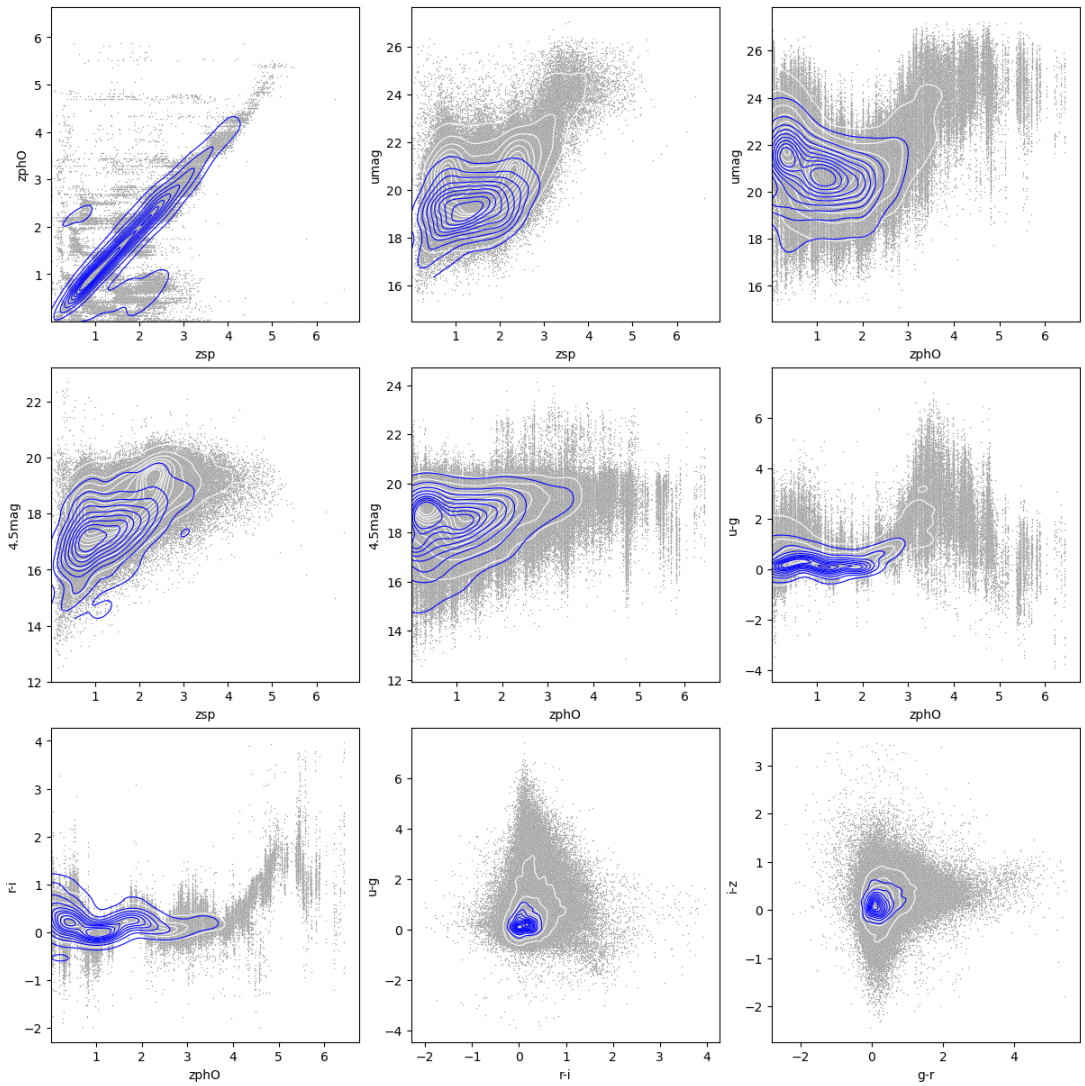

Figure 5: Scatter plots (grey points) between redshifts, magnitudes and colors for all data. Contours were made with kernel density estimation algorithm, each with 10 contours, starting from the level of 5% of relative density. White contours were made without taking weights into account, while blue contours – with weights.

4.3 Robust subsample analysis

4.3.1 Filtering the subsample and analysis of distributions

The following analysis will be restricted to the robust subset of the data, as defined by [Richards et al., 2015]. Namely, "robust" candidates are those which satisfy:

1. `zphOP` > 0.8 and
2. $abs($`gisig`$) \leq 0.95$.

For high-redshift candidates, $3.5 < z < 5$, we further restrict the requirements for robust sources to:

3. `NUV` and `FUV` are *NA*, i.e., nondetections in *GALEX* (real high-z quasars are unlikely to be detected), and
4. `imag` < 22 (point-like objects of the 22nd or greater magnitude have a significant probability of being galaxies at these redshifts).

