The Search for a Standard Candle in the Sloan Digital Sky Survey DR7 Quasar Database

Christopher R. Wahl*
PHS 404 - Senior Project Seminar II
Department of Physics, SUNY Brockport

Dr. Eric Monier
Research Advisor
Department of Physics, SUNY Brockport
(Dated: April 26, 2017)

Quasars are among the brightest objects in the universe. Their intense luminosities arise from an accretion disk surrounding a supermassive black hole. This project seeks to identify a class of quasars with observed magnitudes depending only on distance. The Sloan Digital Sky Survey Quasar database was queried for quasars over the redshift range of [0.46, 0.82], g $\leq$ 19. Spectral flux densities were re-sampled onto regular wavelength intervals and shifted into rest frame, separated by 1 Å. A scaling and χ^2 matching system was used to compare individual spectra to the remaining memebers of the catalog. Matching was conducted independently over Mg II, Hβ emission lines and the central continuum and results were compared current Flat ΛCDM using standard values. Preliminary results are reported for guidance.

I. INTRODUCTION

A. Quasars

Quasi-Stellar Objects (QSO), or Quasars, are a form of active galactic nuclei. Powered by in-falling matter to a central black hole, they are some of the most luminous objects in the universe. Estimated black hole masses are often found 10^7~$10^9 M_\odot$ and event horizons can be discussed in terms of the orbital radius of Uranus.[1]

First discovered in the 1950s as loud radio sources, the low resolution of the measurements prevented isolation of sources. Eventually, Cyril Hazard was able to identify the optical source of Quasar 3C 273[2] (Figure 1) through lunar occultations[1]. Shortly afterward, Maarten Schmidt recognized hydrogen emission lines of the spectrum at a redshift of $z = 0.158$[2]. Using Hubble's 1929 discovery of the relation between distance and recessional velocity, this cosmological redshift, corresponding to ~2 billion light-years distant, firmly established quasars as extragalactic objects.

It is apparent that while even at these distances, quasar luminosities are intense enough to have appeared star-like in initial observations. The matter accretion discs around the black hole can grow to diameters measured in light-days, and powerful relativistic jets expelling material are often found from the central black hole. In some cases, these jets are described in terms of millions of light-years.

Quasars have been discovered in substantial number across the universe, from $z \approx 0.1$ to redshifts exceeding $z = 6$. The light from these objects was emitted billions of years ago, and thus its observation can be used as a window to the early universe.

B. Redshift as a Measure of Distance

In 1929, Edwin Hubble confirmed the relation proposed by Georges Lemaître prior; The recessional velocity of a galaxy and its proper distance[3] are related

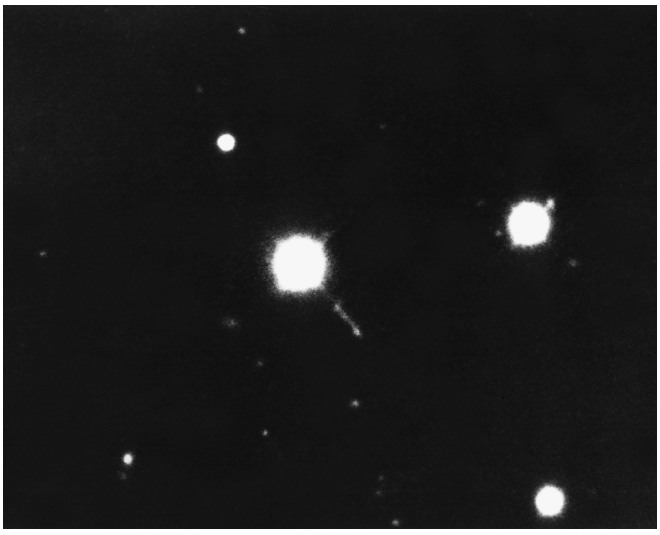

FIG. 1: The first identified quasar: 3C 273 (NOAO/AURA/NSF)

* cwahl2@brockport.edu

[1] The Schwarzschild Radius for a black hole of $10^9 M_\odot \approx 20$ AU where $M_\odot$ indicates the units are in terms of solar masses.

[2] The 273rd object in the Third Cambridge Catalog of Radio Sources

[3] Proper distance is the cosmological separation between objects and so can change with time as a result of the expansion of

by a constant[3]. Working from redshifts measured by Slipher and Humason, along with his own distance measurements, Hubble noticed that there existed a roughly linear relationship between a galaxy's distance and its redshift. That is, the further away an object is the faster it is receding.

$$v(D) = H_0 D \text{ where } \begin{cases} v & \text{: Recessional Velocity} \\ H_0 & \text{: Hubble Constant} \\ D & \text{: Proper distance} \end{cases}$$

The Hubble Constant, originally reported as 500 km s^{-1} Mpc^{-1} by Hubble (though it is not asserted that he himself is the one who titled the constant), as reported measurements increased, its value rapidly converged its current value near 70 km s^{-1} Mpc^{-1}. The Hubble constant, like the fine-structure constant α (and others), may not in fact be constant, but may vary over cosmological time. That is a topic for many other papers.

Redshift is the result of the Doppler effect - which describes how the light emitted from a receding object is elongated as a function of its velocity. In visible light, this represents a shift toward the red end of the spectrum giving rise to the term "redshift.". While this value is affected by both the cosmic expansion as well as any local velocity effects, an object sufficiently far away will have its motion dominated by the inflation.[4]

At these scales, the finite speed of light must be taken into account. Light observed today from an object 1 light-year away, was emitted 1 year ago. Looking at cosmological objects is synonymous with looking back in time (and is appropriately referred to as the "lookback time"). At increasing distances, this relationship does not hold, but in fact scales substantially as a result of the universe continuing to expand while the light traverses it. For example, light emitted 4 billion years ago from an object 4 billion light years away will have actually covered over a distance of over 6 billion light-years before being observed today.

For the ease of discussion, both distances and velocities (being linked) are often described by the redshift term z where

$$\lambda_{observed} = (1 + z)\lambda_{emitted} \tag{1}$$

Notice that emitted wavelengths are elongated with increasing z, indicated that a redshift of $z = 0$ corresponds to *now*. That is, cosmological redshift is ever defined with the present corresponding to $z = 0$ and increases as a function of distance $\propto$ lookback time.

C. Standard Candles

Standard candles are useful for determining scale and distance measures across the universe. These objects have known luminosities and exhibit some unique identifying characteristic. The observed intensity of an object can then be used to determine its distance. Clearly, these are extremely useful for measuring and modeling in cosmology.

A couple of examples include:

- Cepheid Variables: A class of stars which pulsate radially.[5] Their luminosities are related to (and can thus be determined by observation of) their pulsation period. Cepheids were used by Hubble in 1924 to establish conclusively that the Andromeda & Triangulum "spiral nebulae" were in fact separate galaxies, rather than objects within the Milky Way[4]. These can be used as standard candles up to 35 - 50 million light-years, which makes them effective for determining distance in the Local Group.

- Type ia Supernovae: A specific type of supernova which occurs in a binary system containing a white dwarf. The dwarfs accumulate material as a result of matter accretion from their binary companion. Should this star grow beyond 1.44 $M_\odot$- known as the Chandrasekhar limit[6] - it will re-ignite and often trigger a supernova. These explosions produce consistent peak luminosities as a result of the uniform mass at which re-ignition occurs. Type ia supernovas have been used for distance measurements on the order of 300 million light-years which puts them firmly on the super-cluster range. Unfortunately, these events are rare and their light curves peak for only a matter of hours or days at the beginning of the event. If observations are not made over the supernova's peak intensity (big telescopes need to be swung into place, calibrations must be made, and they may simply be on the wrong side of the planet), a reliable distance measure will not likely be determined.[7] Astronomers

the Universe. Compare this to the *comoving distance*, which is factors out universal expansion (although other factors can come into play, such as gravitational attraction).

[4] It's worthwhile to understand that this recessional velocity is not caused by the galaxy's motion; Galaxies are at this scale more or less not moving. The redshifting of their emitted light is actually caused by the expansion of space *between* the objects. Even if the emitter and the observer were perfectly stationary, the light would *still* be redshifted from the cosmological inflation occurring between them. Simply put, the expansion of space is pushing objects apart, even while the objects themselves are stationary.

[5] There are actually two main populations of Cepheid stars - Type i and Type ii, appropriately. They both exhibit the same pulsation-luminosity connection, though the exact relationships differ slightly.

[6] Note that this is **not** the same as the Chandrasekhar *Mass*.

[7] Sometimes humans get lucky and a person like the Minister Robert Evans takes up astronomy as a hobby. He holds the current record of discovering 42 different supernova[5] - including an entirely new type: The Type ib supernova. He features prominently in Bill Bryson's A Short History of Nearly Everything.

have been searching for progenitors of these supernova for more than a century[6].

This project seeks to establish the foundation for finding a reliable set of standard candles at a substantially greater range[8] in the interest of defining a cosmological distance measure independent of redshift alone. The results of any detection of this effect may be applied to more accurately probe currently establish measures and allow deeper refinement of the Cosmological Standard Model.

D. The Sloan Digital Sky Survey (SDSS)

1. Background

The Sloan Digital Sky Survey began calibration observations in 1998 from the purpose built 2.5 m Sloan Telescope, moving to data collection in 2000. Undergoing several phases, the SDSS has imaged more than 5 million objects, including 800,000 galaxies and over 100,000 quasars, covering 35% of the sky.

Located at the Apache Point Observatory in Sunspot, NM, the telescope captures images using a system of five filters - *u, g, r, i, & z*[9] - along with the full spectrum. The camera itself consists of thirty chips, each at a resolution of 2048x2048 pixels, totaling approximately 120 megapixels.[10] It is cooled by liquid nitrogen to 190 K (-80°C) and produces about 200 GB of data each night[7]. As the telescope is ground based, it is only capable of observing light which passes through the Earth's atmosphere. In practice, this corresponds to wavelengths within 3800 Å to 9200 Å - near ultraviolet to infrared.

Each observation makes use of a spectrographic cartridge. These cartridges, or plates, are aluminum discs which are prepared specifically for each portion of sky being observed. Each disc covers 3° of sky and is manually drilled with 640 holes[7]. These holes are individually connected by optical fiber to the camera and correspond to a star, galaxy, or other object of interest. They may be reused for multiple observations, or the same portion of sky may be observed through a different plate, to record different objects in the same field.

The Survey has undergone four distinct phases of observation. Phase III completed most recently in 2014 with Phase IV planned to run until 2020. Each including a number of experiments, early phases set out in search of type Ia supernovae and along with observations investigating the Milky Way's galactic evolution. Recent experiments include attempts map nearby galaxies using spatially resolved spectroscopy, and to make precision cosmological measurements of the early universe. SDSS data are organized and made available in numbered releases with the most recent release of DR13. The most relevant information to this project, inclusive through DR7, was a part of the Sloan Legacy Survey in Phases I & II and completed in 2008. Archived data are cataloged into the Sky Server - a free publicly accessible SQL-powered database and web interface.[11]

With DR7 having been available for nearly a decade, the data within it have been well studied. Additional information, including central black hole mass estimates and improved redshifts, have since been published. Where the initial data files are not modified once stored in the archive server, any corrections must be accessed from a secondary source. A substantial number of properties and refinements have been cataloged into a single location[8], though it should be noted that these values have not been updated since 2011.

