Spectroscopic sky subtraction and data optimization in observation exposures from the Sloan Digital Sky Survey fiber optic telescope

Presented in partial fulfillment of the Undergraduate Honors Research Distinction for The Ohio State University department of Electrical and Computer Engineering.

By

Sarah Folkens

Undergraduate program in Electrical and Computer Engineering

Advisors

Dr. Jennifer Johnson

Dr. Kiryung Lee

Abstract

The Ohio State University is a participating institution in the Sloan Digital Sky Survey (SDSS), whose mission is creating a 3 dimensional map of the universe, including spectra of astronomical objects. Key spectral information includes the strength of emission and absorption lines. The purpose of this research is to identify an improved method of subtracting the unwanted sky signal from the desired star signals using the data collected by SDSS-IV. The project also intends to identify the fewest number of calibration sky fibers that should be used in future SDSS-V observations in order to increase the amount of science data collected throughout the life of the telescope.

Completing this research began with obtaining access to the data, and collaboration with members of the SDSS team: Dr. Johnson at OSU, and Dr. Holtzman at the New Mexico State University. My analysis was guided by Dr. Kiryung Lee at OSU's ECE department. First, a mathematical model of the data characteristics was developed. Next, appropriate methods of sky subtraction were identified, to compare against the current method used by SDSS. This included principal component analysis (PCA), as well as sparse PCA. Evaluation of the methods was performed using synthetic data, followed by evaluation with measured data. Fits files which were generated during SDSS-IV were used to plot and align the fibers' spectra for subtraction. Then an evaluation process was developed to calculate subtraction errors, since the ground truth in this case was unknown. Finally, the minimum number of calibration fibers required to meet noise specifications was identified.

The results began with the synthetic case in which the model was sparse data set of n samples and p pixels in which the samples were assumed similar, with variations in amplitude and white gaussian noise. The SPCA signal representation yielded the lowest RMS error in the synthetic case. Next, the measured data was used and partitioned into training and testing samples. The results from measured data deviated significantly from the synthetic case and the averaging of the four nearest fibers yielded the lowest RMS error. This is the method currently used by SDSS, but which had not been evaluated for error analysis until these experiments. The study concluded that no significant information would be lost by modifying the current practices and reducing the number of calibration sky fibers used from 35 down to 15, which would increase the science data per-observation by 7.5%.

Acknowledgements

This study depended on the work and assistance of many contributing parties of the Sloan Digital Sky Survey team. In particular, advisors Dr. Jennifer Johnson and Dr. Kiryung Lee actively guided this thesis project. Dr. Jon Holztman was also a frequent correspondence when discussing the data collected by SDSS-IV. Finally, many thanks go out to Sten Hasslequist and the other members of the APOGEE effort who invited me to share these findings at their November telecon.

Table of Contents

List of Figures

List of Tables

Background and Motivation

The Sloan Digital Sky Survey (SDSS-V) uses a plate populated with optical fibers in the focal plane of the telescope with a 2.5-meter mirror. The fibers collect flux within three wavelength bands and store the data for processing. In general, flux from stars will display narrow peaks (emissions) and most often troughs (absorptions) at certain wavelengths, depending on their elemental composition and relative velocities. These absorption and emission lines are the desired observation data collected from the SDSS-V telescope. However, since SDSS-V is a ground-based telescope, the light travels through the earth's atmosphere before being collected. The telescope also picks up additional signals like moonlight scattering off dust in the solar system. This results in unwanted emission and absorption lines which must be removed in order to include only the star's data in the results. The process of removing these unwanted signals is referred to as sky subtraction [1].

The current method of sky subtraction used by SDSS-V is the following: 35 out of the 300 total fibers used in each observation plate are scattered around the plate and aimed towards relatively empty places in the sky field. After an observation, the signals due to the earth's atmosphere needed to be removed from the science data. Since it was expected that the sky-only spectra would vary slightly based on their locations, the four closest sky fibers were averaged and subtracted from each science fiber. Thus, roughly ten percent of the observation capacity during each exposure (ten percent of the observation plate) was dedicated to observations of blank sky for use during sky subtraction.

While this averaging method makes intuitive sense, there is currently no quantification for its effectiveness or efficiency. The amount of sky fibers used, their positioning in the plate, and the post-processing method of averaging and subtraction are without justification. This study seeks to identify a means of measuring and comparing the success of sky subtraction methods. It also endeavors to implement an improved sky subtraction algorithm for post processing. The importance of this study is that it would lead to a more precise estimate of the sky subtraction, and allow the scientists to use the fewest possible number of fibers for sky collection so that the maximum amount of fibers can be used to collect science data.

Significance

In order to push the limits of exploring the surrounding universe, astronomers have continuously sought to improve both the amount and precision of the information available to scientists in the field. In this study we hope to increase the amount and precision of information collected by the Sloan Digital Sky Survey (SDSS-V) including the strength of emission and absorption lines in the spectra of stellar objects. By developing a method of evaluating and maximizing the accuracy of sky subtraction, the confidence level and precision of the SDSS-V data can be more fully understood and better estimated. Furthermore, once the method of subtraction is optimized, the number of fibers used to collect science data can also be optimized. For example, if only 15 sky fibers were required to accurately subtract sky from the remaining 285 fibers, then the efficiency of each observation would increase from 90% to 95% science data. This would increase the amount of data collected by 5% throughout the life of the telescope. Therefore, even though the cost of this project was relatively small as compared to other costs, the impact has potential significance to both the amount and quality of data collected by the SDSS-V project.

This design has the potential to have environmental, social, and economic impact as part of the greater SDSS effort. Some impacts are difficult to predict. For instance, it could be that the science gleaned from SDSS will illuminate physical truths about the nature of the galaxy and further empower space exploration, or even colonization. Or it could be that the technologies developed within the SDSS effort, including the sky subtraction technologies discussed within this paper, could be applied to other fields: medical, communications, economic, etc. For example, this project used analytical methods which might be useful to a researcher studying social problems such as housing discrimination and systemic injustices. The impact of this project, therefore, is undeterminable. Both discoveries and technologies developed within SDSS and this paper should not be exclusionary, but rather shared and used to impact the wider public in a positive way. For this reason, the fidelity and communication of this project was given high priority throughout the life of the project.

Problem Statement

This project tested the assumption that the current method of sky subtraction used by the Sloan Digital Sky Survey (SDSS-V) was the most accurate and efficient method. The current method of subtraction can be visualized using the graphic below [2].