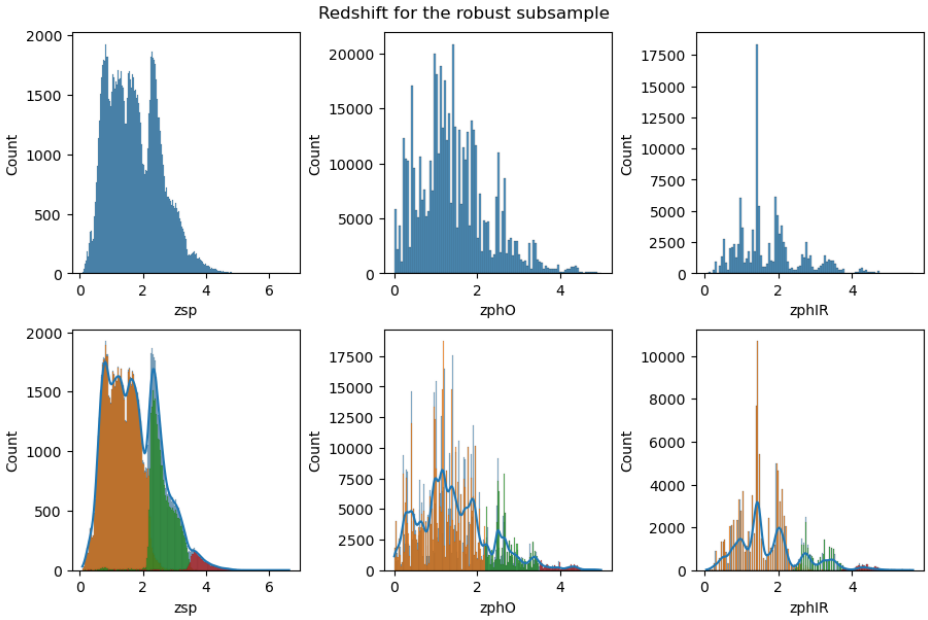

Figure 6: Spectroscopic redshift (`zsp`) and photometric redshift (`zphO` and `zphIR`) distribution in the robust subsample. In the upper row, the bin numbers are 200 for spectroscopic and 100 for photometric redshifts. In the lower row, the bin number is 200 for each plot. Orange (slightly rotten) bins represent low-redshift objects, green – mid-redshift and red – high-redshift. The average distribution (KDE) was plotted with the blue line in the lower row.

We mark robust sources with a new boolean column `robust` and define the following division with respect to the photometric redshift `zphO`:

$$\begin{aligned} &\text{low-redshift object:} && 0 \leq \texttt{zphO} < 2.2, \\ &\text{mid-redshift object:} && 2.2 \leq \texttt{zphO} < 3.5, \\ &\text{high-redshift object:} && 3.5 \leq \texttt{zphO} < 5, \end{aligned}$$

based on the division given by [Richards et al., 2015] where the cutoff at $z = 5$ has been applied because redshifts greater than ~ 5.5 require additional care during analysis.

The total number of objects satisfying these conditions is 516169. Distribution of the robust objects in each spectral class and redshift range is shown below:

Class	low z	mid z	high z	all
U	345490	48152	7858	401500
QSO	78736	32246	2975	113957
STAR	73	339	48	460
GALAXY	164	53	28	245
CELG	1	4	1	6
??	0	1	0	1

There is still a small number of contaminants in the robust subset. For inspecting the relationships in the sample of confirmed quasars and quasar candidates, diminishing the influence of contaminants, we will have these objects unflagged as "robust" and flagged as "robust contaminant". Hence, we are going to eventually work with the subset of 515457 objects of *U* and *QSO* classes only.

Spectroscopic and photometric redshift distributions for the robust subset are shown in Figure 6. We can notice that `zsp` reveals a new peak at $z \approx 1$ in this subsample. The photometric redshifts still have a "shredded" shape (due to the algorithm in the original paper) but, after smoothing with the moving average, `zphO` could resemble a bell-shaped