2. File Format

The Sloan system is highly automated. Telescope data are processed through a computational pipeline before being stored in the archive server. Each quasar is compared to a template spectrum originally developed by Vanden Berk, et al[9]. Specific features, such as oxygen III emission lines, are identified and their offset from the rest frame template is used to determine redshift. The data, filter magnitudes, observational comments and additional information from the pipeline analysis are packed into Flexible Image Transport (FIT) files. Queries to the Sky Server,[12] either by browsing or via the SQL interface, can be used to select quasars with desired characteristics and a simple `wget` command employed to mass-download the results.[13]

---

[8] Closer to the 10,000 million light-years scale.

[9] *u*: Ultraviolet (3542 Å), *g*: Green (4770 Å), *r*: Red (6231 Å), *i*: Near-Infrared (7625 Å), *z*: Infrared (9134 Å)

[10] In the year 2000, an $800 point-and-shoot digital camera had a resolution of around 3 MP.

[11] The most recent releases and information are available at `http://www.sdss.org`, though it can be difficult to discern which sets can be found where. If it is not abundantly clear, Phase III releases (DR8 through DR12) are available at `http://www.sdss3.org`. Phases I & II are part of the archived database - which is no longer being updated - and the webpages are stored as-is at `http://classic.sdss.org`. At the time of this writing only DR13 from Phase IV has been made available. Somewhat confusingly, it is also listed at `sdss3.org`, despite DR12 being the final Phase III release. Moreover, the DR11 & 12 links will also go to `sdss.org`, while DR8, 9 & 10 remain at `sdss3.org`. They also link off to releases DR1 through 7, though those all go, appropriately, to the archive server at `classic.sdss.org`. This kind of interaction is somewhat characteristic of the SDSS. A well maintained, standardized SDSS wiki would be a phenomenal resource and a fantastic project for a group of undergraduate students.

[12] The Phase I & II Sky Server interface is located at the Catalog Archive Server address: `http://cas.sdss.org/dr7/en/tools/search/`

[13] An *extremely* useful url for this is `http://das.sdss.org/www/html/das2.html`. Specifically, a list of spectrographic fibers was generated from an Sky Server table join and this site was used

The FIT format is intentionally flexible and its uses extend far beyond what will be addressed in this project. However, while the *method* of reading and writing data is standardized, the data itself being stored - and its format - is left unto the group providing it. Where the SDSS hosts vast amounts of data - free and available at anytime to any person interested - it is in their interest to minimize the size of that data in both storage and delivery bandwidth. Specifically in the case of this project, the consequences of re-expanding the data into usable values adds a layer of complexity, which will be discussed further on in Section IV B.

This project makes use of spectrographic observations of quasars. The data of interest are the measures of intensity of quasar light (flux density) corresponding to the wavelength of observation. An example spectrum is given in Figure 2. The general format of a one-dimensional spectrum file name, as stored by the SDSS, is

`spSpec-`*mmmmm-pppp-fff*`.fit`

where $\begin{cases} mmmmm : \text{Modified Julian Date} \\ pppp : \text{Plate number used for observation} \\ fff : \text{Fiber number observed from} \end{cases}$

This format of MJD[14]-PLATE-FIBER is preserved for use as a unique namestring identification for each spectrum used in this project[11]. The Survey calibrates each

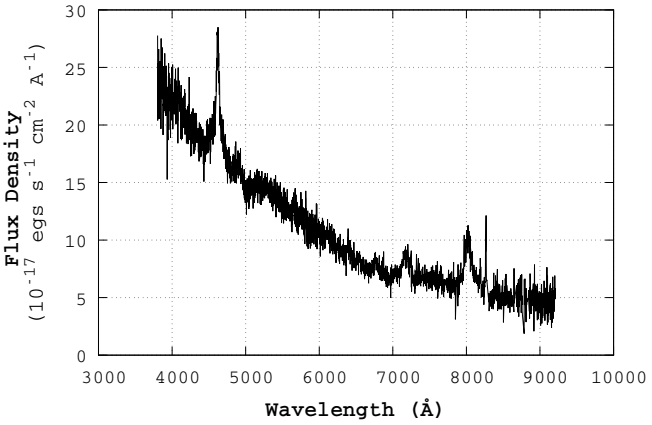

FIG. 2: The observed frame spectrum of QSO 53385-1944-476

spectrum and resamples the flux densities into wavelength bins described the dispersion formula given in Equation 2.

FIT files include a header layer and number of data layers. The data layers used here are "aligned," which is to say that the information at the n^{th} pixel in layer 1 corresponds to information at the same point in layer 2, 3, etc. In regards to the one-dimensional QSO spectra, the specific data layers are

1. `Flux Density`: The data y-axis data displayed in Figure 2.
2. `"Continuum-Subtracted" Flux Density`: A "flattened" form of the first later. This layer was not used in this project.
3. `Flux Density Uncertainty`
4. `Error Bitmask`: A description of known error flags. The use of this is discussed in Section IV.

Notably absent, here, is a layer describing the wavelengths themselves - individual wavelength values are not stored in the 1-D spectrum. Rather, the data to generate them are stored in the file header.

Headers are the storage location of "singular" information, such as the observation date and conditions, object ID, position, redshift, units and magnitude. More applicable, the coefficients of the wavelength dispersion equation (2) below are stored in the header.

$$\lambda_i = 10^{c_0 + c_1 i} \text{ where } \begin{cases} \lambda_i : i\text{th wavelength} \\ c_0 : \lambda_0 \text{ coefficient} \\ c_1 : \text{Dispersion coefficient} \end{cases} \quad (2)$$

Specifically, three important values are for generating the wavelengths are extractable from the header: Both coefficients, stored as '`coeff0`' & '`coeff1`' given in the equation above, and the total number of pixels (i.e. total number of datapoints) in the spectrum - '`naxis`'. A user familiar with Python can picture the header as analogous to the `dict` type; A key of '`coeff0`' returns the corresponding floating point value. The data layers are - in Python - accessible simply as arrays of length `naxis`.

Regarding the wavelength bins, these are clearly **not** spaced by constant intervals, as evident by the exponential form of this dispersion equation. Thus, forming a rest frame spectrum binned by 1 Å spacing (to provide a common scheme from which to compare all spectra) is something of a challenge. Speculation as to the reasoning behind using this exponential dispersion is an exercise left unto the reader.

to generate the file list passed to `wget` to download the results. The SDSS sets a maximum of 10,000 requests per session, so any larger queries should be split into multiple commands.

[14] **MJD**: *Modified Julian Date*, or the number of days since midnight on November 17, 1858. It can be calculated directly by subtracting 2,400,000.5 days - the number of days since 1 Jan 4713 B.C.E. (beginning of the Julian calendar) - from the current Julian date. The $^1/_2$ day factor arises from the Julian Calendar beginning at noon rather than midnight. Further information as to *why* this date was chosen can be found in source [10], but it effectively amounts to ease of storing values in computer memory. MJD was introduced by the Smithsonian Astrophysical Observatory in 1957 to record the orbit of Sputnik. There is also a *Truncated* Julian Date, which is given TJD = MJD - 4000, resetting the epoch to midnight, 24 May 1968. In both cases, the interest is in simplifying chronological calculations.

A sample header, along with further explanation of the stored information can be found on the Sloan archive server[12]. It is important to be aware that error mask codes, header contents and numerous other factors, while consistent between Phases I & II (i.e. through DR7), may not be in later data releases and thus the information provided here may not be applicable.

II. RESOURCES

A. Hardware

Developmental hardware and lightweight processing was achieved using a consumer grade laptop. More computational intensive programs offloaded onto one or more a desktop computers. Most methods targeted a maximum RAM usage of 8 gigabytes, and in practice the longest processes lasted 36+ hours. Specific hardware listing can be found in Table IV at the end of this document.

B. Software

Project code was written almost exclusively in Python. Initially beginning in Python 2.7, the project was migrated to Python 3.6 64-bit. No effort was made to maintain backwards compatibility after the migration, primarily do to type guidance. In the current inception, Python 3.6 is an absolute minimum requirement resulting from the use of `f"{...}"` formatted strings. In addition to included base operators and built-in types, NumPy,[15] AstroPy,[16] and Gnuplot.py[17] libraries were also used. Package management was controlled via the Anaconda 3 distribution, which also included the Python interpreter.

Both Windows 10 and Linux Mint 18.1 - a variant of Ubuntu 16.04 LTS - were used as operating systems. Care was taken to control for the operating system on initialization in common, base methods. Thankfully, Python's `os` module, at the time of import, invokes appropriate file controls for parsing file paths for the operating system. The base project path was maintained constant relative to the library code, thus establishing its location could be reached with the same relative path regardless of whatever parent path the entire project used.

Nearly all pathing information is stored in the `common.constants` file. This is where the `os` library is imported and where all others will import the `os.path` methods such as `join` from.[18]

It is strongly recommended that a person intending to make use of this library operate it solely on an appropriate Linux and Anaconda 3 distribution[19] until such a time that they are well familiarized with its structure, design and cross-platform Python operation. Should a Windows installation be the only option - noting that Anaconda 3 is available for Windows - then the user should ensure that the entry file of any program running a multiprocessing-based method is guarded by:

```
if __name__ == "__main__":
        from multiprocessing
                        import freeze_support()
        freeze_support()
```

This guard is required under Windows as a result of how the operating system handles the `fork` process for multiple threads. It must be invoked in the entry file, before any multiprocessing methods are called. It is not required under Linux distributions. A wise programmer inserts it regardless if they suspect cross-platform operation at any time.

A custom `Spectrum` class was written to extend Python's built-in `dict` type. Data are stored by keys of wavelength corresponding to values in the form of a `tuple` of (flux density, error). All values are `float` type. Additional fields are included for a namestring identification, redshift value and filter magnitude in g, with appropriate getter and setter methods. A substantial number of methods for scaling, aligning, AB magnitude determination and more are also included in the class. The user should be familiar with this class as it effectively forms the foundation of the library. A number of methods may have fallen out of use as the project has evolved, but many others (`scale, align, getWavelengths` among many) are used in nearly every operation.

Both source code and the data library were synchronized across all hardware by an on-site Git remote.

---

[15] http://www.numpy.org/, Version 1.11.3

[16] http://www.astropy.org/, Version 1.3

[17] http://github.com/oblalex/gnuplot.py-py3k, Ported to Python3. Original package at http://gnuplot-py.sourceforge.net/

[18] Due to the way the `os` module is designed, methods such as `join` and `split` should not be directly imported (i.e. one should **not** call `from os.path import join`). Instead the entire package should be imported - that package import process is how Python sets up path handling for the operating system currently being used. Desired methods can be (and are) reassigned to shorter calls such as `join = os.path.join`.

[19] **Warning:** The Gnuplot.py package is not maintained by Anaconda for Python 3 and must be installed manually. It depends upon an existing Gnuplot installation. Under Windows, the library will likely need to be pointed to this installation path. Alternatively, the user may wish to learn Python's extremely feature-rich `matplotlib`. Gnuplot was used by the author based on familiarity and consistency when plotting manually, outside of Python code.