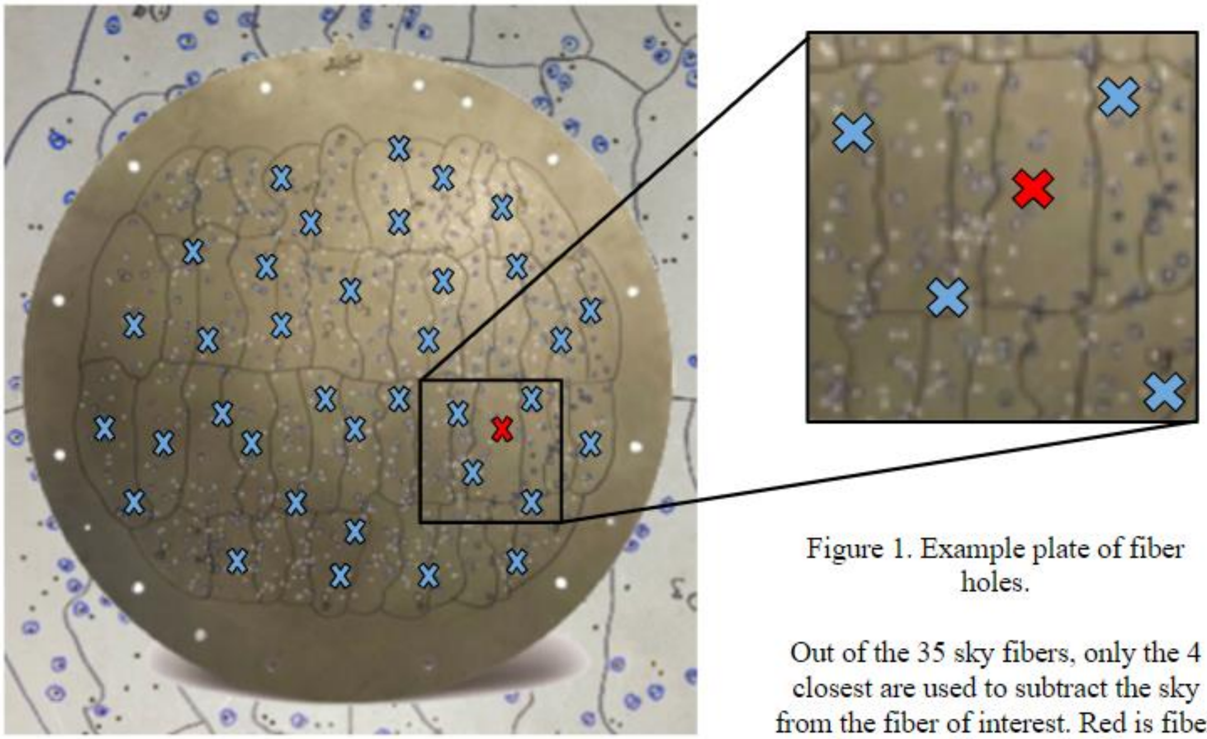


Figure 1. Example plate of fiber holes.

Out of the 35 sky fibers, only the 4 closest are used to subtract the sky from the fiber of interest. Red is fiber of interest, blue are sky fibers.

A range of possible data manipulation methods was researched and considered for viability. The goal was to reduce the sky noise to less than 1 part in 100, such that it had levels less than other sources of noise; noise sources beyond those from sky subtraction were beyond the scope of this project. The goal was also to minimize the number of sky fibers used per plate to accomplish the desired noise levels due to subtraction. The following section outlines the methods used in pursuit of a sky subtraction method that would satisfy these two goals.

Design Process and Methodology

This design process is part of the much larger SDSS data pipeline effort. Detailed below are the preliminary investigations, parameter justifications, simulated results, measured results, and conclusions within the scope of this project.

Preliminary Investigation

The project began with developing an understanding of the problem. Dr. Jennifer Johnson was a member of the SDSS team and initiated this project to access the spectral data, quantify the sky subtraction error as it was currently done, and explore ways to minimize the number of sky fibers used per observation.

The process began with obtaining access to the data. This included creating accounts for data.sdss.org, sdss.org/dr16, and trac.sdss.org. Next, it was important to understand the data processing pipeline for the spectral data collected. First, the raw data underwent dark subtraction and flat fielding as initial calibration steps. Next, the 1-dimensional spectra were extracted and paired with the corresponding wavelength vectors, which were calibrated. The data accessed for this project was the 'ap1D' fits files, which came in the pipeline before sky subtraction and telluric correction were applied. Figure 1 in the Appendix shows sample spectra as plotted in Matlab and Figure 2 shows a flow diagram of the data pipeline.

Next, the data needed to be extracted and made useful. Matlab was selected to do the research and prototyping for this project. There was documentation of the ap1D data model in sdss.org but it was not always accurate. There was also documentation on which of the 300 fibers in each exposure were sky fibers. However, it was not documented that the fibers are indexed from the bottom upwards. These hurdles were overcome throughout the project as needed, and detailed documentation was written up as a reference for future use. Indeed, the reference documents made have already been used by an associate professor at Chungnam National University and saved him much time in finding the data he needed. The Documentation section of this report has more information on the data access process.

The current method of subtraction also required some explanation. After consulting Dr. Jennifer Johnson and Dr. Jon Holtzman it was found that the subtraction process used four sky fibers, averaged them, and then directly subtracted the resulting representation of sky from the signal of interest. There was no resizing or fitting of the signals. Furthermore, the four fibers averaged were selected in the following way: the sky fibers were measured in right ascension (RA) and declination (DEC) with reference to the fiber of interest. The algorithm would then select the four closest sky fibers, but exclude those that were more than 75 fibers away in slit position. The slit position corresponded to the ordering of the signals 1-300. The method of obtaining the RA and DEC positions of fibers was also detailed in the documentation guides and can be found also in the Matlab script, 'nearestFour.m'.

Preliminary investigation of the data included visual inspection of the spectra of 35 sky fibers in a single exposure. The data was visually very similar. It consisted of a large dynamic range in which a few large peaks were orders of magnitude greater than the continuum. Star fibers were also visually inspected and were found to exhibit characteristic features, as marked in Figure 4 below.

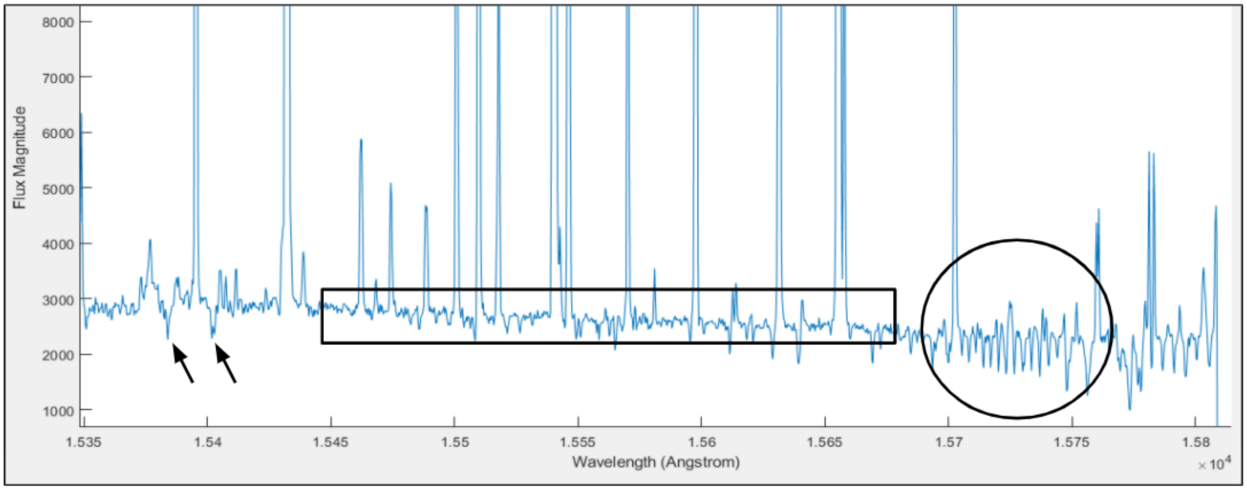


Figure 4. A sample Halo star spectra. Typical features of stars include: absorption lines (left), the eye-shaped peaks-and-troughs feature (right), and a gently sloping continuum (center).

Also included in preliminary investigation was a measure of the amount of variance accounted for in the first principal component of a set of sky fiber data. It was found in these initial explorations that more than 99% of the variance was accounted for in only the first principal component. This seemed to indicate that a single principal component could be used to represent the sky signal, which was to be subtracted out. Further discussions observed that the number of sky fibers, typically 35, was much less than the number of pixels in each 1-

dimensional signal, 2048 pixels. This indicated that the data sets used to characterize the sky signal should be classified as sparse.

Sparse methods of principal component analysis have been explored previously in the literature. *Sparse Principal Component Analysis* by Zou, Hastie, and Tibshirani was the foremost paper consulted to understand the methods used to address sparsity in the data [5]. These methods included the Lasso and the Elastic Net approach. Fortunately, Moshin Ali had previously published code in Matlab which implemented these algorithms, when provided a range of parameters [4]. All analysis performed in Matlab on both synthetic and measured data for this project leveraged this 2016 script by Ali.

Parameter Gathering

The parameter gathering occurred in two stages: the characterizing of the data, and the selection of appropriate parameters as input to the principal component analysis (PCA) and sparse principal component analysis (SPCA) algorithms.