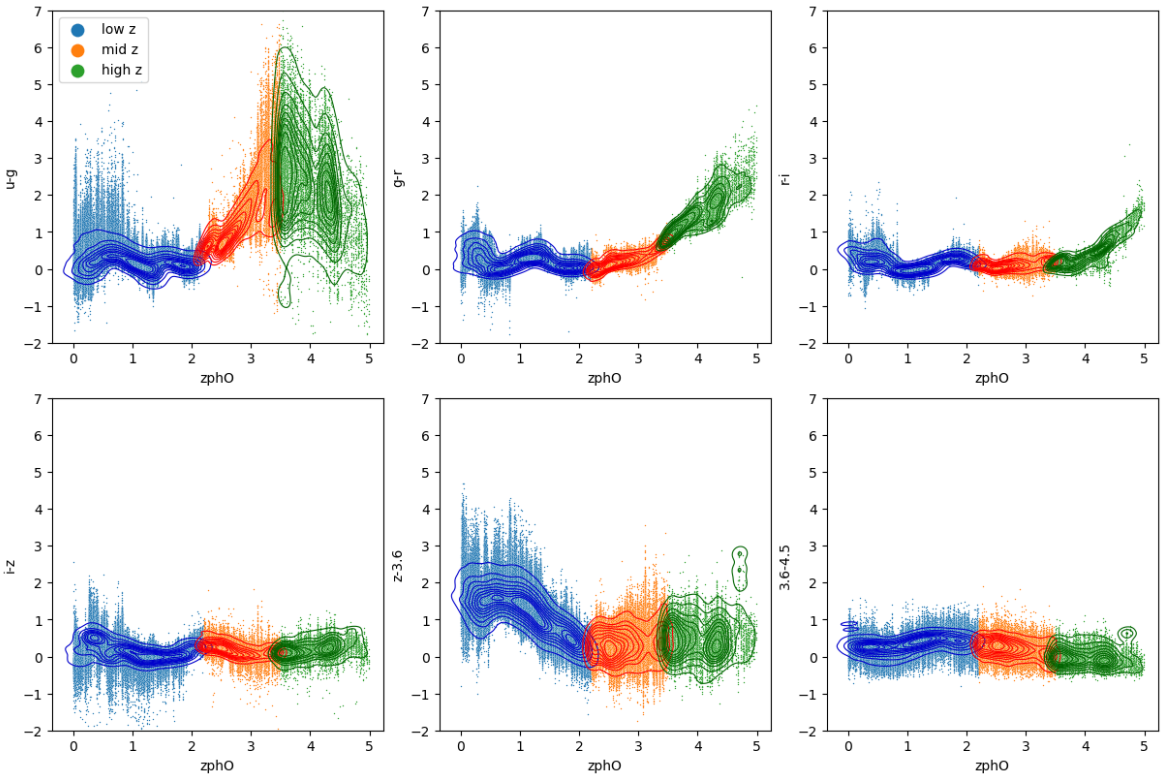

Figure 7: Scatter plots of colors versus `zphO` for the robust subsample. Contours were calculated with weights and have the same settings as before, i.e., 10 contours starting with 5% of relative density.

distribution to some extent. For other variables, limiting to the robust subsample decreases the skewness (gets the skewness parameter closer to zero) in most of the cases. Moreover, it rules out outliers and the points near the boundaries of the most dense areas of points in scatter plots in Figure 5, decreasing the variance of points in those correlations.

4.3.2 Analysis of correlations

Figure 7 shows scatter plots of colors versus the photometric redshift `zphO` in the robust subsample. For each of the three redshift classes, separate KDE contours were prepared with weights calculated as before. A strict nonlinear relationship is visible in each plot, suitable for fitting a polynomial, which would help derive redshift out of the photometric measurements with no spectrometry available. The plots for `u-g`, `g-r`, `r-i` and `z-3.6`, show good variability over `zphO` which would be promising for predicting the unknown redshift. For `g-r` and `r-i`, the high-redshift points stand out of the constant level at lower redshift values, while for `u-g` and `z-3.6`, they are the low-redshift objects that stand out from the rest. Unfortunately, `u-g` shows a large variance of color for high redshift. Finally, the variability of `i-z` and `3.6-4.5` is near to constant with respect to `zphO`.

In Figure 8, color-color diagrams are shown for the robust subsample split into the three classes of redshift values. For each class, separate weighted KDE contours were made. The separation of regions with different redshifts is clearly visible in each plot. To get a more accurate classification, one could construct 3D color-color-color plots (or even higher-dimensional spaces) where well-defined regions of redshift regimes can be possibly fitted. Combining color-color and color-redshift plots one could derive a redshift progression curve in the color-color space which could enhance the redshift classification and provide a deeper understanding of the quasar color-color space.