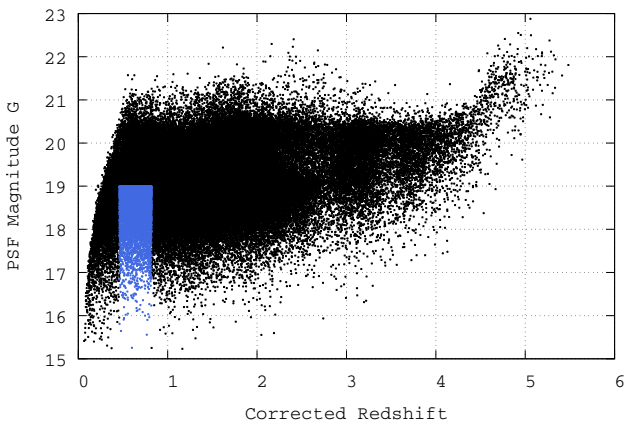

FIG. 3: *Black*: The distribution of all QSOs in DR7 of the SDSS. *Blue*: The catalog portion of interest.

III. REDUCTION OF THE SDSS CATALOG

The Sloan Digital Sky Survey, being ground based, limits the range of observed wavelength. That is, only the light which makes it through the atmosphere is viewable. Thus, it is important to select a portion of the catalog which:

- Contains a substantial number of QSO spectra
- Covers a wavelength range where identifying features may be exhibited
- Is bright enough to support a high signal-to-noise ratio

The redshift range of [0.46, 0.82] was selected to ensure both magnesium ((II)) and hydrogen Balmer-β were contained in all spectra, with a 200 Å buffer. At magnitudes dimmer than $g > 19$ [20] signal-to-noise quickly becomes problematic and so quasars beyond that limit were also excluded. Where available, any quasars flagged as containing broad absorption lines were also removed.

This first pass selection criteria reduces quasars of interest to below five thousand. A plot of the PSF magnitude[21] in filter g vs corrected redshifts of the QSOs in Survey is given in Figure 3, with the reduced catalog highlighted.

Over the course of data processing, a number of objects were removed from the catalog due to substantial errors, or missing portions of the spectrographic data.

[20] For persons unfamiliar, magnitudes are a logarithmic measure of an object's brightness and they are defined in such a way that *larger*, more positive values, are *dimmer* and smaller values, brighter. For example, the sun has an apparent magnitude of -27. Pluto's is 13.

[21] Note: This value - the point-spread-function magnitude - is drawn from the DR7 Properties catalog[8] and is *not* the same as the **fiber** magnitude stored in the file headers. Unless otherwise specified, any g filter magnitude is given as fiber, not PSF.

IV. DATA PREPARATION

A. Initial Conversion & Error Masking

The files corresponding to quasars of interest were cloned from the SDSS Sky Server. AstroPy's built in FIT reader methods were used to access the apertures and header information in each file. Each spectrum was given a unique identification string based on the header descriptors MJD-PLATE-FIBER - effectively the SDSS assigned filename, with `spSpec-` and `.fit` portions removed. Calibrated flux densities and corresponding errors were extracted from the first and second layers, respectively. Sloan coded bitmasks of:

SP_MASK_FULLREJECT

SP_MASK_BRIGHTSKY

SP_MASK_NODATA

were applied at each point, discarding any matches. The entire selection of SDSS DR7 error bits, codes and descriptions are listed in Table VII for reference.

A bitmask is simply a binary string with a specific position flipped. For example, the SP_MASK_FULLREJECT bit - number 18 - corresponds to a value of 2^{18}, or 262144. In binary, this is 1000000000000000000, or the 18th bit flipped.[22]

In this manner, a point in a spectrum's error mask can be labeled with multiple flags corresponding to multiple bitmasks. A masked value of 0...010100 has both the 2^{nd} and 4^{th} bits flipped, indicating that point has both BADFLAT and MANYBADCOL flags.

Using these as intended, specific errors can be quickly identified. Most languages have a bitwise AND operator (in Python, this is simply the & operator). The bitwise AND combines two bit arrays such that where they both have 1s, the resulting array has a 1. Otherwise, the result array has 0. I.E.:

```
Array 1:  11000101  - Spectrum bit errors
Array 2:  01001101  - Undesired flags
------------------
 Result:  01000101
```

Combining the corresponding error bits into a single value, this AND operator can filter a desired mask in one pass. If a point's corresponding mask doesn't contain the flags for the errors of interest, the resulting AND operation returns zero. If they do, some non-zero result is returned. In pseudocode this is simply,

```
if error_mask_at_WL & desired_mask != 0:
        skip this datatpoint
```

[22] Where the rightmost bit in this string is the *zeroth* bit which when flipped corresponds to $2^0 = 1$ (in binary, 000...001), this string is 19 characters long.

Generally, the mask codes used are found over less than a few Angstroms. Periodically, the discarded array extends over a large number of points, or the information is not even available in the original file (Figure 4). These are not readily detected without iterating through each spectrum and determining large gaps in wavelengths.

All spectra with contiguous discarded data of portions ≥ 100 Å were removed from the catalog. Over emission lines, this limit was reduced to ≥ 10 Å. This second pass reduced the number of interest to 4,079 objects.

These objects where stored in the `Spectrum` wrapper previously discussed. The dictionary portion was keyed by wavelength, storing a `tuple` of `( flux density, error )` corresponding to that wavelength key. Namestrings, redshift and g magnitude were stored separate from the `dict` portion of the class as class variables. In practice, the separation of namestring, etc, as non-keyed dictionary values allowed easy iteration through a spectrum's data without needing to explicitly exclude these identifiers each time.

Files were stored, initially, in ASCII format. A simple one-line header maintained the non-flux values (namestring, etc). Spectrographic data followed as a standard comma separated value layout. An example of this layout is given in Figure 15. Eventually, the `pickle` library was utilized and the serialized form was written/read directly in binary format. This form reduced loading speeds by as much as 70%. In conjunction with Python's `asyncio` asynchronous read/write libraries, loading the entire library from disk was reduced to less than a minute. Simple, linear looping through ASCII file reading (and thus substantial string-to-float conversions) could easily take several minutes, even from SSDs. This delay becomes significant when repeatably loading the library fresh from the disk.

Wrappers for both text and binary forms were written and organized into a `fileio` library.

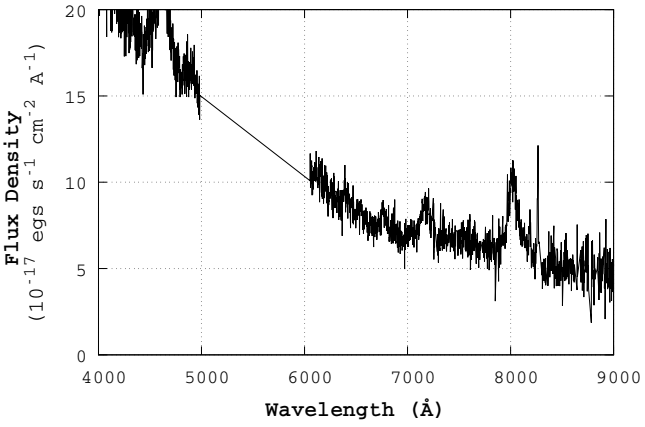

FIG. 4: Large arrays of missing or masked spectrographic data in a spectrum fails the required criteria for the catalog.

B. Resampling and Shifting to Rest Frame

To make comparisons between spectra, they must be shifted from their observed frame into their rest frame. Redshift values are encoded into the individual files as provided from the SDSS, though for a number of spectra, Hewett and Wild were able to provide improved values[13]. An additional requirement is the resampling of flux densities onto a common wavelength λ and wavelength interval $\Delta\lambda$. However, a consequence of the logarithmic wavelength dispersion (Equation 2) used by the SDSS is that a direct shift to rest frame, followed by a binning process prevents a proper statistical comparison.

To control for this nonlinear distribution, all spectra must be resampled in their observed frame onto a constant wavelength separation. Once the dispersion is linearized, the spectra can them be shifted into rest frame. A challenge to minimize resampling arises here. With the desire to have all spectra flux densities dispersed to integer wavelengths with common $\Delta\lambda_{rest} = 1$ Å spacing, it is valuable to determine what this spacing, $\Delta\lambda_{obs}$ will be in the observed frame. Additionally, it is important to determine what λ'_{obs} will return an integer value when shifted to λ'_{rest}. Clearly, these are a function of the individual spectrum's redshift, z.

Once these two values are established, flux densities can be resampled on to wavelength bins spaced by $(1+z)\Delta\lambda_{rest} = \Delta\lambda_{obs}$ beginning at λ'_{obs}. After this binning process is complete, the spectrum may be shifted to the rest frame. At this point, the spectrum will be in the rest frame with integer wavelength bins λ'_{rest} spaced 1 Å apart.

The spectrum objects are written to the project at each stage. Their file extension is used to differentiate between unmodified, binned, and rest-frame-and-binned objects as `.spec`, `.bspec` and `.rspec`, respectively. The `fileio.spec_load_write` package contains `rspecLoader` and `bspecLoader` methods which load the appropriate file based on MJD-Plate-Fiber ID alone. Unless intentionally interested in emitted-frame spectra, all operations are computed with binned-and-rest-frame, `.rspec` objects.

The `Spectrum` class does not differentiate between the three, as all are pickled forms of that class.

C. Band Magnitudes

An AB magnitude was generated for each spectrum directly from the spectrographic data. Unlike other magnitude systems which are generally defined relative some base, AB magnitudes are generated directly from measurements in absolute units. With the project's goal being to identify a set of quasars with brightness can be determined as a function solely of redshift, this system allows an independent method of establishing that brightnes. Equation 3 returns AB magnitude from units

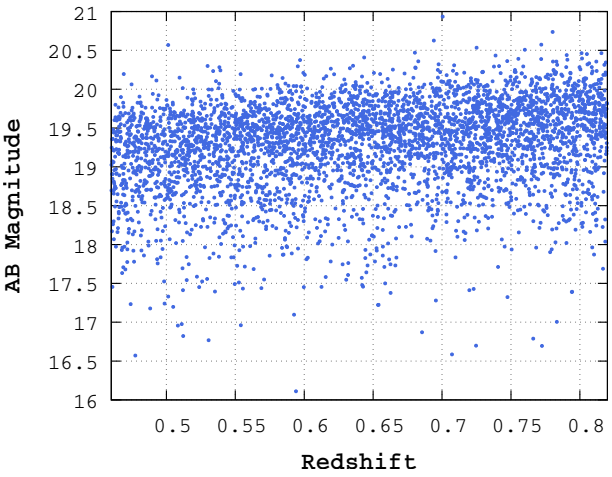

FIG. 5: Redshift and AB magnitude distribution of the reduced SDSS catalog.

of Janskys.

$$m_{AB} = -2.5 \log_{10}(f_\nu) + 8.9 \tag{3}$$

The flux densities stored in SDSS spectra can be converted from their source, [erg cm^{-2} s^{-1} Å^{-1}] to [Jy] by Equation 4.

$$f_\nu \text{ [Jy]} = 3.34 \times 10^4 \left(\frac{\lambda}{[\text{Å}]} \right)^2 \frac{f_\lambda}{\left[\text{erg cm}^{-2} \text{ s}^{-1}\text{Å}^{-1}\right]} \tag{4}$$

A reliably flat portion of the spectrum, centered on 4767 ± 18 Å, was determined visually. An average flux density (in Jy) was determined by using all flux densities within an 18 Å radius. Equation 3 was used to express this value in AB Magnitude terms.[23] Uncertainty was determined by the standard deviation of AB magnitudes over the same range.