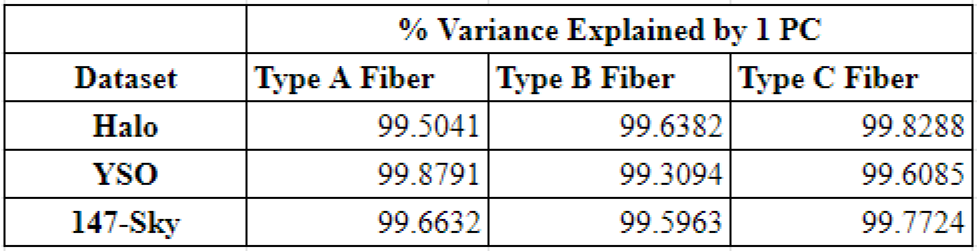
As previously mentioned, the dataset for the sky signals was classified as sparse. The number of samples, n, was far fewer than the number of pixels (or variables), p, such that $n<<p$. Additional assumptions were made about the data. First, it was assumed that the sky was similar across the observation plate. This assumption was supported by discussions with Dr. Jennifer Johnson, visual inspection of the data, and the high percentage of variability contained in a single principal component as seen in initial investigations. However, it was assumed that the amplitude could vary by a scaling factor. This was assumed because of possible imperfections in the fiber positions and the nature of the sensors, which operate based on photon excitation which is quantized. Thus, the sky vectors could be considered as each having the same direction at varying magnitudes. Finally, it was assumed that white gaussian noise would be added to each occurrence of the sky signal due to the AWGN from the measurement instruments.

All these assumptions led to a characterization of the dataset. This characterization supported the hypothesis that sky subtraction would have the least amount of error using the SPCA method, followed by the PCA method, and finally the averaging of four fibers method. This conclusion assumes that PCA would out-perform averaging since it would decrease the influence of noise of the resulting sky representation. Based on our model, we also asserted that sparse PCA could out-perform PCA by restricting the search space in pursuit of a better generalization.

The input parameters then needed to be selected for the PCA and SPCA algorithms. For PCA this was just the number of principal components used to represent the sky signal. For SPCA this included: number of principal components, whether to use soft thresholding, the number of iterations, and the number of non-zero variables.

To determine the number of principal components to use, the PCA was computed for a set of sky fibers. The percent variance accounted for by each principal component was also calculated. This is a built-in Matlab functionality. The results are shown in the table below.

Table 1. Three samples of percent variance contained in a single principal component.



	% Variance Explained by 1 PC		
Dataset	**Type A Fiber**	**Type B Fiber**	**Type C Fiber**
Halo	99.5041	99.6382	99.8288
YSO	99.8791	99.3094	99.6085
147-Sky	99.6632	99.5963	99.7724

With such a high percentage of the variance included in just one principal component, it was determined that the first principal component only would be used. This was later re-checked visually by plotting the RMS error rates of using 1 PC versus using 2 PCs. There was visually negligible difference. Therefore, the choice of using just 1 PC for the signal representation was confirmed.

Next, the parameters were selected for using the SPCA algorithm. First, it was determined that soft thresholding would be used because in this instance n<<p. 35 fibers in the training set was much less than the 2048 pixels per sample.

Cross validation was used to select the number of non-zero variables to use when computing the SPCA. This was first done with synthetic data, then measured data. A set of sky signals was partitioned into five groups. In each of five iterations, one fifth of the signals were used a training data, and the remaining signals were used as testing data. The RMS error was calculated for each iteration of the cross validation process and averaged. This method was applied repeatedly, each with a different number of non-zero variables. The resulting trend is shown below, with synthetic results on the left and measured results on the right.

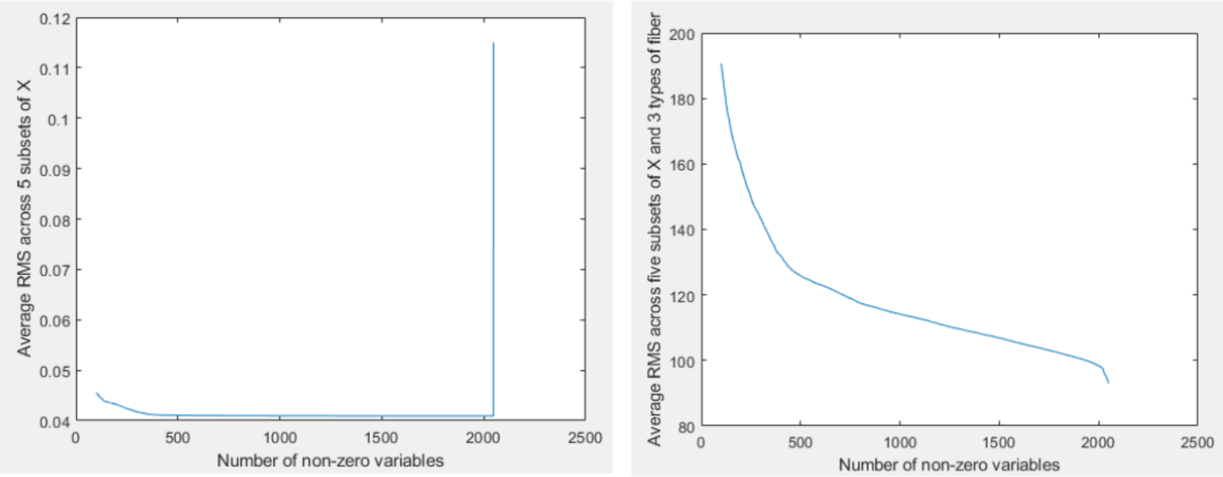


Figure 5. Cross validation was used to test the error rates at a range of 'stop' values, the number of non-zero variables. The synthetic results (left) show clear gains in accuracy up to about 450. The measured results (right) are less clear.

Looking at the plot on the left, the trend suggested that there was significant benefit in increasing the number of variables used up to about 450. Beyond 450 variables the trend shows a near-linear behavior with very little gain in accuracy. This suggests that the significant information is contained in only about 450 non-zero variables when performing SPCA. Notice also that the plot on the right, using measured data, showed a similar trend but with much less clear delineation in the exponential and linear regions. For the purposes of comparing the results of averaging, PCA, and SPCA methods a value of 450 non-zero variables was used based on the results from this cross-validation process.

Simulation and Prototyping

The project first simulated sky signals. A true sky signal representation was created. Next, it was modified to simulate the different signals that the sensors would record across the telescope's plate. The signals were multiplied by a random scalar and then AWGN was added with a SNR of 10 dB. This level of noise was greater than that expected from the telescope. Images of the synthetic signals can be found in Figure 6 in the Appendix.

The simulation used the assumption that a single, true sky signal was unchanging across the observation plate. Therefore, each of the sky subtraction methods was evaluated by subtracting that method's representation of the sky signal from the true signal, all synthetically created as described above.

The method of averaging four fibers was performed first. In the current SDSS method of sky subtraction, no fitting of the signal is applied so after averaging, the signal was directly subtracted. The result is shown below.