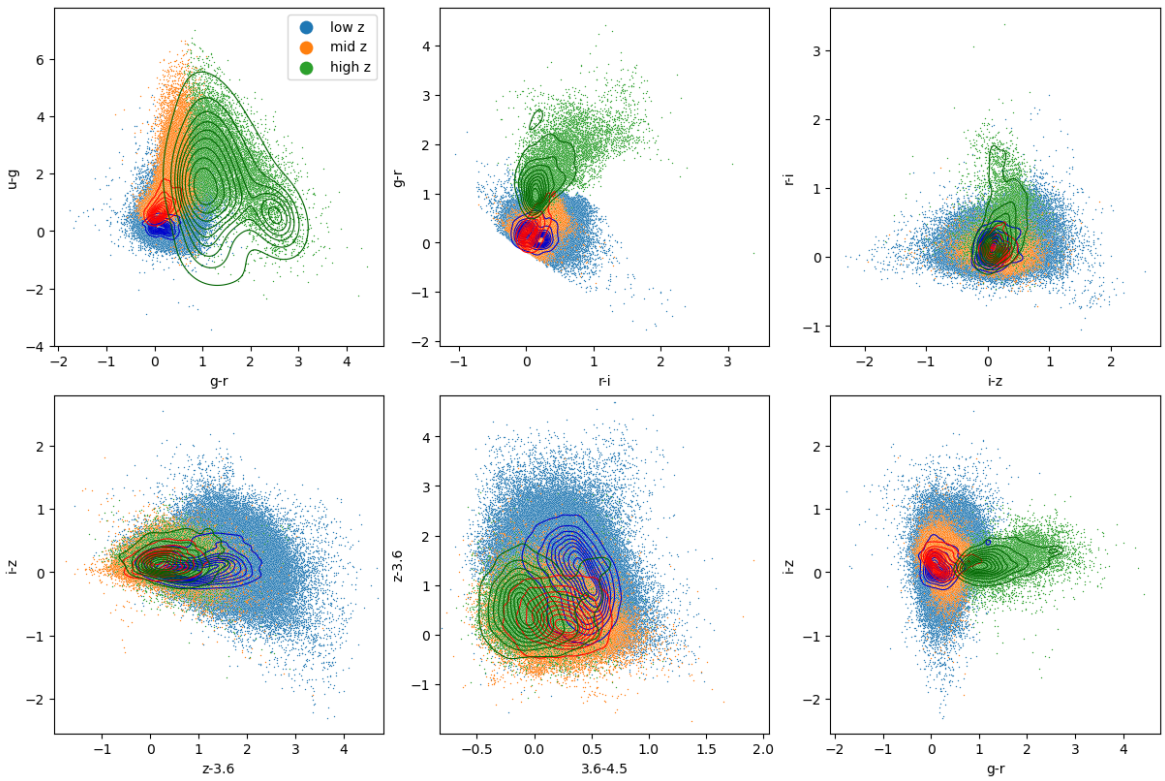

Figure 8: Color-color plots for the robust subsample. Contours were calculated with weights and have the same settings as before.

Finally, in Figure 9, the correlation between chosen *UKIDSS*/*VHS* and *SDSS*/*Spitzer* magnitudes is depicted. We can see a strict linear relationship. Moreover, Figure 10 shows the correlation heatmap between these magnitudes based on the Pearson linear correlation coefficient. It shows that the *SDSS* and *Spitzer* magnitudes are very well linearly correlated with the *YJHK* magnitudes, with the exception of `umag` for which disturbances from the linear trend can occur, as can be seen in the upper left panel of Figure 9. Such a relationship can enable us to build a regression model for obtaining unknown values of near-infrared *YJHK* magnitudes out of known values of *ugriz* and mid-infrared magnitudes. Moreover, in Figure 9, the separation of the redshift regimes is visible which can decrease standard deviation (i.e., uncertainty) for such regression significantly, provided the redshift of an object is known. We should, however, be aware that the *YJHK* magnitudes in our data set have not been corrected for Galactic extinction (as the *ugriz* and the MIR magnitudes).

5 Statistical hypotheses and a significance test

The analysis we performed so far was summarizing the main characteristics of the data mostly with visual methods. However, some of the research problems need to be resolved with formal hypothesis tests. Here, we formulate three important null hypotheses H_0 that need to be tested in order to conduct further analysis.

1. H_0*: Objects of all spectral classes have an equal mean of each color.* We should find a way to extract contaminants in the unclassified (U) objects. It would be ideal, if there were any parameter which differentiated classes at a significant level. Such a parameter could be for instance the colors; if this hypothesis were rejected, there would be a possibility for a machine learning algorithm to seek likely contaminants within the U class.