The distribution of the reduced catalog is given in Figure 5. Unless otherwise specified, all magnitudes hereinafter referred to are AB magnitudes generated as described above.

D. Properties Catalog

The DR7 Properties Catalog tabulated by Yue Shen, et al [8], is contained entirely in a single FIT file. This is very useful for obtaining a copy of the data. The format necessarily stores each entry in an array - and each entry itself a sub-array of 130+ entries including indexed virial black hole mass estimates based on different emission lines (and in some cases, for multiple sources), equivalent widths of EM lines, observational data, and many other values (along with corresponding uncertainty).

This is somewhat cumbersome, however, for making individual queries, as it requires iterating through the primary array, joining MJD-Plate-Fiber strings from each sub-array to identify the object as one desired, then extracting the value/uncertainty of interest from the individual array indices as desired. With more than 100,000 objects, this is slow. Even keeping in only the quasars of interest, this would require knowing the indices of repeatedly used values such as redshift (and its uncertainty) while also loading large amounts of often irrelevant data and header.

Python's built in `dict` built-in data structure is particularly useful for this situation. The `dict` structure is not only inherently fast, it is (like all things pythonic) flexible. This local catalog of information regarding the reduced SDSS group is held in the code under the `catalog` package with the variable name `shenCat`. It is an example of what is described in this project as a "`namestring dict`." This design is simply a `dict` structure with keys of spectrum `namestring` strings, as described in Section IV A. The values are sub-dictionaries themselves, with corresponding keys and values as desired. A "literal" form example is shown at at the end of this document in Figure 14 in Python format. There are methods in the `fileio.list_dict_utils` which are designed to write and read these compound dictionaries to a CSV format.

Specifically, in the case of the `shenCat` variable, the following information is stored:

Key	Description
ab, ab_err	AB Magnitude and Error
bh_hb, bh_hb_err	Log BH mass estimate from Hβ line
bh_mgii, bh_mgii_err	Log BH mass estimate from Mg II line
ew_hb ew_hb_err	Equivalent width of the narrow Hβ line
ew_mgii ew_mgii_err	Equivalent width of the narrow Mg II line
lum_hb, lum_hb_err	Log line luminosity of Hβ line
lum_mgii, lum_mgii_err	Log line luminosity of Mg II line
gmag	Fiber magnitude in g filter
z, z_err	Hewett Wild corrected redshifts

Additional properties can be added as desired. The `shenCat` variable is actually a `dict` wrapped with a customized generic `catalog` class containing methods to re-

[23] SDSS QSO flux densities in DR7 are given as 10^{-17}erg cm^{-2} s^{-1}Å^{-1} with [λ] already defined in Å. In this case, the method in Eq. 3 can be simplified to: $f_\nu = 3.34 \times 10^{-13} \cdot \lambda^2 \cdot f_\lambda$, using the values directly from Sloan.

write/backup itself (serialized in `pickle` format and as a `JSON` file which can be directly imported as it stores the literal declaration of the catalog), after adding new properties.

The values for AB magnitude were generated directly from the spectrographic data, as described previously. Fiber g magnitudes are extracted from the FIT header of each file. With the generation of AB magnitudes, the necessity of the `gmag` key is somewhat outdated for use. The `*_err` keys access the uncertainty for their corresponding value - i.e., the redshift of an object is given by `z` $\pm$ `z_err`.

V. ANALYSIS

A. Matching System

A primary spectrum is selected to be compared against the entirety of the catalog. The primary method of discriminating results was a modified version of Pearson's χ^2 test, given in Equation 5.

$$\chi^2 = \sum_i \frac{\left[(f_{obs,i} \pm \epsilon_{obs,i}) - (f_{pri,i} \pm \epsilon_{obs,i})\right]^2}{f_{pri,i}} \tag{5}$$

where

f_{pri} : Flux density of primary spectrum

f_{obs} : Flux density of the input comparison

ϵ : Corresponding uncertainty

Matching process was conducted primarily over emission lines, with Mg II and Hβ being the heaviest weighted. OIII and Hγ are available as secondary and tertiary discriminators, though their efficacy is weak. The continuum, which at the given redshift range was effectively contained between Mg II and Hβ lines, was also available as a discriminatory option.

The χ^2 test results were most accurate conducted individually over the different identification areas. The ranges of the individual discriminants were small, in some cases only 50 Å. By operating the test over all, or multiple discriminants concurrently, substantial differences may be masked and discriminatory resolution greatly reduced.

To make proper comparisons, the two spectra must be scaled relative to each other. The scaling range used coincided with the AB magnitude determination range, and for the same reasons. The spectrum corresponding to f_{obs} was multiplied at all points by a scale factor determined by the ratio of the two AB-bands.

When performing mass comparisons against a primary spectrum, this process ensures all resultant χ^2 values resolved on the same scale. Proper comparisons between each resulting set requires that all primary spectra be on a common luminosity scale, as well. This standardized AB magnitude realization was established later in the project, so an efficient manner of mass-matching and scaling was not as thoroughly developed as it could have been.

B. Composite Generation

A composite spectrum is the result of "averaging" a number of spectra together. A process similar to the one undertaken in this method was used to investigate quasar spectra by Vanden Berk, et al[9] and the paper is valuable material. The same group also used this technique to generate the template spectra used by the Sloan pipeline to identify quasar redshifts.

A composing method was passed a list of input spectra, which were then scaled onto a mutual luminosity. The flux densities at any shared wavelengths were averaged, with uncertainties determined from the standard deviation of that group. Singular data points were passed along with corresponding uncertainty.

This differs from the process used by Vanden Berk, et al. Rather than averaging the points arithmetically, the group made use of the geometric mean.

$$\text{Geometric: } \bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 x_3 \cdots x_n} \tag{6}$$

$$\text{Arithmetic: } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

The paper argued this middle mean is better for preserving the slope-shaped features, though it was also reduced in sensitivity to emission lines. Some work was attempted with this method, though no significant difference was noticed, but composite spectra were substantially reduced in focus during later work. The geometric mean proved difficult to work with, stemming from inability to incorporate any value ≤ 0 - values not out of the ordinary in noisy, dim spectra.

The primary advantage of a composite is the substantially reduced noise the spectrum (see Figure 6). In theory, this reduced-uncertainty and reduced-noise spectrum should yield better fitting results, or at least act as a stronger discriminator.

Practicably, it is useful to consider the composite spectrum as a template. If the contributing spectra are observations of the same object - effectively the desired standard candle - then their resulting composite would serve as an increasingly accurate example spectrum. This can be used to inspect the catalog for matches which were missed due to the limitations of comparing imperfect observations. Conversely, they are important in discriminating against false-positives.

A standard candle template will eventually need to be developed as a composite. Where a single spectrum cov-

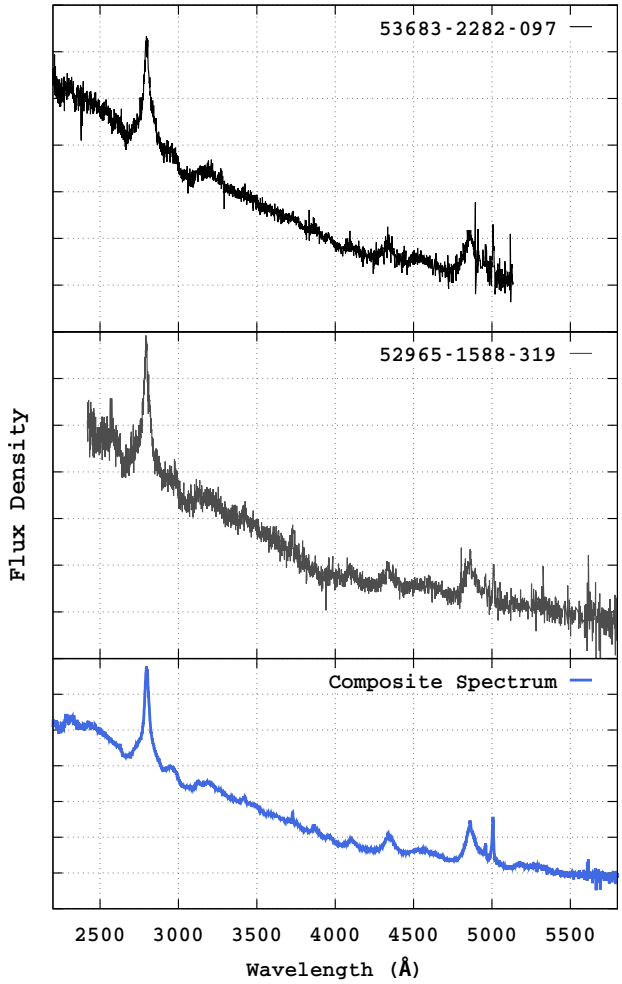

FIG. 6: **Top, Middle**: Two of many spectra which contribute to forming the composite in the bottom panel. **Bottom**: Note the increased wavelength range and reduced noise compared to the above individual spectra.

ers only a wavelength range bounded by the Survey's observational limits, a generated composite's wavelength range is comprised by the entirety of its constituent entries (see Figure 6). However, until the luminosity - redshift relation of this candle is quantified, the redshift of the composite will not be readily available for comparison.

C. Modeled Evolution

The AstroPy library includes a cosmology package which is capable of generating parameters from the Flat Lambda Cosmological Development Model (FΛCDM). The package model was initialized using the accepted values for the Hubble Constant $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_M = 0.3$ and $\Omega_\Lambda = 0.7$. With a known initial redshift and magnitude, an expected evolution of a QSO spectrum's magnitude solely as a function of redshift was generated.

Luminosity distance D_L is related to the comoving distance between two points in an expanding spacetime. It is a parameter which can be extracted from the FΛCDM using the redshift of a QSO relative the Earth. The absolute magnitude, M, of a QSO can be determined using D_L (in megaparsecs) from its apparent magnitude, m_0, at known redshift z_0 as given in Equation 8.

$$M = m_0 - 5\left[\log_{10} D_L(z_0) - 1\right] \tag{8}$$

With M established as a constant and the FΛCDM suppling D_L as a function of redshift, this equation can be rearranged to superficially "move" an object to a different redshift *(read: distance)* and "observe" its brightness.

$$m_e(z) = M + 5\left[\log_{10} D_L(z) - 1\right] \tag{9}$$

where

m_e : Evolved apparent magnitude at redshift z

M : Absolute magnitude, determined in Eq 8

$D_L(z)$: Luminosity distance (Mpc) as a function of redshift

This model-generated expected evolution is useful as a guide for directing the standard candle search and understanding results.

The use of "absolute magnitude" here is actually a misnomer, as the apparent magnitudes used by this method are not actually apparent magnitudes, but AB magnitudes. However, the "absolute" magnitude is being used only as a standard from which to determine evolved magnitudes. It is used, effectively, as a placeholder from which the changing luminosity distance can be defined and the input magnitude system is preserved.

VI. SEARCH PROCESSES

A. Catalog Evaluation

While it was expected that the 4,000+ QSOs remaining in the reduced catalog comprised a large enough distribution over the magnitude/redshift range of interest to yield a standard candle effect, it was necessary to substantiate that. By imposting a "perfect" effect in agreement with the FΛCDM, an evaluation of the number potential matches was made.