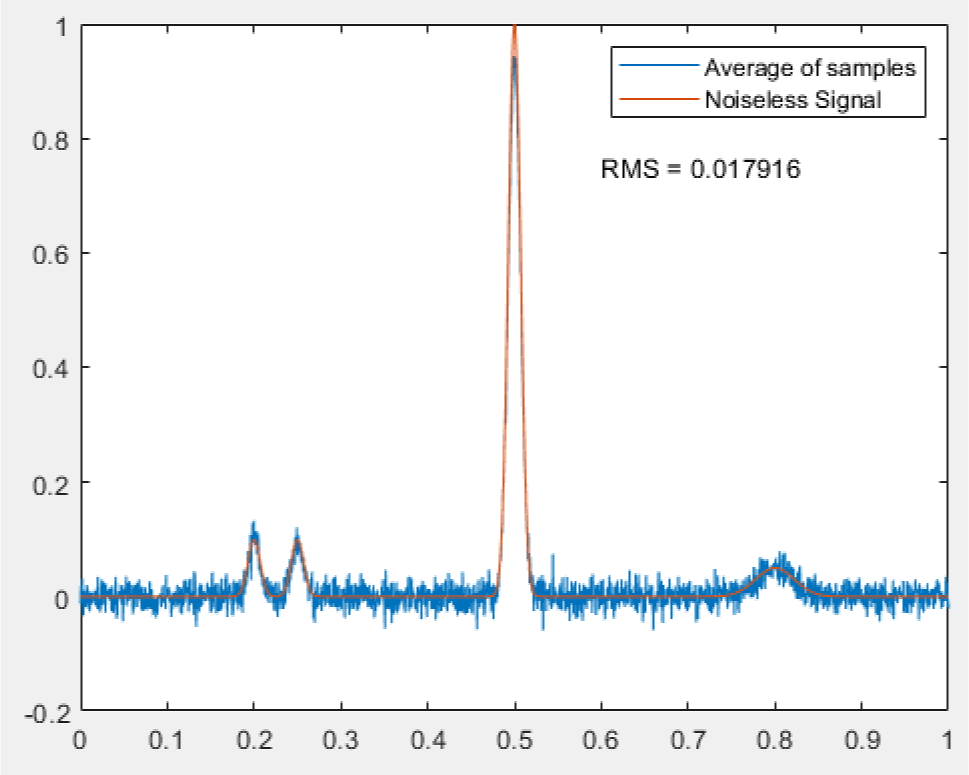


Figure 7. Simulated results of the current subtraction method: represent sky by averaging the 4 closest fibers.

The plot above shows the somewhat noisy representation of the sky signal as results from averaging. It also shows the RMS error calculated after subtraction. This method resulted in 0.017916 RMS error, which was the highest error of the three tested in this study.

Next the PCA of the synthetic sky signals was calculated and the first principal component was used to represent the signal. Here, the resulting sky representation was scaled to best fit the true sky signal before subtraction, since the PCA is assumed to be representative of the signal up to a scalar multiple. The result is shown below for the simulation.

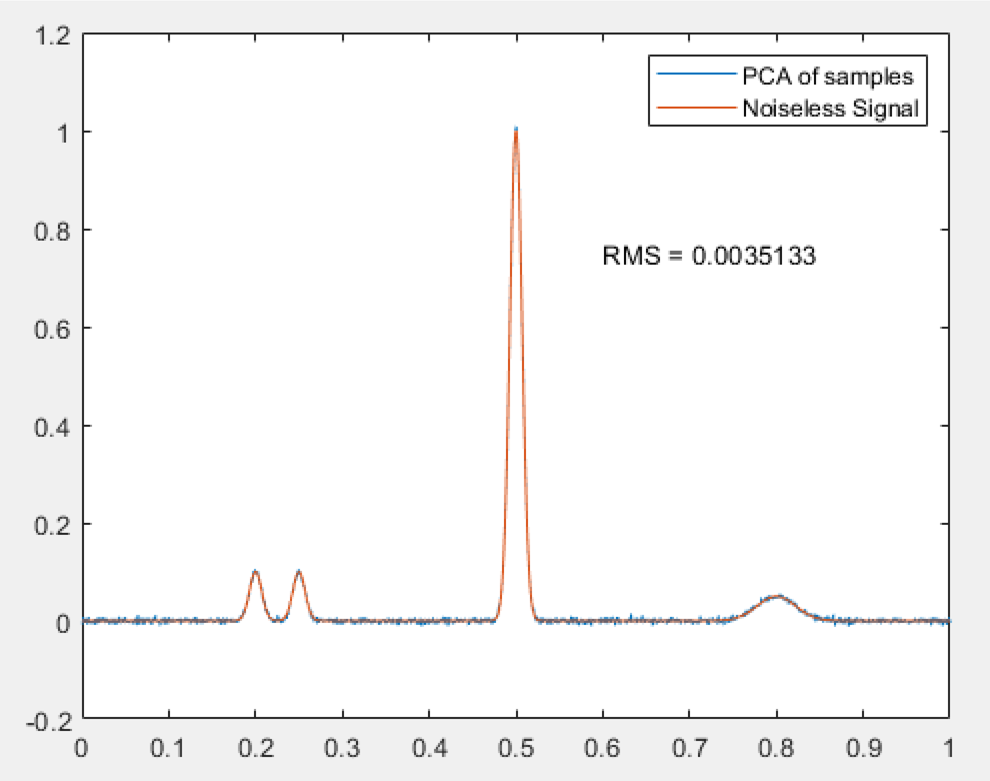


Figure 8. Simulated results of the PCA subtraction method: represent sky using the first principal component.

The plot above shows a sky signal representation very similar, visually, to the averaging method. The RMS error calculated was 0.0035133. This was less error than the averaging method, but as is shown next, more error than the SPCA method in this simulation.

Finally, the SPCA of the synthetic sky signals was calculated and the first principal component was used to represent the sky signal. 450 non-zero variables were used, and the representative signal was again scaled before subtraction. The result is shown below for the simulation.

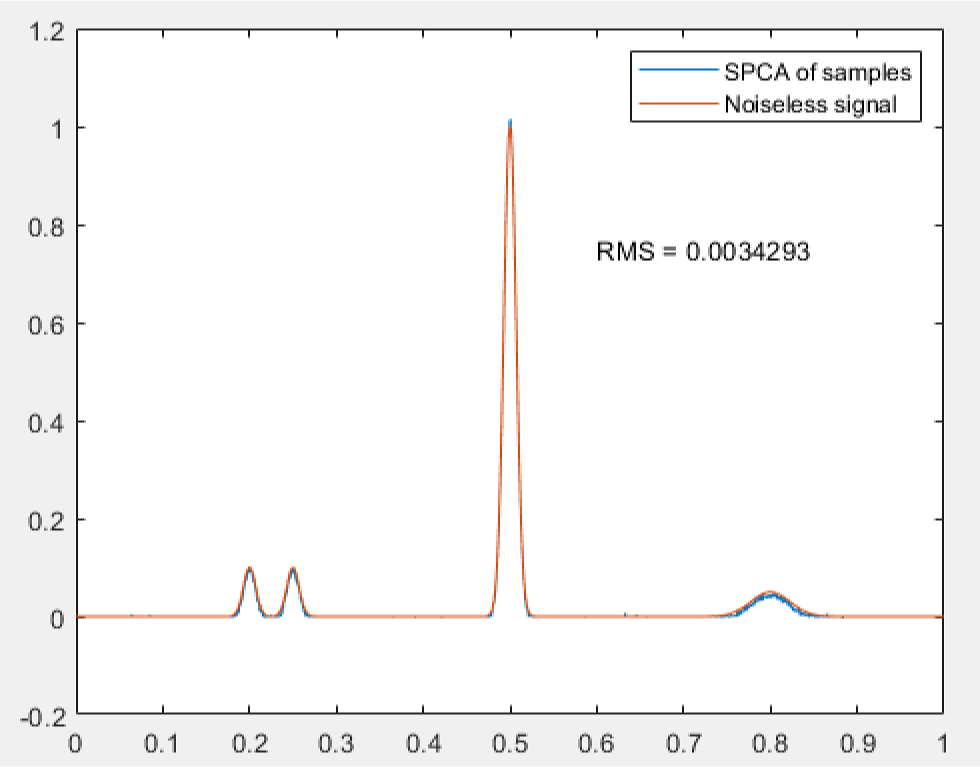


Figure 9. Simulated results of the SPCA subtraction method: represent sky using the first principal component with 450 non-zero variables.

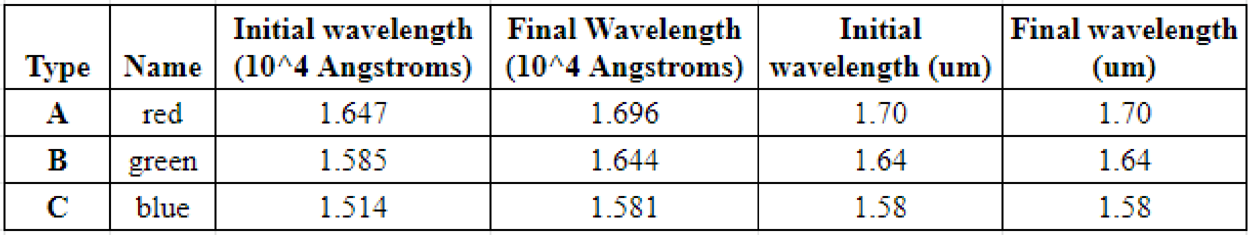
The result shown above is characteristic of using a sparsity model in the sky representation: it is visually a cleaner signal, with less over-fitting than the averaging or PCA methods. These results supported the initial hypothesis that the SPCA method would result in the best representation of the sky signal and the lowest RMS error, 0.0034293. However, this was the simulated case and no measured data was used. The next section describes the results for the measured case and discusses the implications of deviation from the simulated results, which were based on data model previously described.