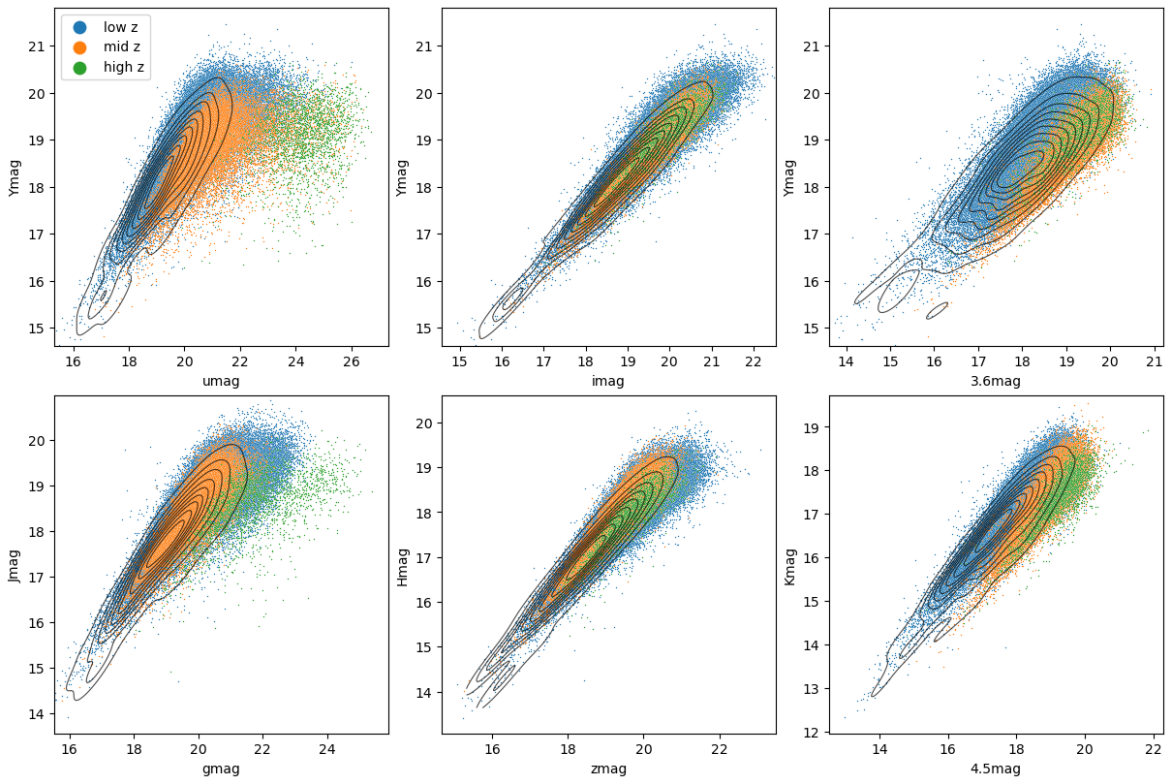

Figure 9: Scatter plots of *UKIDSS / VHS* magnitudes versus *SDSS / Spitzer* magnitudes for the robust subsample. Contours were calculated with weights and have the same settings as before.

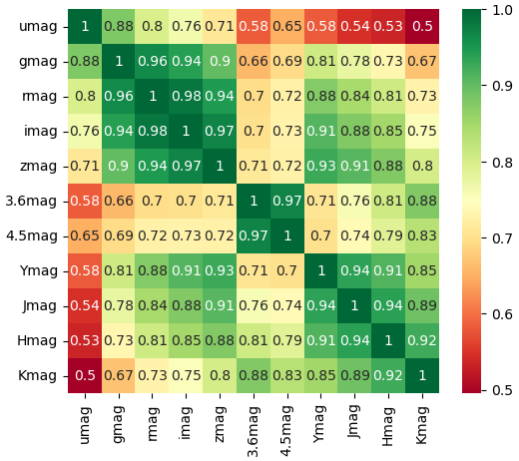

Figure 10: Correlation matrix of *UKIDSS / VHS*, *SDSS* and *Spitzer* magnitudes with Pearson linear correlation coefficients for the robust subsample.

2. H_0: *The mean* `zsp` *is equal to the mean* `zphO` *for the objects having the spectrometry performed.* We need to check if the algorithm for calculating the photometric redshift out of the relationship of the *ugriz* colors and the spectroscopic redshift utilized by [Richards et al., 2015] gives reliable results at a given significance level. Rejection of this hypothesis would mean that the algorithm works incorrectly (for example, when the colors vs. `zsp` relationship has too little variability to be efficient in designating the photometric redshift).