For each spectrum, its AB magnitude and uncertainty was evolved from $z = 0.46$ to $z = 0.82$. Any object with magnitude beyond $\pm 0.5\sigma$ this range was removed and number of remaining objects tallied. Numerous quasars yielded thousands of potential matches, though they generally resided in high redshifts where AB magnitude uncertainty is large. These counts reduced quickly, but still

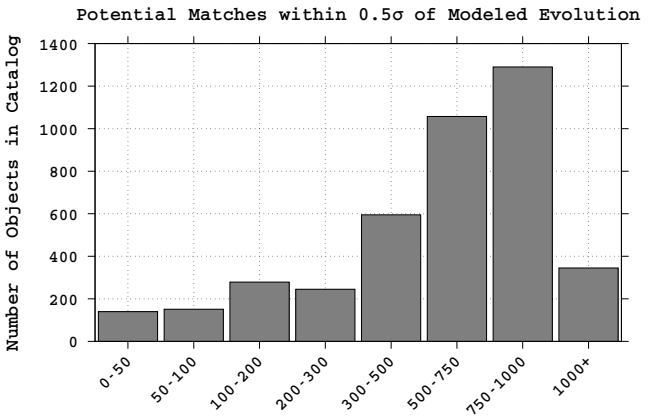

FIG. 7: Distribution of potential matches within 0.5σ of the modeled evolution process.

remained in several hundred with more than half the objects returning $500 - 1000$ potential matches. The distribution by redshift bin is given in Figure 7. In the appendix, Figure 16 shows the relationship of these potential counts to AB magnitude as well.

A strong effect could be imposed on the vast majority of spectra in the catalog. The reduced catalog's distribution is robust enough to expect a standard candle effect, should it exist, to be detectable.

B. Organize by Redshift Bin

Initial testing began by organizing the catalog by redshift bins spaces $\Delta z = 0.01$, with bin populations 100-120 in number. If the effect exists at a probability greater than 1-in-100, it was expected that most bins would contain at least one result.

A composite spectrum was generated from the members of the lowest redshift bin, $z = 0.46 - 0.47$. It was matched through the remaining bins. The best match AB magnitudes were plot against their corresponding redshifts. No pattern was evident.

The composite was used to find the most "average" quasar in the first bin, and match that successively though remaining bins. Again, the best matched spectra were selected and a magnitude v. z plot was generated. No pattern was directly evident.[24]

Working with the lowest redshift bin as the source, its entire population was matched through the remaining successive bins and the best matching result was kept. The matching process was conducted across the entire mutual spectrographic dataset of the objects.

The AB magnitudes of these results were plot as a function of redshift and compared against the modeled evolution. Overwhelmingly, the method was heavily favoring the brightest spectra in each bin. With the increased noise, then corresponding reduced discriminatory power, which would become characteristic of the higher redshift bins, the method was more apt to develop scatter plots than return an indication that matches were dimming with increased distance.

After experimentation and review of spectrographic comparison plots the masking effect of large-evaluation range discussed previously became evident (see: Section V A). Matching was furthermore limited to specific identifying features - specifically emission lines - and only supplemented with continuum as a final pass.

C. Full Catalog Search

Concurrently, the imposition of a specific redshift bin structure was lifted and matching results were restricted solely by the χ^2value returned. Using Equation 3, all spectra were scaled to an AB magnitude of 20 (more specifically, the corresponding flux density at the central AB range wavelength of $\lambda = 4767$ Å).

A generalized matching process was developed. Wavelength ranges were tailored to corresponding emission lines, accounting for Doppler broadening and a portion of the spectrum which "runs-up" to the line itself. The Mg II line was used as the first filter. Well matched results from that test were passed to be selected by Hβ. A final pass over the continuum was performed on the remaining spectra. The three regions being entirely independent of each other, the order of filter application with respect to results is inconsequential. For computational efficiency, the smallest wavelength ranges were applied first and results filtered before passing to the next range.

[24] During this investigation, composites were generated for each redshift bin. When their spectra were organized onto a single plot, a small absorber feature was noticed. Its wavelength position in these rest frame composites receded blueward with increasing z.
Returning the composites into the observed frame, this feature was centered at ~5550 Å, Broadening was noticed, though it is almost certainly attributed to the 'averaging' of redshifts in each composite.
A cursory search was made to identify any known atmospheric absorbers at this position, but none were found. Smaller, but still evident absorption features were noticeable when plotting the composites three-dimensionally in the following schema, and viewing from an overhead angle. {x: Wavelength, y: Redshift, z: Flux Density}.
It is desirable to refer to these small, unclassified, but still noticeable (assumed atmospheric) absorption lines somewhat colloquially as the "Monier Lines," in honor of the project research advisor (who is receptive to the idea, if not only in the interest of levity). They are evident in Figure 19 at the end of this report.

1. χ^2 Considerations

In all cases, the filter mechanism was simply $\chi^2 \leq$ `MAX_VALUE`. Where Pearson's χ^2 process is comparing direct values (rather than frequencies of occurrence in statistical bins, as originally designed), a difficulty emerges in comparing the results from one spectrum to another. The values used to produce the results discussed later are given in Table III Most importantly, the filter limits used here were done so *without* accounting for overlap; i.e. χ^2 calculation was conducted over the flux densities alone, without accounting for uncertainty strictly to find *any* kind of result.

This is accomplished without modification to the library code, as the `chi` method accepts an uncertainty multiplier, `n_sigma`. This parameter defaults to 1, but overlapping can be disabled by passing a value of 0. The inclusion of uncertainty rapidly increases the number matches and can yield some strange-looking sets without *very* tight maximum χ^2 values. Further refinement from these results will provide guidance for how to quantify overlapping flux densities inclusive of uncertainty.

Numerous maximum χ^2 values were used over the course of development. Those used for discussion of results in this paper emerged from substantial trial-and-error, guided by experience gained from the development process. It is speculated that these limits are substantially more restrictive - if not only based on the excluded uncertainty - than those that will be needed in the end. They are also imperfect.

The returned value itself is dependent both on the differences in spectra, but on their scaled flux densities themselves. Consider a pair of spectra, mutually scaled, which at a common wavelength, have respective flux densities of $f_p = 5$ and $f_o = 3$. Their χ^2 at that single point is

$$\chi^2 = \frac{(5-3)^2}{5} = 0.8$$

However, scaling these values by a factor ten and processing again returns $\chi^2 = 8$, owing to the spare multiple of 10 coming through the squared numerator. Clearly, hard-and-fast limits are undesirable without some specific quantification in relation to the intensities.

Developing some additional, non-χ^2 filter will allow restricted values to be loosened, passing a greater number of objects to the new filter, opening the potential for more matches.

This process was brute-forced - applied to every object in the catalog against every *other* object in the catalog. Generally, the Mg II line was effective as an initial filter, often cutting returned counts to less than 500, sometimes 10% of that, though on rare occasions over 1,000 QSOs passed. The second filter, Hβ, also proved to be a moderate discriminant, with the number of emergent results usually less than fifty. A final continuum pass served to either completely remove any remaining matches, or at least reduce them to single digits.

For any spectra which returned nonzero matching counts, the resulting matches were written to the disk before proceeding. A running total was appended as well, in the event of power loss, or other unplanned interruption (which were wont to occur at the most critical times), by appending each count to a CSV. The matching process alone estimates $\sim 10^{10}$ comparisons which even on a 12-logical core processor could take > 24 hours. In the event of a break - planned or otherwise - this running total could be read in and those spectra already processed, skipped. This also allowed inspection of match count distribution throughout the process.

2. Result Reduction

The final counts returned 1855 objects with more than one match, ~1000 with double digits. Of those, excessively high counts of 40+ (at least, excessive from experience) arose for nearly half.

Of these ~500, reduction was accomplished through a combination of direct observation and line fitting with SciPy's `curve_fit( fit_function, xdata, ydata, **kwargs )`. A useful feature of this method is that it allows the `fit_function` to handle any number of parameters. The results of those best-fit parameters are returned in a list corresponding to the method's input order. I.e. for a given method which returns only one `float`:

```
def fit_function( x, a, b, c, d ):
        ....
        return x

[ a, b, c, d ] = curve_fit( fit_function,
                            xdata, ydata )[ 0 ]
```

`curve_fit` actually returns two values: `popt & pcov`. The first, already described above, is the optimized parameters (in this case, the list `[ a, b, c ,d ]`, though this can be extended by simply defining the function to be fit as `f( x, *popt )` and passing the optimized parameters). The second, `pcov` is a 2d array giving the covariance of the `popt` parameters.

While the raw method can be used as described above, the reader is directed to the `anaylsis.slope_fit` package in the library which serves as a wrapper for this process. It also includes pre-made linear, quadratic and logarithmic function fitting methods, with generic wrappers for custom implementation. Simplified constants are returns as a `tuple`, and uncertainties are available with the flip of a boolean field.

From model guidance and the evolved magnitude equation 9, this expected to be logarithmic in nature. This is still somewhat speculative if not only on the basis of luminosity distance (the true logarithmic factor) not being directly proportional to redshift; Recall that D_L is dependent upon the model of the expanding universe.

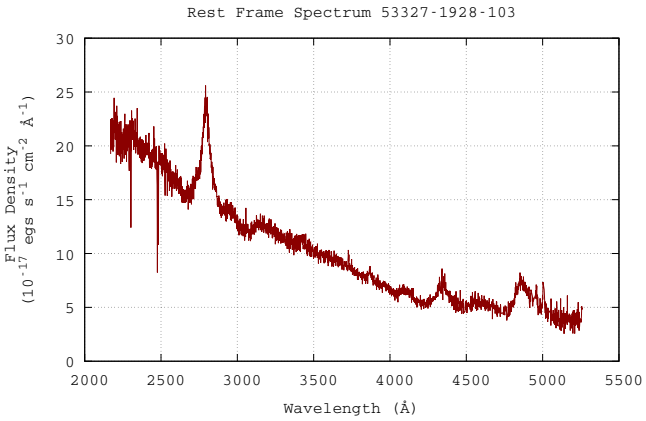

FIG. 8: Rest frame of primary Spectrum 53327-1928-103

However, for the limited catalog range the model suggests a range of ∼2.5 megaparsecs. In comparison, from the catalog range to a redshift of $z = 2$, D_L extends an additional 10 Mpc. At $z = 3$, D_L it is 20 Mpc from the catalog. At least for this initial search, inaccuracies in the cosmological model will most likely be well within the uncertainty of these results.

Each matching set's AB magnitude vs. Redshift data was plotted and fit to

$$\texttt{a}\ \log_{10}(\ \texttt{x}\ )\ +\ \texttt{b}$$

for constants a and b. Similarly, an R^2 value for each fit was determined. Visual inspection filtered through the sets with the highest R^2 to their optimized fit function, allowing for $\pm 2\sigma$ in the AB Magnitudes.

VII. RESULTS AND DISCUSSION

A candidate set was found in the results to primary spectrum 53327-1928-103. Noting its properties in Table I, it is a high redshift, dim spectrum. Its plot shown in Figure 8, it is notable that the Mg II appears to exhibit a larger than usual Doppler broadening along with reduced peak intensity relative the rest of the spectrum. Furthermore, the oxygen III pair (at 4959 Å and 5007 Å) is almost nonexistent and the continuum appears mostly devoid of any standout emission. All of these aspects, however, may be indicative of the high redshift-high uncertainty nature of the spectrum itself - particularly that the O III line is very near the spectrum data limit.