Results from Measured Data

The hypothesis of these tests was that the SPCA method would result in the best representation of the sky signal and the lowest RMS error. This hypothesis was based on the data model previously described: a similar sky signal across the observation plate, with differences by a scalar magnitude and AWGN at the sensor level, and sparse in dimension. It was also based on the simulated results of the synthetic case, as described above.

The measured data used to test this hypothesis was gathered by the SDSS team during SDSS-IV and it used three types of fiberoptic cables at each location in the plate. The three fiber types corresponded to different wavelengths collected by the sensor. A table describing the three fiber types is shown below.

Table 2. Types A, B, and C fibers used in measured data results.



Type	Name	Initial wavelength (10^4 Angstroms)	Final Wavelength (10^4 Angstroms)	Initial wavelength (um)	Final wavelength (um)
A	red	1.647	1.696	1.70	1.70
B	green	1.585	1.644	1.64	1.64
C	blue	1.514	1.581	1.58	1.58

Each observation plate (shown previously, Figure 1) was populated with a type A, B, and C fiber in each hole. The location of the holes was described by their right ascension (RA) and declination (DEC) coordinates in the sky. Each plate in SDSS-IV included 35 holes for sky fibers, which point to areas of blank sky; this paper refers to these as sky fibers. The non-sky fibers on the plate are often referred to as fibers of interest, science fibers, or star fibers.

The current SDSS practice had been to use 35 sky fibers, spread out over each observation plate, to be used in sky subtraction. The number of fibers, 35, could be changed in future observations as needed. Therefore, the data used to test the hypothesis was a special dataset that included not 35 sky fibers, but 147 sky fibers, allowing multiple training sizes to be tested.

In each test, some of the sky fibers were designated as training data, and some of them were designated as test data. The RMS error was calculated by subtracting the sky signal representation from each of the fiber signals in the test set. Since the test set was also comprised of sky fibers, this resulted in a test of “sky minus sky” and the perfect method would yield a theoretical RMS result of zero after subtraction. The test was performed using 35 sky fibers since that was the current method, but also 15 fibers and 112 fibers.

These tests were performed in Matlab. To access the code, see the Documentation section. A flow diagram describing the Matlab tests is in Figure 10 in the Appendix. The results were performed in two ways. First, the average RMS error was measured for each method and with 15, 35, and 112-fiber training sets. The RMS error values by method and fiber type are in Table 3 in the Appendix, and a plot of the Type B per-pixel errors is in Figure 11 in the Appendix. A summery of the results are shown in the graph below.

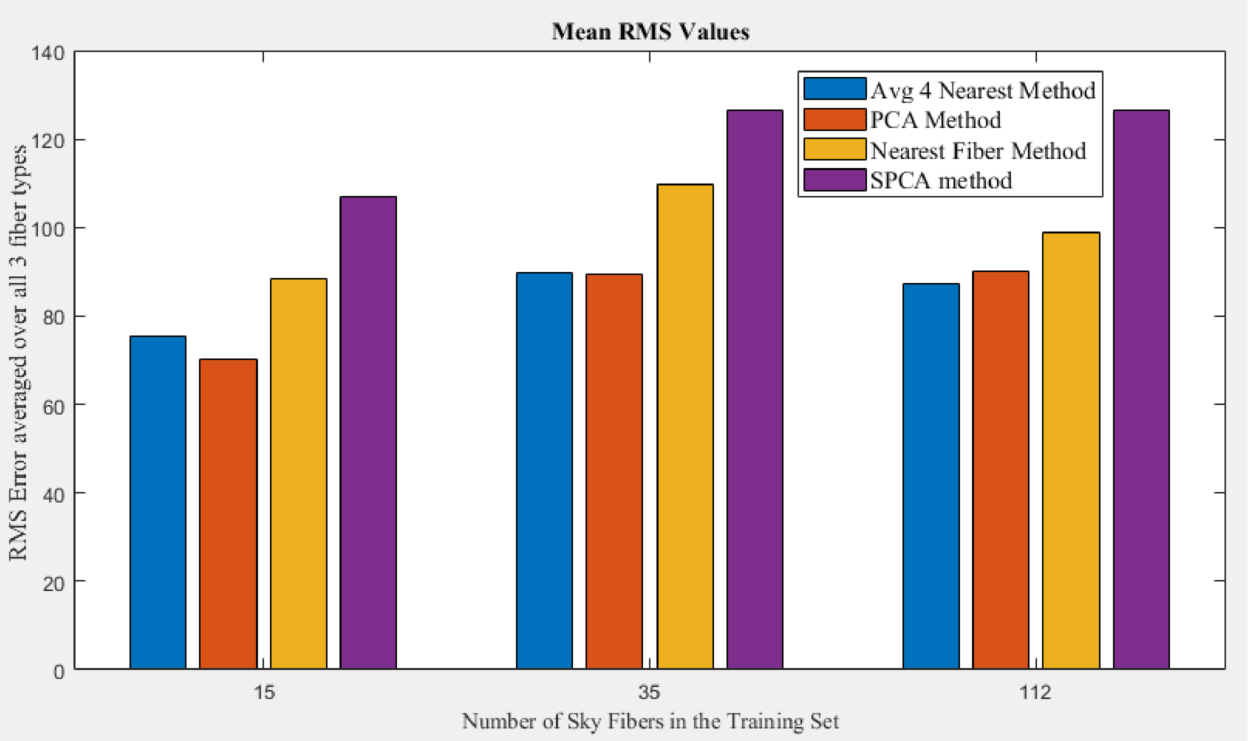
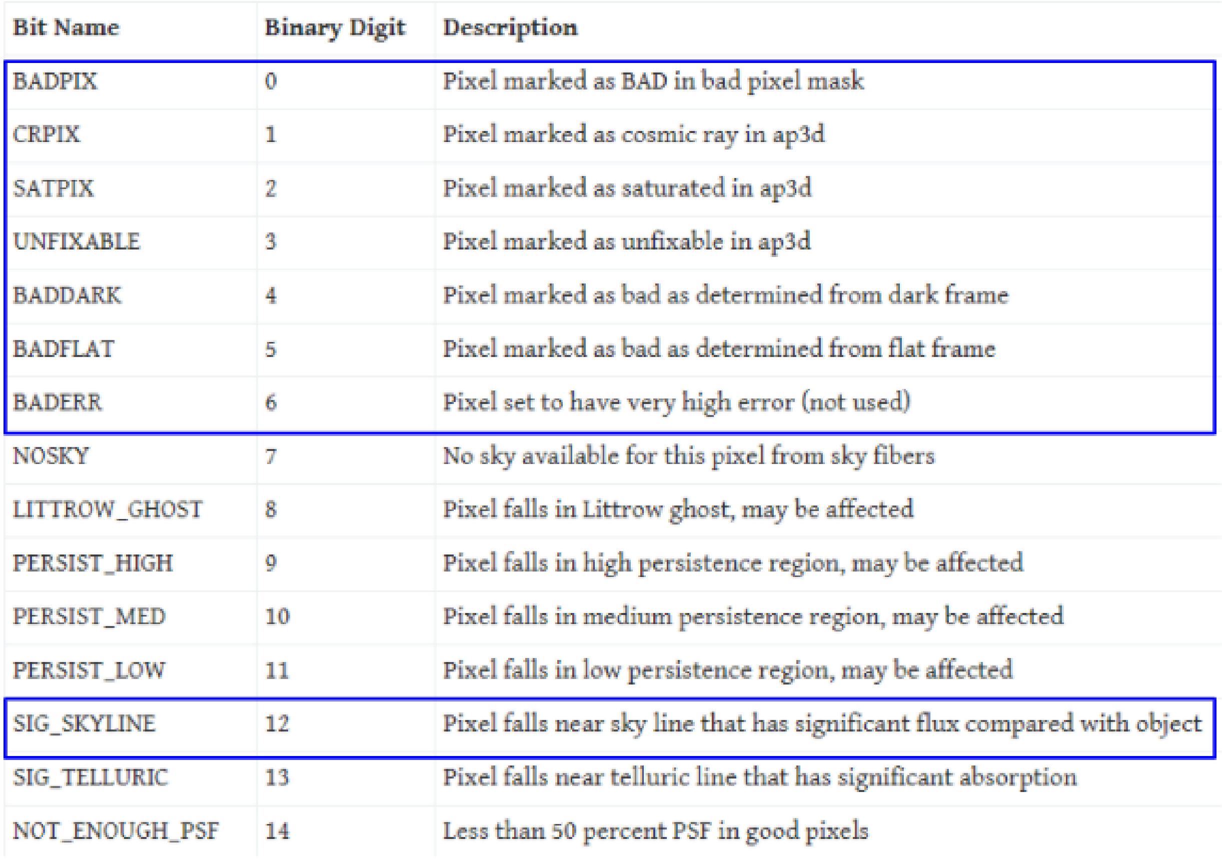