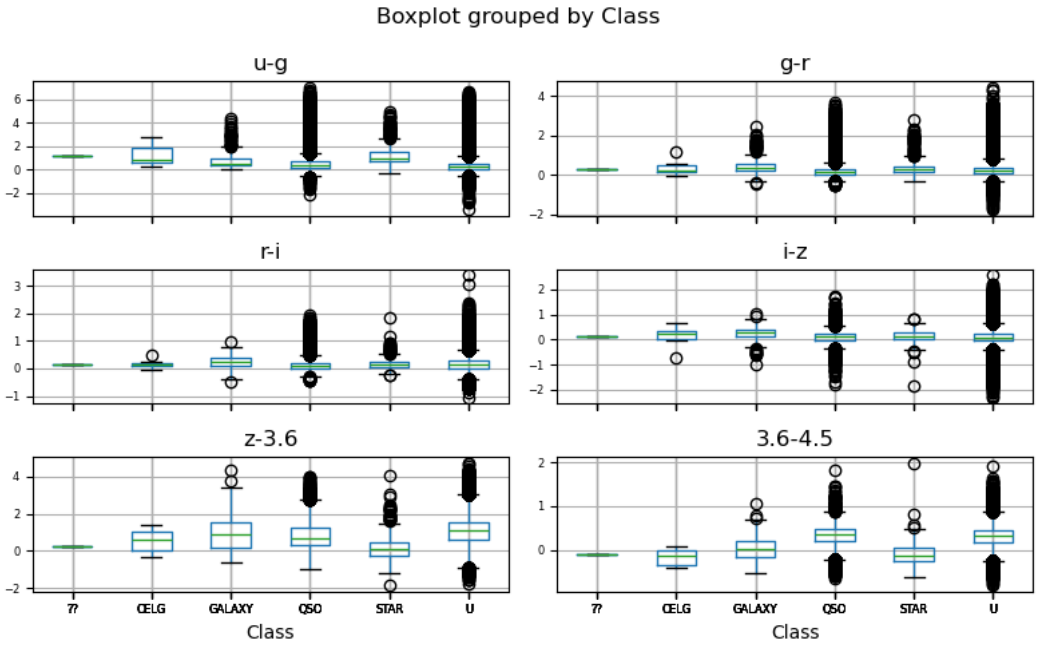

Figure 11: Box plot of the colors in the "robust" and "robust contaminants" subsamples grouped by the class of the objects.

3. H_0*: The mean* `zphO` *is equal to the mean* `zphIR` *for the objects having the* JHK *photometry performed.* We also ought to test the algorithm of [Richards et al., 2015] for calculating the photometric redshift utilizing the *ugrizJHK* magnitudes. Rejection of this hypothesis would mean that the algorithm works incorrectly when including/excluding the *JHK* colors (for example, when the colors vs. `zsp` relationship with/without *JHK* have too little variability to be efficient in designating the photometric redshift).

We choose the first hypothesis of the above three to be tested. The alternative hypothesis H_1 in this case is:

H_1*: Objects of all spectral classes do not have an equal mean of each color.*

In other words, there exists at least one spectral class that has a different mean of at least one color than other classes. The form of the alternative hypothesis indicates that it is a *composite hypothesis*, so we cannot construct the *region of acceptance* nor determine the *type II error* – thus we need to perform the *significance test* (according to Fisher's paradigm, which is currently considered to be a part of the Neyman-Pearson theory) where we deal with the *critical region* for rejecting the null hypothesis H_0 only and the region of acceptance of H_0 is not defined.

Figure 11 shows the boxplot of colors in the robust subsample (including the contaminants) grouped by the class of the objects. The equality of these distributions between classes will be tested for each color separately.

The natural and simplest choice for the test model for the above hypotheses is the *one-way analysis of variance (ANOVA)* which utilizes the *Fisher-Snedecor F distribution* as the test statistic for determining whether the means in given groups of data are statistically equal. However, this model has the following assumptions which must be met by data:

1. Response variable residuals are normally distributed (or approximately normally distributed).

2. Variances of populations are equal.

3. Responses for a given group are independent and identically distributed random variables.

Starting from the end, the third assumption is believed to be met by the data. Next, the second assumption is unknown to be met, however, since we do not have information about variances of the populations. This is an argument against performing the ANOVA test.

Another counter-argument comes from examining the first assumption. The basic test statistic for checking the normality of the distribution of a given sample is the *chi-squared test statistic*, but a more responsive approach is provided by the *Kolmogorov-Smirnov test statistic*. Let us perform the latter with a significance level of 0.01. The test rejects the hypotheses that our samples obey the normal distribution for any color with the p-values equal to 0 for each of the colors (this right-tailed test statistic is greater than 0.0373 for each of the colors, while the critical value for the significance level of 0.01 is equal to 0.0023). Hence, the assumptions for the ANOVA test are not satisfied by the color variables in our sample.

The non-parametric alternative for ANOVA is the *Kruskal-Wallis rank test* which does not assume the distribution of data to be normal. It assigns ranks to the data points and uses the H *statistic*. For the number of samples k greater than 3 and the number of examples n_i in the ith sample greater than 5, where $i = 1, \ldots, k$, the H statistic has the *chi-squared distribution with* $k - 1$ *degrees of freedom* – under the null hypothesis that all samples come from the same population. This test is right-tailed.

We used the `scipy.stats.kruskal()` function to perform the Kruskal-Wallis significance test; we choose the significance level to be 0.05. For the test, we chose four samples of objects: robust unclassified, robust quasars, robust galaxies (contaminants) and robust stars (contaminants). The samples of robust CELGs and objects hard to interpret were excluded due to too small number of examples, also when relative sizes of samples are taken into account.

The Kruskal-Wallis test statistic gave results greater than 154.8 for each of the colors. The critical value of the chi-squared statistic with 3 degrees of freedom for the significance level of 0.05 is equal to 7.815 which is much less than each of our results. Equivalently, all the p-values are less than $2.5 \cdot 10^{-33}$ – much less than the significance level. This indicates that we should reject the null hypothesis that the colors of objects of different classes have the same distribution. Indeed, the result suggests the samples may come from different populations, so there should be a way to classify hidden contaminants within the U class in the color-color plots.