Where the high redshift catalog has characteristically high uncertainty, this often leads to greater populations in matched sets. However, what separates this 21-member set apart is the uncharacteristic alignment in their AB magnitude vs. redshift evolution.

A composite spectrum, generated from the matched set, is given in Figure 9. The O III line is clearly visible and the Mg II emission broadening relative its peak

Redshift:	0.750837 ± 0.000641
Fiber Magnitude (g):	18.4317
AB Magnitude:	19.0903 ± 0.0964
$\log_{10}(M_\odot)$ **BH** Hβ:	9.286 ± 0.040
$\log_{10}(M_\odot)$ **BH** Mg II:	9.278 ± 0.026

TABLE I: Properties of Spectrum 53327-1928-103

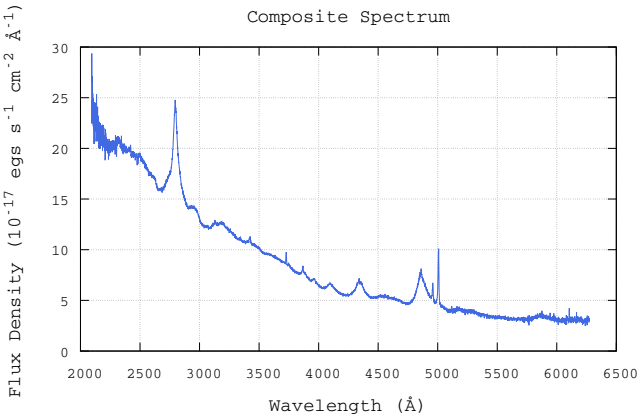

FIG. 9: Composite Spectrum from matched set to 53327-1928-103

is more indicative of usual results. Some slight emission peaks are visible along the continuum, though in those terms it still appears mostly featureless - which is *not* often observed. If this set does mostly describe a standard candle group, that does not necessarily indicate that *all* members given in this process are a correct. Similarly, there may be members missing. That is to say, the composite displayed may be incomplete, inclusive of erroneous matches, or both.

The Magnitude v. Redshift distribution is far from definitive, but still displays some promise. The grouping plot itself is given Figures 10 and 11, with data provided at the end of this report in Table V. While the primary spectrum is dim in comparison to the catalog as a whole, Figure 11 clearly shows it is near the brighter portion of the distribution for its redshift. This, in turn, leads i o*very* bright low-redshift matches along the expected range - which are clearly found. However, the low redshift catalog in this bright range is low in potential matches already.

This especially bright set is in a somewhat inconvenient position. The topmost panel of Figure 16 implies that at redshift ≈ 0.75, AB magnitudes of ∼19.5 will have the greatest potential number of matches - at least insofar as the FλCD model predicts. Following the shape of the

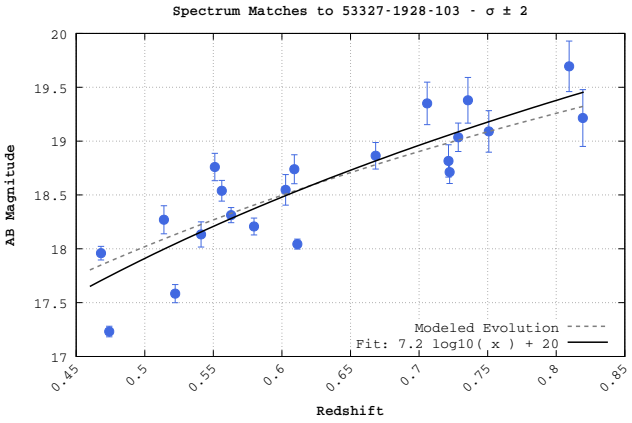

FIG. 10: Matches to Spectrum 53327-1928-103, at $\pm 2\sigma$

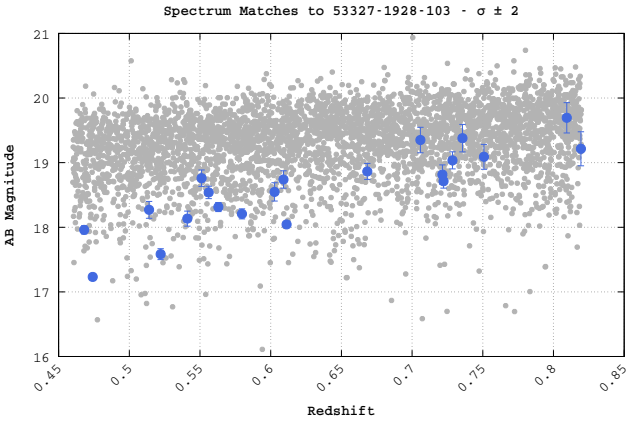

FIG. 11: Matches to Spectrum 53327-1928-103 as compared to the entire reduced catalog distribution.

redshift shaded arcs, these potential matches fall off very quickly (note the logarithmic y axis). At an initial AB magnitude of ~19, the potential matches are more than halved. While no selection was made to these matches regarding the modeled evolution, it is worthwhile to understand that a potentially greater number of effected objects may occur with a dimmer set.

Using supplemental data cataloged by Shen, et al discussed previously[8], the black hole mass estimates for all matches were extracted. The estimates considered were made from both Hβ and Mg II lines, each from the same respective source (the DR7 properties catalog includes multiple sources for Hβ line mass estimates). The $\log_{10} M_{\odot}$ central black hole mass estimates are given in Figures 12 and 13. If these object are all the same class, it should be reasonable to expect that they are of similar mass.

This similarity does, in fact, seem to manifest. In the case of both Hβ and Mg II estimates, all objects are within an order of magnitude of each other. When compared to the catalog distribution as whole, the matches have some alignment. Also as expected, the masses of these QSOs are in the larger portion of mass distribution, which falls in with their relative brightness.

Further match sets were also found through the Full Catalog search process. Though counts and R^2 values for the fit were not nearly as substantial, they are in need of further analysis. At this point, however, a better course of action may be to use the above described set for the purposes of refining results as greatly as practicable. If strong correlation can be discerned for a known set, the specifics would be extremely valuable to specifically quantify χ^2 limits as well as develop other non-χ^2 filters.

An additional group was found matching to Spectrum 54259-1832-281, with members and data listed in Table VI. The number of matches to the primary spectrum is lower - half that of the previously discussed set, though the fit is notably tighter - as evident in Figures 17 and 18 at the end of this document. Further analysis is left

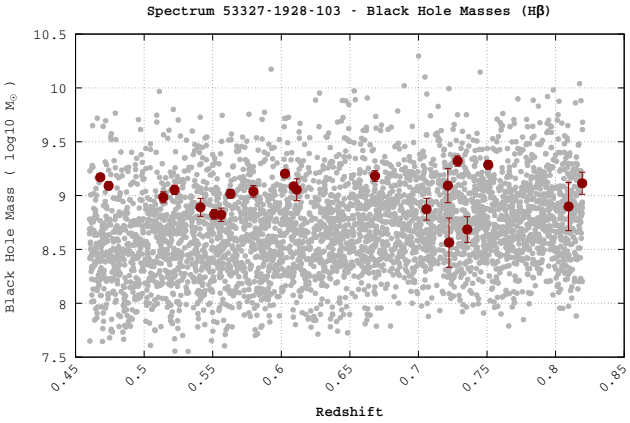

FIG. 12: Black hole mass estimates based off Hβ for Spectrum 53327-1928-103 matches

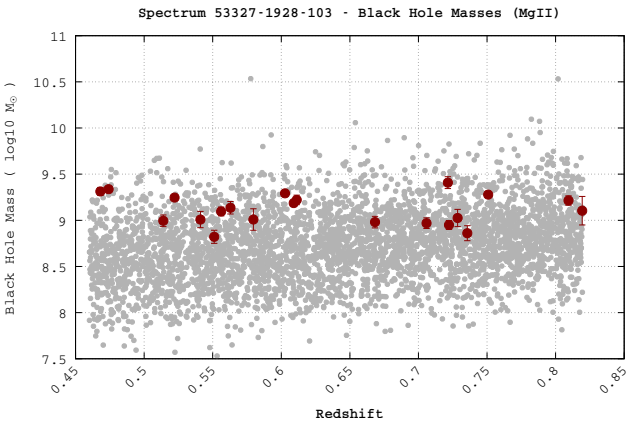

FIG. 13: Black hole mass estimates based off Mg II for Spectrum 53327-1928-103 matches

to a later date.

VIII. CONCLUSIONS AND FUTURE STUDIES

There appears to be evidence of at least one set of standard candle quasars with inclinations that further groups may also exist. While some cursory discussion of characteristics of interest was conducted above, a proper quantification of these possibly identifying features (and any others found) is necessary. At this time, no assertion is made that the discussed features are characteristic or indicative of the specific set given.

A. Identification of Characteristic Features

Should identifying features be established, it is assumed that they will not be completely independently selectable. This implies that they would be applied as additional (if not primary) filters, independent of χ^2 processes. With the added discriminatory ability that these identifying passes may provide, the χ^2 restrictions currently imposed may be relaxed, reducing the likelihood of eliminating desirable objects. This may also control for the high-uncertainty bias noticed, although a proper quantification of χ^2 values may reveal the source of this bias.

B. Quantification of χ^2 Values

With the discussed inherent dependence of χ^2 values both magnitude and the total number of points being matched over, these are greatly in need of specific contextualizing. Some measure, perhaps in terms of `Flux Density`$^{-1}$ `Å`$^{-1}$, will need to be developed to give quantification as to why χ^2_1 is scientifically/statistically acceptable, while χ^2_2 is not.

C. Moving Average

Other filter methods of interest may include some kind of moving average over an effective range. This would entail "sliding" across two spectra with a, say, 10 Å range in 5 Å increments. An average of the flux densities in this range for each spectrum would be determined and compared. Should be average fall within some acceptable range, then the process is continued by advancing the range position by 5 Å and repeating until complete.

A "quality of match" value may be discerned beyond the simple boolean matched/not-matched by summing the difference of these averages. This is similar to the χ^2, but address the differences between limited and extensive matching ranges.

A similar process was developed prior to understanding the bitmask described earlier in the interest of detecting large spikes in the spectrum - something of a hot pixel detector method. This process was abandoned immediately after the bitmask system became clear, but it was implemented effectively. Should that code be salvageable from deep inside some git repository somewhere, it will be explicitly included with the rest of the library.

D. Slope Fitting

Slope fitting portions of the spectrum may also be effective. Vanden Berk, et al., discusses the different portions of the spectrum used for fitting and identification[9] and so that will not be rehashed here except to say that it may be a useful process. The DR7 Properties catalog also contains power law slope fit values around certain emission lines, but for other ranges SciPy's `curve_fit` method should have no difficulty providing accurate fits.

A similar process was undertaken, though not followed through due to time restrictions, by dividing the spectra and applying linear fits. The argument was based on the idea that if the slope of this divided spectrum was exceedingly flat, the two spectra were commonly shaped. In practice, this process did not require any mutual scaling (in fact, doing so would lose the relation between their luminosities), though point-to-point alignment is still critical. Often, spectra which were "bowed" would average to a flat linear slope and so a second-order polynomial fit was applied to select against these values. While this slope fitting can quickly become a rabbit hole across the entire spectrum, the division process - or the mean flux density of the divided spectrum - may be useful to compare relative intensities of limited portions such as emission lines.