Figure 12. RMS values for the methods tested, averaged across fiber type.

Notice that the measured results differed greatly from the synthetic results. The averaging method yielded the lowest RMS error, whereas the SPCA method had the highest error rates. The exception what the Type A fiber using 15 sky fibers had a high error for the averaging method, which caused the PCA to have the lowest score for the 15-fiber case. Follow-up tests were done in order to explore these discrepancies between the measured and synthetic results. The follow-up tests can be found in the following section.

The second type of results found were most meaningful to the SDSS team. A mask of known good-pixels and bad-pixels for a typical stellar spectra was applied to the RMS per-pixel error vectors. Essentially, it was known and accepted that certain pixels would have no chance at an accurate data reading. The chart below shows the reasons pixels were excluded as well as which bit masks were used by this research project.

Table 4. Pre-existing pixel masks, the masks used in this result are mark in blue [3].



Bit Name	Binary Digit	Description
BADPIX	0	Pixel marked as BAD in bad pixel mask
CRPIX	1	Pixel marked as cosmic ray in ap3d
SATPIX	2	Pixel marked as saturated in ap3d
UNFIXABLE	3	Pixel marked as unfixable in ap3d
BADDARK	4	Pixel marked as bad as determined from dark frame
BADFLAT	5	Pixel marked as bad as determined from flat frame
BADERR	6	Pixel set to have very high error (not used)
NOSKY	7	No sky available for this pixel from sky fibers
LITTROW_GHOST	8	Pixel falls in Littrow ghost, may be affected
PERSIST_HIGH	9	Pixel falls in high persistence region, may be affected
PERSIST_MED	10	Pixel falls in medium persistence region, may be affected
PERSIST_LOW	11	Pixel falls in low persistence region, may be affected
SIG_SKYLINE	12	Pixel falls near sky line that has significant flux compared with object
SIG_TELLURIC	13	Pixel falls near telluric line that has significant absorption
NOT_ENOUGH_PSF	14	Less than 50 percent PSF in good pixels

As mentioned, the above masks were taken for a typical star signal such that if a pixel at a given wavelength was typically discarded it was not included in the error analysis described here. Next, the flux of the continuum was isolated so that the error rates could be compared to the magnitude of the continuum at that location. The continuum here refers to the sort of sloping shape that the signal makes which underlies all of its high frequency peaks and troughs.

The continuum of the star signal was isolated by using a median filter followed by a lowpass filter. The median filter removed all peaks and troughs from the signal and the lowpass filter smoothed out the continuum. The level of continuum was used to determine the threshold for acceptable error. Through discussions with the SDSS team, including Dr. Jennifer Johnson and Dr. Jon Holtzman, it was determined that the goal of the sky subtraction was to reduce the noise to less than 1 part per 100. This implied that the number of photons measured in the continuum at a given wavelength should be at least 100 times greater than the error incurred by the sky subtraction process.

The results were a tally of the number of pixels which exceeded this threshold, for each subtraction method, and also plots of the RMS error functions and the locations where the sky subtraction failed to meet this standard. A plot of the averaging method's results using 35 sky fibers is shown below.

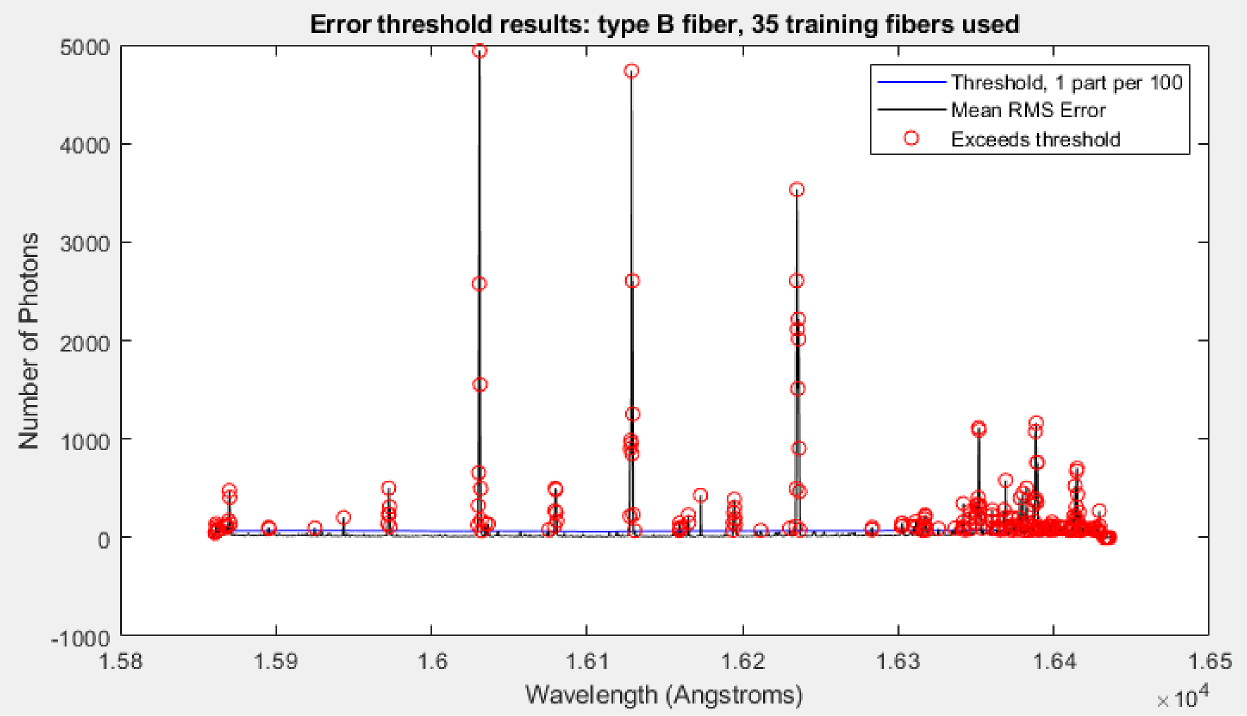


Figure 13. Using the current method of averaging 4 fibers out of 35, the error threshold is often exceeded.

The results plotted above are based on the current method of subtraction. The red circles show where the error rates exceeded the threshold. The error rate at many of the pixels exceeded this threshold. It was found that 242 out of 2015 of the viable pixels (after masking the 2048) exceeded the error threshold on average, using the current method used by SDSS for the type B fiber.

Additional Tests

The tests designed to evaluate subtraction methods produced different results in the measured case than in the simulated case. Tests were performed to diagnose this difference. The first test was a sanity check, to confirm the accurate calculations of the averaging, PCA, and

SPCA and guard against a design error. It was known that the PCA should perform at least as well as the averaging method if the test set was the same for both. So instead of using only four fibers for the averaging, a test was performed using all 35 fibers for both the averaging, PCA, and SPCA methods. As expected, when the averaging used all 35 fibers instead of the nearest 4, the PCA method had a lower error.