6 Summary

The exploratory data analysis was performed for the catalog of quasar candidates published by [Richards et al., 2015]. The initial data processing started with data wrangling, followed by data cleaning for removing duplicates and outliers, and then finished with feature engineering creating new features essential for further analysis, for example, colors (differences of magnitudes) or the monomials of redshift for regression.

The total number of objects from the catalog we got after the cleaning is 885449 among which 733669 are spectrally unclassified (the U class) objects, 150443 are spectrally classified quasars (the QSO class) and 1337 are spectrally classified objects other than quasars (contaminants). The U class can be the aim of the future machine learning investigation on quasars so the main correlations between the variables were examined in this class, along with the QSO class. The main correlations between data points for all the examples in the cleaned catalog are shown in Figure 5. Among others, examples of color-redshift and color-color correlations are shown.

We focused on the robust subset of the data set, as defined by [Richards et al., 2015]. The total number of robust objects is 516169, with 401500 U-class and 113957 QSO-class objects.

Figure 7 shows the nonlinear correlations between colors and the photometric redshift `zph0`, while in Figure 8 color-color plots are shown and in Figure 10 the linear correlations between *UKIDSS/VHS* magnitudes and *SDSS/Spitzer* magnitudes can be seen.

Three statistical hypotheses about the data were formulated. One of them, claiming that objects of all spectral classes have an equal mean of each color was tested with the Kruskal-Wallis significance test and resulted in the rejection of this hypothesis.

6.1 Next steps

The next steps in analyzing this data would include fitting a polynomial function into the color-redshift relationship using regression algorithms to obtain a prediction on redshift for an object with photometry (i.e., colors are known) and without spectrometry performed (i.e., no direct measurement of redshift is available). The colors for which the color-redshift relationship is most variable may be chosen to create a multidimensional training space.

Furthermore, classification algorithms may be used to differentiate objects with respect to given features in the color-color plots. One of the examples of these features can be redshift, as depicted in Figure 8. Additionally, by utilizing the color-redshift relationship, one can derive a redshift progression line in multidimensional color-color spaces. Moreover, we showed by hypothesis testing that there is likely a separation of distribution of colors between different spectroscopical types of objects which yields a possibility to classify in color-color spaces whether an object is a quasar or not.

Finally, the regression of the linear relationships between magnitudes from Figure 10 can help to estimate lacking magnitudes based on known magnitudes at different passbands.

6.2 Data quality

The data set used contains a large number of examples which are astronomical objects acquired by various telescopes. Some of the most important measurements have their errors given (e.g., magnitude) which is helpful in the estimation of the true value of the measured variable. The catalog is complete in the sense that no essential value is missing.

One of the disadvantages of this data set is the irregularity of the distribution of spectroscopic redshift, `zsp`, caused by merging various surveys, each aiming at different ranges of redshift (see Figures 1 and 6 for all the cleaned data and the robust subset of it, respectively). This may be one of the reasons for the observed skewness in brightness distributions. Another reason for this may be the Galactic extinction. Although, the authors performed the correction for the extinction, the knowledge of this phenomenon might likely have improved since the date of publication, thus it would be optimal to use this quasar catalog along with the current Galactic extinction data for improved extinction reduction. Finally, it seems that there is a need for data with a more robust estimation of photometric redshift which agrees better with the spectroscopic redshift (see the top left plot in Figure 5), does not tend to be quantized (discretized) into certain values only (as can be seen, for instance, in any plot with `zph0`) and has less uncertainty.

References

[Richards et al., 2015] Richards, G., Myers, A., Peters, C., Krawczyk, C., Chase, G., Ross, N., Fan, X., Jiang, L., Lacy, M., McGreer, I., Trump, J., and Riegel, R. (2015). Bayesian high-redshift quasar classification from optical and mid-IR photometry. *The Astrophysical Journal Supplement Series*, 219.

Table 1: Features in optical+MIR photometric quasar catalog in cand.dat. This table is a copy of Table 2 in [Richards et al., 2015].