E. Line Luminosity and Equivalent Widths

In the interest of addressing the Doppler broadening described in the matched set above, some relation may be substantiated through equivalent width or full-width at half-maximum measurements. These are already tabulated in the DR7 QSO Properties[8].

These values were only recently extracted from the DR7 Properties catalog. At the time of this writing, no investigation has been made, so it is speculated that if these lines are discriminatory, the information may prove useful.

F. Extension to Further Redshifts and Databases

The extension of this project to further redshifts, or even non-Sloan datasets, may also provide insight to the effect. If the effect occurs at a low probability, an increase in the number of available quasars of interest would correspond to an increased likelihood of detection. Should an affected set exist in the current reduced catalog, both

the fit and modeled evolution may provide guidance for detection outside this initial redshift range.

However, some method of establishing similarity across these sets will need to be developed. Considering this redshift range (0.46, 0.82) was selected specifically to be inclusive of Mg II and Hβ emission lines, some connecting anchor will be necessary to ensure matched objects have some mutual correlation. The magnesium II line χ^2 fit is useful, but far from independently reliable. Considering the luminosities of this catalog already exist at the brightest edge of the SDSS for the redshift range, it is not recommended to investigate lower redshifts. In short, higher redshifts will result in increasingly blueward rest frame spectra, quickly eliminating the visibility of Hβ and redward features.

It is also recommended to continue to operate in the Sloan survey rather than extending to other databases - at least until an effect can be properly identified. Much can be said for equally calibrated spectra, available from the same location and packed in the same format, but most of that conversation speaks for itself. There may be some value if a 1:1 identification system exists between the SDSS and some other QSO database, though it is doubted that a great deal of light will be generated from all that heat.

IX. ACKNOWLEDGMENTS AND THANKS

This project was initially made possible by the Richard V. Mancuso Summer Research Scholarship. The author would like to thank Dr. Eric Monier for his guidance, insight and (well tested) patience. Additional thanks go out to the faculty in the Department of Physics of The College at Brockport.

[1] C. Hazard, M. B. Mackey, and A. J. Shimmins, Nature **197**, 1037 (1963).
[2] M. Schmidt, Nature **197**, 1040 (1963).
[3] E. Hubble, Proceedings of the National Academy of Sciences **15**, 168 (1929).
[4] E. P. Hubble, The Observatory **48**, 139 (1925).
[5] B. Smith, "Sky the limit on galaxy quest," `http://www.theage.com.au/national/sky-the-limit-on-galaxy-quest-20090109-7dp0.html` (2009), Accessed: 2017-03-15.
[6] R. Kotak, in *RS Ophiuchi (2006) and the Recurrent Nova Phenomenon*, Astronomical Society of the Pacific Conference Series, Vol. 401, edited by A. Evans, M. F. Bode, T. J. O'Brien, and M. J. Darnley (2008) p. 150.
[7] J. E. Gunn, W. A. Siegmund, E. J. Mannery, R. E. Owen, C. L. Hull, R. F. Leger, L. N. Carey, G. R. Knapp, D. G. York, W. N. Boroski, *et al.*, The Astronomical Journal **131**, 2332 (2006).
[8] Y. Shen, G. T. Richards, M. A. Strauss, P. B. Hall, D. P. Schneider, S. Snedden, D. Bizyaev, H. Brewington, V. Malanushenko, E. Malanushenko, *et al.*, The Astrophysical Journal Supplement Series **194**, 45 (2011).
[9] D. E. Vanden Berk, G. T. Richards, A. Bauer, M. A. Strauss, D. P. Schneider, T. M. Heckman, D. G. York, P. B. Hall, X. Fan, G. Knapp, *et al.*, The Astronomical Journal **122**, 549 (2001).
[10] Digital Equipment Corporation, "Why is Wednesday, November 17, 1858 the base time for OpenVMS?" `http://www.slac.stanford.edu/~rkj/crazytime.txt` (1999), Accessed: 2017-02-28.
[11] Sloan Digital Sky Survey, "Reading individual spectra files (spSpec files) - SDSS DR7," `http://classic.sdss.org/dr7/products/spectra/read_spSpec.html` (), accessed 2017-02-03.
[12] Sloan Digital Sky Survey, "Survey Interface File for Spectroscopic Output 1-d Spectrum," `http://classic.sdss.org/dr7/dm/flatFiles/spSpec.html` (), accessed: 05 Jan 2017.
[13] P. C. Hewett and V. Wild, Monthly Notices of the Royal Astronomical Society **405**, 2302 (2010).

X. FIGURES & TABLES

```
"53345-0123-387" :
        { "ab" : 19.21256
        "ab_err" : 0.00213
        "z" : 0.5676
        "z_err" : 0.0076
        ... }
"52678-4678-115" :
        { "ab" : 20.95123
        "ab_err" : 0.01691
        "z" : 0.63329
        "z_err" : 0.08821
        ... }
"56348-2232-556" :
        { ... }
```

FIG. 14: A concept example of how the `namestring dict` format in the project is laid out.

χ^2 FILTER LIMITATIONS	
Emission Lines	15
Continuum	100
Standardized AB Magnitude	20

TABLE III: Maximum χ^2 values used in the full catalog search process.

```
namestring=51608-0267-078,z=0.512379,gmag=18.665100
wavelength,flux density,error
3799.0960480000003,12.744149684906006,1.65735006332
3800.608427,34.6014986038208,6.1407995224
3802.120806,31.26889991760254,5.767977714538574
          .
          .
          .
\eof
```

FIG. 15: Example of text storage format for a spectrum

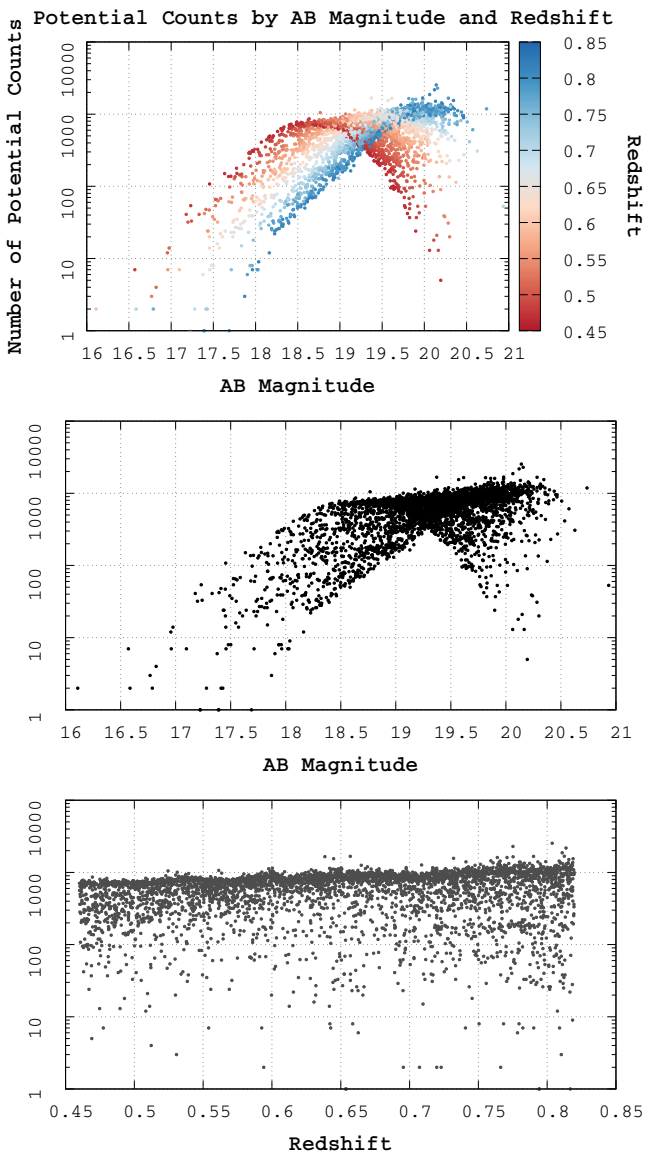

FIG. 16: The count of potential matches to each object in the catalog along its modeled evolution, within $\pm 0.5\sigma$. **Top**: Distribution by AB Magnitude, shaded to show Redshift relation. **Middle**: The potential count distribution, solely by AB Magnitude. **Bottom:** The potential count distribution solely by Redshift.

PROJECT COMPUTATIONAL HARDWARE

	Laptop	Primary Desktop	Standby Desktop
Model:	Asus UX330UA-AH54	Custom Build	Custom Build
Processor:	Intel Core i5-7200U	Intel Core i7-6800K	Intel Core i7-2600K
RAM:	8 GB	32 GB	8 GB
Storage:	1 TB SSD	1 TB SSD	512 GB SSD
Operating System:	Dual Boot: Windows 10 64-bit, Linux Mint 18.1	Windows 10 64-bit	Linux Mint 18.1

TABLE IV: A breakdown of the computational hardware used in this project.

MATCHING SET AND PROPERTIES TO SPECTRUM `53327-1928-103`

MJD-PLATE-FIBER ID	REDSHIFT	AB MAGNITUDE	$\mathrm{LOG}_{10}M_{\odot}\mathrm{H}\beta$	$\mathrm{LOG}_{10}M_{\odot}\mathrm{Mg\ II}$
53466-1675-004	0.46806 ± 0.00051	17.95 ± 0.031	9.170 ± 0.015	9.312 ± 0.043
52000-0554-553	0.47411 ± 0.00051	17.23 ± 0.024	9.090 ± 0.036	9.337 ± 0.028
52370-0793-549	0.51397 ± 0.00053	18.26 ± 0.065	8.984 ± 0.050	8.994 ± 0.062
53877-2512-582	0.52216 ± 0.00055	17.58 ± 0.042	9.053 ± 0.043	9.246 ± 0.022
53521-1714-571	0.54106 ± 0.00054	18.13 ± 0.058	8.890 ± 0.083	9.008 ± 0.088
54144-1845-637	0.55110 ± 0.00054	18.75 ± 0.063	8.825 ± 0.044	8.821 ± 0.070
52824-1393-056	0.55613 ± 0.00057	18.53 ± 0.048	8.822 ± 0.062	9.095 ± 0.048
53474-2007-171	0.56297 ± 0.00055	18.31 ± 0.035	9.016 ± 0.040	9.135 ± 0.067
53795-2216-298	0.57965 ± 0.00055	18.20 ± 0.039	9.039 ± 0.053	9.00 ± 0.11
54505-2662-548	0.60271 ± 0.00056	18.54 ± 0.070	9.204 ± 0.020	9.292 ± 0.028
52468-0729-143	0.60906 ± 0.00056	18.73 ± 0.067	9.085 ± 0.026	9.186 ± 0.029
52258-0412-423	0.61123 ± 0.00056	18.04 ± 0.023	9.05 ± 0.10	9.220 ± 0.051
54499-2616-517	0.66831 ± 0.00061	18.86 ± 0.061	9.182 ± 0.049	8.980 ± 0.062
51986-0294-555	0.70589 ± 0.00062	19.35 ± 0.098	8.87 ± 0.10	8.968 ± 0.057
52670-1206-290	0.72143 ± 0.00075	18.81 ± 0.074	9.09 ± 0.15	9.407 ± 0.064
53819-2226-431	0.72224 ± 0.00061	18.71 ± 0.053	8.56 ± 0.22	8.949 ± 0.048
52228-0721-281	0.72856 ± 0.00061	19.03 ± 0.065	9.320 ± 0.047	9.025 ± 0.093
52991-1270-380	0.73563 ± 0.00061	19.30 ± 0.10	8.68 ± 0.11	8.860 ± 0.081
53327-1928-103	$\mathbf{0.75083 \pm 0.00064}$	$\mathbf{19.09 \pm 0.096}$	$\mathbf{9.286 \pm 0.040}$	$\mathbf{9.278 \pm 0.025}$
53464-1674-428	0.80941 ± 0.00063	19.60 ± 0.11	8.89 ± 0.22	9.214 ± 0.050
54242-2750-160	0.81940 ± 0.00067	19.20 ± 0.13	9.11 ± 0.10	9.10 ± 0.15