The next test performed was to test the new hypothesis that the simulation model assumptions were incorrect. More precisely, this new hypothesis was that the true sky signal was in fact not similar across the observation plate, but instead varied in a significant way based on RA and DEC position. If true, this would explain the superior performance of the averaging method, since that method excluded all but the closest fibers in RA and DEC when generating a sky signal to be subtracted. To test this hypothesis, k mean squared Bayesian clustering was used. It was expected that by using the elbow method and the silhouette method, appropriate groupings could be made in the dataset, and that then the PCA of these groupings would be more successful than the PCA of the entire plate of sky fibers.

Indeed, when this analysis was performed on the dataset before the data was normalized to a single lambda variable, the result showed that grouping the data in two groups scored above .75 on the silhouette test, and produced spatially meaningful groupings. The silhouette test and groupings by fiber location for this are shown below.

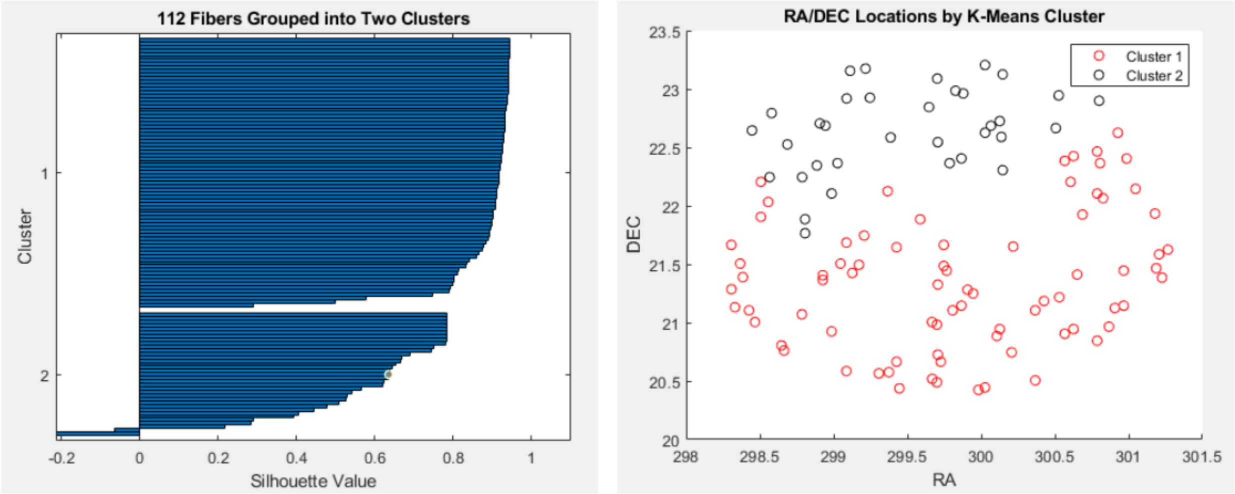
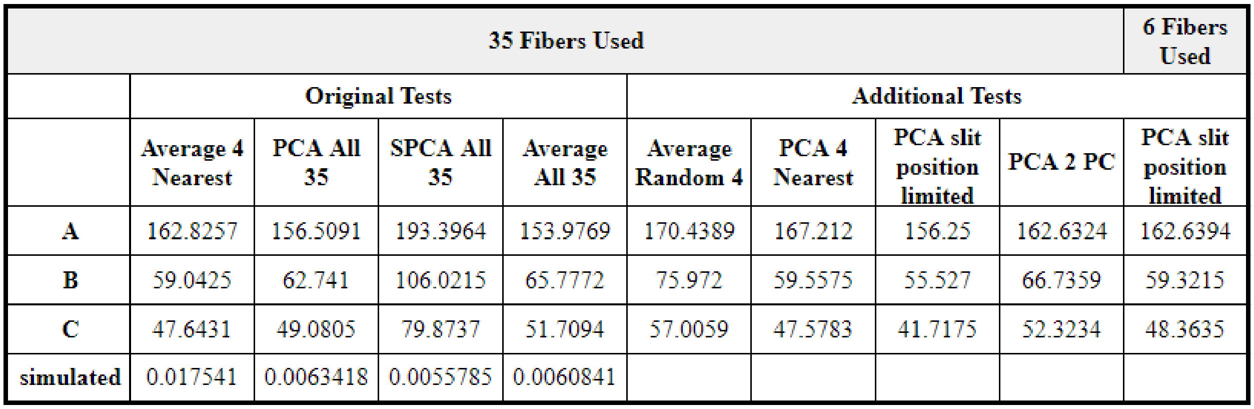


Figure 14. For the non-normalized dataset, k means clustering formed well-scoring groups.

However, this it was found that once the data was normalized to a single lambda variable, as it is in all the calculations of RMS error, this meaningful grouping declined and the clusters became weak or included all the data except a few outlier into a single cluster.

The problem remained that an improvement was made when the nearest four fibers method was used: when that algorithm was removed, performance dropped. So the next test performed was to use the same method of selecting four fibers, but perform PCA analysis on the data instead of averaging. The performance using this method had a lower RMS error than the current method of averaging four fibers. Then, the algorithm for selecting fibers was modified to sweep through values of 1 to 35 in order to see the optimum number of nearest-fibers to use in the PCA. Again, there was a surprise. The optimum number of fibers was all 35 training fibers, and the error score was lower than the initial PCA tests. The only difference between the original PCA measurements and the new PCA measurement was the slit position limit imposed on the fibers used to calculate the PCA. The algorithm excluded those that were more than 75 away in slit position. Note that slit position is different than RA and DEC position: it is the order that they are fed into the spectrograph's slit, vertically aligned. It was found that when the PCA method excluded all the fibers in the training set that were more than 75 away in slit position from the fiber of interest, that the RMS error was minimized. A range of limiting values were tested, but 75 was roughly optimal, excluding some variation based on the three types of fiber. The table below shows the key comparisons and test result.

Table 5. RMS error values: optimized by PCA with limited slit position imposed.



35 Fibers Used									6 Fibers Used
	Original Tests				Additional Tests				
	Average 4 Nearest	PCA All 35	SPCA All 35	Average All 35	Average Random 4	PCA 4 Nearest	PCA slit position limited	PCA 2 PC	PCA slit position limited
A	162.8257	156.5091	193.3964	153.9769	170.4389	167.212	156.25	162.6324	162.6394
B	59.0425	62.741	106.0215	65.7772	75.972	59.5575	55.527	66.7359	59.3215
C	47.6431	49.0805	79.8737	51.7094	57.0059	47.5783	41.7175	52.3234	48.3635
simulated	0.017541	0.0063418	0.0055785	0.0060841					

Recall, the current method within SDSS to subtract the sky from fibers of interest was to populate each plate with 35 sky fibers and average the four nearest in RA/DEC and slit position to subtract from each science fiber. The table above shows that using slit position limiting, the PCA method could achieve the same error rates as the current method with only 6 calibration fibers. You will also notice in the table above that tests were done to see if using 2 principal components would solve the slit position problem independently, but the error rate was still lower by limiting the slit position to 75.

Unfortunately, the improvements in RMS error using the slit position limited PCA method did not per-pixel error below the desired rates. Using 35 training fibers, the number of beyond-threshold pixels decreased from 242 to 221 out of the 2015 usable pixels. Using 112 training fibers this could be further reduced to 195 pixels. No scenario tested was able to reduce the error rate completely below the desired threshold.

Conclusions

This study began by framing the problem to be addressed: how to optimize the sky subtraction process in order to minimize error and maximize the number of science fibers in each observation. First, the data was characterized as sparse, homogenous within a scaling factor, and with added white Gaussian noise. This characterization was used to justify the hypothesis that sparse principal component analysis and principal component analysis should outperform the current method of subtraction. The current method of sky subtraction used by SDSS was to take the four closest sky fibers in slit position and RA/DEC position, average them, and subtract them from the fiber of interest.