Column	Name	**Short**	Description
1	R.A.	**RAdeg**	Right ascension (J2000)
2	Decl.	**DEdeg**	Declination (J2000)
3	CLASS	**Class**	Spectral classification (QSO, GALAXY, STAR, CELG, ??, or U)
4	ZSPEC	**zsp**	Spectroscopic redshift (if known)
5	U_MAG	**umag**	SDSS u-band AB magnitude, corrected for Galactic extinction
6	G_MAG	**gmag**	SDSS g-band AB magnitude, corrected for Galactic extinction
7	R_MAG	**rmag**	SDSS r-band AB magnitude, corrected for Galactic extinction
8	I_MAG	**imag**	SDSS i-band AB magnitude, corrected for Galactic extinction
9	Z_MAG	**zmag**	SDSS z-band AB magnitude, corrected for Galactic extinction
10	CH1_MAG	**3.6mag**	3.6 μm AB magnitude, corrected for Galactic extinction
11	CH2_MAG	**4.5mag**	4.5 μm AB magnitude, corrected for Galactic extinction
12	U_MAG_ERR	**e_umag**	Error on u-band magnitude
13	G_MAG_ERR	**e_gmag**	Error on g-band magnitude
14	R_MAG_ERR	**e_rmag**	Error on r-band magnitude
15	I_MAG_ERR	**e_imag**	Error on i-band magnitude
16	Z_MAG_ERR	**e_zmag**	Error on z-band magnitude
17	CH1_MAG_ERR	**e_3.6mag**	Error on 3.6 μm magnitude
18	CH2_MAG_ERR	**e_4.5mag**	Error on 4.5 μm magnitude
19	U_FLUX	**Fu**	SDSS u-band flux density in nanomaggies
20	G_FLUX	**Fg**	SDSS g-band flux density in nanomaggies
21	R_FLUX	**Fr**	SDSS r-band flux density in nanomaggies
22	I_FLUX	**Fi**	SDSS i-band flux density in nanomaggies
23	Z_FLUX	**Fz**	SDSS z-band flux density in nanomaggies
24	CH1_FLUX	**F3.6**	3.6 μm flux density in microJy
25	CH2_FLUX	**F4.5**	4.5 μm flux density in microJy
26	U_FLUX_ERR	**e_Fu**	Error in u-band flux density
27	G_FLUX_ERR	**e_Fg**	Error in g-band flux density
28	R_FLUX_ERR	**e_Fr**	Error in r-band flux density
29	I_FLUX_ERR	**e_Fi**	Error in i-band flux density
30	Z_FLUX_ERR	**e_Fz**	Error in z-band flux density
31	CH1_FLUX_ERR	**e_F3.6**	Error in 3.6 μm flux density
32	CH2_FLUX_ERR	**e_F4.5**	Error in 4.5 μm flux density
33	YAPERMAG3	**Ymag**	Y-band Vega magnitude from UKIDSS or VHS
34	JAPERMAG3	**Jmag**	J-band Vega magnitude from UKIDSS or VHS
35	HAPERMAG3	**Hmag**	H-band Vega magnitude from UKIDSS or VHS
36	KSAPERMAG3	**Kmag**	K-band Vega magnitude from UKIDSS or VHS
37	YAPERMAG3ERR	**e_Ymag**	Error in Y-band magnitude
38	JAPERMAG3ERR	**e_Jmag**	Error in J-band magnitude
39	HAPERMAG3ERR	**e_Hmag**	Error in H-band magnitude
40	KSAPERMAG3ERR	**e_Kmag**	Error in K-band magnitude
41	FUV_MAG	**FUV**	GALEX FUV magnitude (AB)
42	FUV_MAG_ERR	**NUV**	GALEX NUV magnitude (AB)
43	NUV_MAG	**e_FUV**	Error in FUV magnitude
44	NUV_MAG_ERR	**e_NUV**	Error in NUV magnitude
45	GI_SIGMA	**gisig**	Indicator of distance from mean $g-i$ color at ZPHOTBEST
46	EXTINCTU	**Au**	Extinction in SDSS u band
47	STAR_DENS	**den***	Star density from KDE algorithm
48	QSO_DENS	**denq**	Quasar density from KDE algorithm
49	ZPHOTMIN	**b_zphO**	Minimum photometric redshift (ugriz)
50	ZPHOTBEST	**zphO**	Best photometric redshift (ugriz)
51	ZPHOTMAX	**B_zphO**	Maximum photometric redshift (ugriz)
52	ZPHOTPROB	**zphOP**	Probability of ZPHOTBEST being between min and max
53	ZPHOTMINJHK	**b_zphIR**	Minimum photometric redshift (ugrizJHK)
54	ZPHOTBESTJHK	**zphIR**	Best photometric redshift (ugrizJHK)
55	ZPHOTMAXJHK	**B_zphIR**	Maximum photometric redshift (ugrizJHK)
56	ZPHOTPROBJHK	**zphIRP**	Probability of ZPHOTBESTJHK being between min and max
57	LEGACY	**Leg**	Indicates if object is in the SDSS Legacy footprint
58	SDSS_UNIFORM	**Uni**	Indicates if object was selected according to Richards et al. (2002)
59	PRIMTARGET	**Prim**	SDSS primary target selection flag; see Richards et al. (2002)
60	PM	**pm**	Proper motion in milliarcseconds per year
61	DUPBIT	**DB**	Bitwise flag indicating low-z (2^0), mid-z (2^1), and high-z (2^2) sources