TABLE V: Redshift-sorted members of the matching set to Spectrum `53327-1928-103`

MATCHING SET AND PROPERTIES TO SPECTRUM 54259-1832-281

MJD-PLATE-FIBER ID	REDSHIFT	AB MAGNITUDE	$\mathrm{LOG}_{10}M_{\odot}\mathrm{H}\beta$	$\mathrm{LOG}_{10}M_{\odot}\mathrm{Mg\ II}$
53084-1380-159	0.50957 ± 0.00055	18.417 ± 0.037	8.87 ± 0.16	8.973 ± 0.039
53472-1971-397	0.54046 ± 0.00055	18.877 ± 0.066	8.921 ± 0.054	9.044 ± 0.050
54259-1832-281	**0.55330 ± 0.00055**	**18.937 ± 0.059**	**9.008 ± 0.029**	**9.122 ± 0.038**
53054-1436-525	0.56198 ± 0.00059	19.114 ± 0.072	9.011 ± 0.076	8.836 ± 0.055
54477-2611-496	0.63432 ± 0.00058	19.169 ± 0.053	8.776 ± 0.055	8.825 ± 0.046
51999-0535-160	0.66097 ± 0.00060	19.412 ± 0.078	8.980 ± 0.057	9.068 ± 0.062
54230-2172-451	0.66612 ± 0.00072	19.384 ± 0.070	8.742 ± 0.050	9.001 ± 0.064
54052-2432-205	0.67718 ± 0.00059	19.178 ± 0.062	9.164 ± 0.062	9.125 ± 0.068
53144-1702-513	0.73432 ± 0.00063	19.43 ± 0.11	9.20 ± 0.13	9.156 ± 0.053
52228-0721-583	0.74303 ± 0.00069	19.505 ± 0.095	8.943 ± 0.23	8.991 ± 0.043
52288-0520-473	0.74498 ± 0.00076	19.446 ± 0.079	9.350 ± 0.047	9.389 ± 0.074
54139-2425-558	0.79295 ± 0.00078	19.921 ± 0.107	9.090 ± 0.102	8.932 ± 0.056

TABLE VI: Redshift-sorted members of the matching set to Spectrum 54259-1832-281

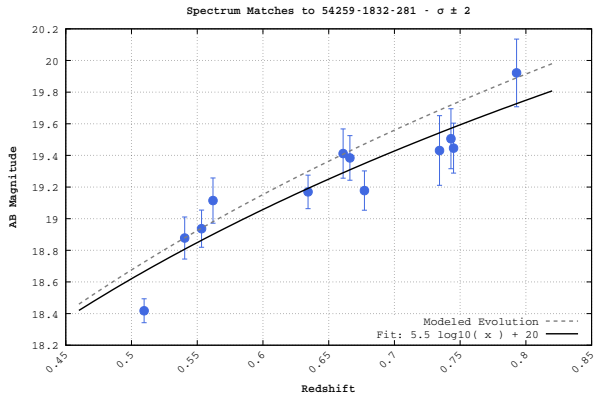

FIG. 17: Matches to Spectrum 54259-1832-281 with fit and expected evolution.

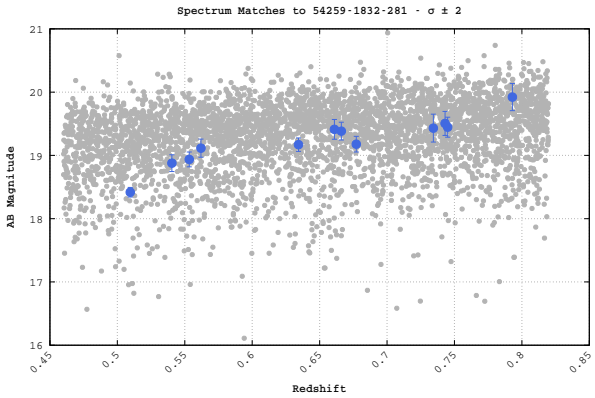

FIG. 18: Matches to Spectrum 54259-1832-281 as compared to the entire catalog.

SDSS DR7 ERROR MASKS

BIT	HEX	SDSS CODE	DESCRIPTION
None	0x000	SP_MASK_OK	No issue detected
0	0x001	SP_MASK_NOPLUG	Fiber not listed in plugmap file
1	0x002	SP_MASK_BADTRACE	Bad trace from routing TRACE320CRUDE
2	0x004	SP_MASK_BADFLAT	Low counts in fiberflat
3	0x008	SP_MASK_BADARC	Bad arc solution
4	0x010	SP_MASK_MANYBADCOL	More than 10% pixels are bad columns
5	0x020	SP_MASK_MANYREJECT	More than 10% pixels are rejected in extraction
6	0x040	SP_MASK_LARGESHIFT	Large spatial shift between flat & object position
16	0x10000	SP_MASK_NEARBADPIX	Bad pixel within 3 pixels of trace
17	0x20000	SP_MASK_LOWFLAT	Flat field less than 0.5
18	0x40000	SP_MASK_FULLREJECT	Pixel fully rejected in extraction
19	0x80000	SP_MASK_PARTIALREJ	Some pixels rejected in extraction
20	0x100000	SP_MASK_SCATLIGHT	Scattered light significant
21	0x200000	SP_MASK_CROSSTALK	Cross-talk significant
22	0x400000	SP_MASK_NOSKY	Sky level unknown at this wavelength
23	0x800000	SP_MASK_BRIGHTSKY	Sky level > flux + 10*(flux error)
24	0x1000000	SP_MASK_NODATA	No data available in combine B-spline
25	0x2000000	SP_MASK_COMBINEREJ	Rejected in combine B-spline
26	0x4000000	SP_MASK_BADFLUXFACTOR	Low flux-calibration or flux-correction factor
27	0x8000000	SP_MASK_BADSKYCHI	$\mathrm{Chi}^2 > 4$ in sky residuals at this wavelength
28	0x10000000	SP_MASK_REDMONSTER	Contiguous region of bad chi^2 in sky residuals
30	0x40000000	SP_MASK_EMLINE	Emission line detected here

TABLE VII: Bitmask codes and corresponding error used in SDSS DR7 Spectrum files

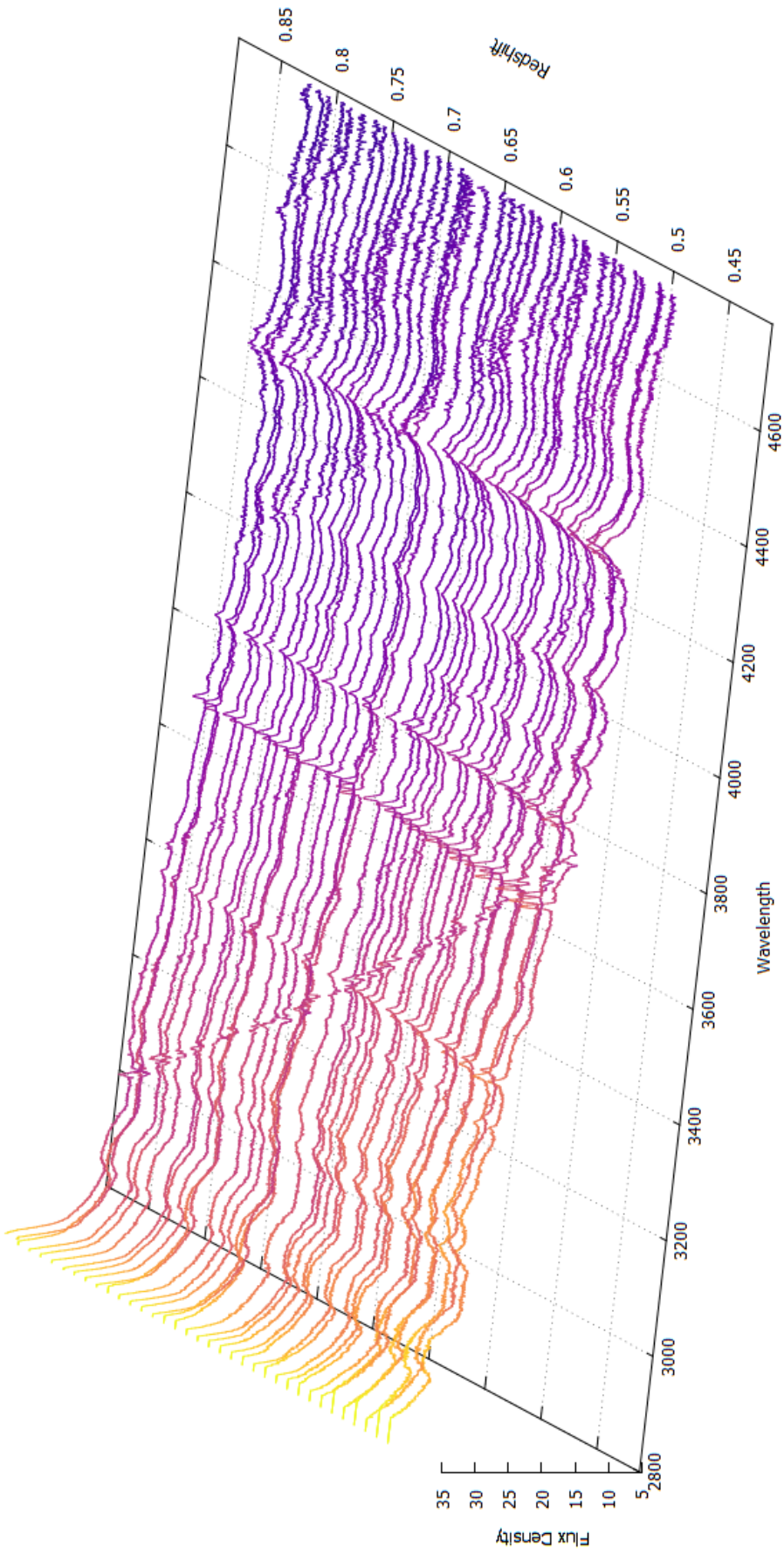

FIG. 19: A series of composite spectra generated by Redshift $\Delta z = 0.01$ bins. The "Monier Lines" can be observed as small spike in the continuum beginning near ≈ 3800 Å at $z = 0.46$, proceeding blueward with increasing redshift to ≈ 3000 Å at $z = 0.82$. This receding pattern with increasing redshift can be seen, though much weaker, can be observed in numerous places along the spectra.
This relation indicates the absorber is likely both stationary and isotropic, leading to the speculation that the Monier Lines are atmospheric absorbers.