The first set of results were found through simulation. These results confirmed the hypothesis that for a dataset matching the assumptions listed above, the SPCA had the lowest RMS error. Cross validation of the simulated data was used to select 450 for the number of non-zero variables in the SPCA algorithm.

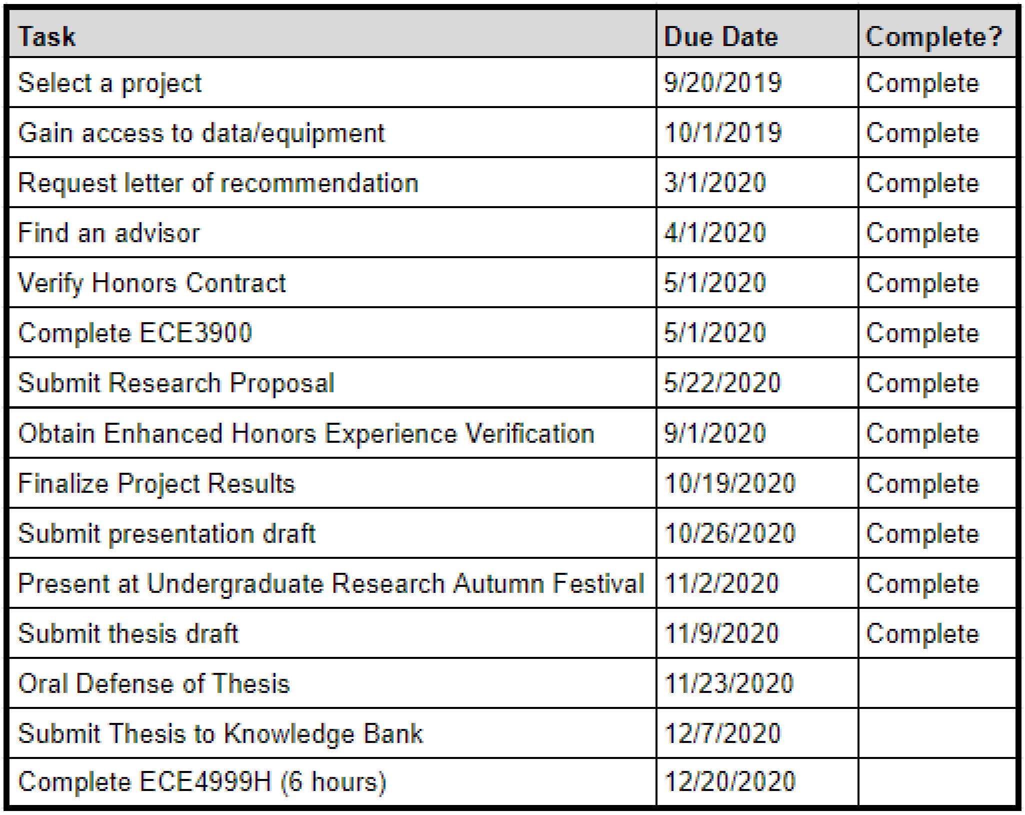
The second set of results were found using measured data. Here, a dataset with an unusually high number of sky fibers was used so that the number of training fibers (sky fibers) could be varied. The percent of variance was measured for the dataset and it was found that more than 99% of variance was explained in the first principal component; therefore, only one principal component was used. Cross validation was used on the measured dataset, but it was less of a clear distinction between exponential and linear sections of the trend; therefore 450 non-zero variables were again used in the measured case. The results showed that between the three methods tested, the averaging of the four nearest fibers produced the best RMS error results. It was further found that for this scenario, approximately 12 percent of the pixels exceeded the threshold of 1 part per 100 photons of error.

The final set of results explored the disparities between the simulated and measured results. It was found that the PCA method yielded less error than the averaging method when the training set was limited. The optimum limiting conditions were determined: use all available training fibers except those more than 75 away in slit position, calculate the PCA for each fiber of interest, then subtract the resulting signal representation. The number of sky fibers in the training set could be reduced to 6 fibers and still maintain the error rates resulting from SDSS's current methods. This would be a 10% increase in the amount of science data collected over the coarse of the telescope.

Documentation

Documentation of this project was shared within the group via a shared GoogleDrive and MicrosoftTeams project page. Documentation of this project included a task breakdown table with due dates, Google Drive meeting notes, Microsoft Team discussion boards, email collaborations, and how-to data access summaries. Meeting notes, discussion boards, and email collaborations which document the acquisition of information pertaining to this project and the stages of its progress can be accessed upon request. A table of the high-level tasks are in Table 6 below.

Table 6. Task breakdown of the capstone requirements and other project milestones.



Task	Due Date	Complete?
Select a project	9/20/2019	Complete
Gain access to data/equipment	10/1/2019	Complete
Request letter of recommendation	3/1/2020	Complete
Find an advisor	4/1/2020	Complete
Verify Honors Contract	5/1/2020	Complete
Complete ECE3900	5/1/2020	Complete
Submit Research Proposal	5/22/2020	Complete
Obtain Enhanced Honors Experience Verification	9/1/2020	Complete
Finalize Project Results	10/19/2020	Complete
Submit presentation draft	10/26/2020	Complete
Present at Undergraduate Research Autumn Festival	11/2/2020	Complete
Submit thesis draft	11/9/2020	Complete
Oral Defense of Thesis	11/23/2020	
Submit Thesis to Knowledge Bank	12/7/2020	
Complete ECE4999H (6 hours)	12/20/2020	

The table shown above documents the high-level task management of the project. It was referenced and updated throughout the life of the project to help team members coordinate and meet deadlines.

To view the data access How-To files, please follow the link below.

https://drive.google.com/drive/folders/1oLwWwSmEydAPMx6SBL8oPNuGfID_GoOZ?usp=sharing

Finally, to view or download the Matlab code referenced here, please follow the link below.

https://drive.google.com/drive/folders/1aI7ol0ioBdUzH7gdI47RWGXWhD2AcfOT?usp=sharing

Implications and Future Work

The results from this project could continue on in a few directions. First, the results found in this study could be re-tested for a variety of observation conditions. For example, it may be that a certain number of fibers is adequate for Halo-type observations, while more fibers are required for YSO-type observations. This futher study will be made much easier by the prototype scripts described and referenced in this project, since they only need to be applied to new datasets in order to compare results.

Second, the sky subtraction method described in the conclusion of this report could be implemented in the SDSS data pipeline for SDSS-V. The method that was found here to be most efficient was PCA with limited slit positions for the training data. If this was implemented in SDSS-V, fewer sky fibers could be used while still maintaining the same level of error.

Of course, the optimized subtractions described here can also be used for the already-collected SDSS-IV data. Although, we cannot increase the number of science fibers after-the-fact, we can get improved error rates by using the PCA with limited slit position method as found in these studies.

Finally, this project may have many less-direct implications in future work. As mentioned, the data access guides created during this project have already been used by fellow SDSS members to streamline the learning curve when accessing sky signals from the data available. Also, some of the observation made in this project would be of interest to those scientists who are not just trying to subtract out the sky signals, but who wish to study the sky signals and characterize it. The characterization of the data, and the modifications to assumption made here could be very useful to a wide variety of interested parties in the future.

Bibliography

[1] *APOGEE Visit Spectra Reduction*, Sloan Digital Sky Survey. Accessed on: Mar. 25, 2020. [Online]. Available :https://www.sdss.org/dr15/irspec/apred/.

[2] *SDSS Voyages*, Sloan Digital Sky Survey. Accessed on: April 16, 2020. [Online]. Available: https://voyages.sdss.org/for-educators/ground-control/my_plate/.

[3] *SDSS Bitmasks*, Sloan Digital Sky Survey. Accessed on: Nov. 15, 2020. [Online]. Available: https://www.sdss.org/dr16/algorithms/bitmasks/#APOGEE_PIXMASK.

[4] Ali, Moshin, *Sparse Kernel Principal Component Analysis*, Aug. 31, 2016. [online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/58939-sparse-kernel-principal-component-analysis.

[5] Zou, Hastie, and Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15, pp. 265-286, 2006.

Appendix

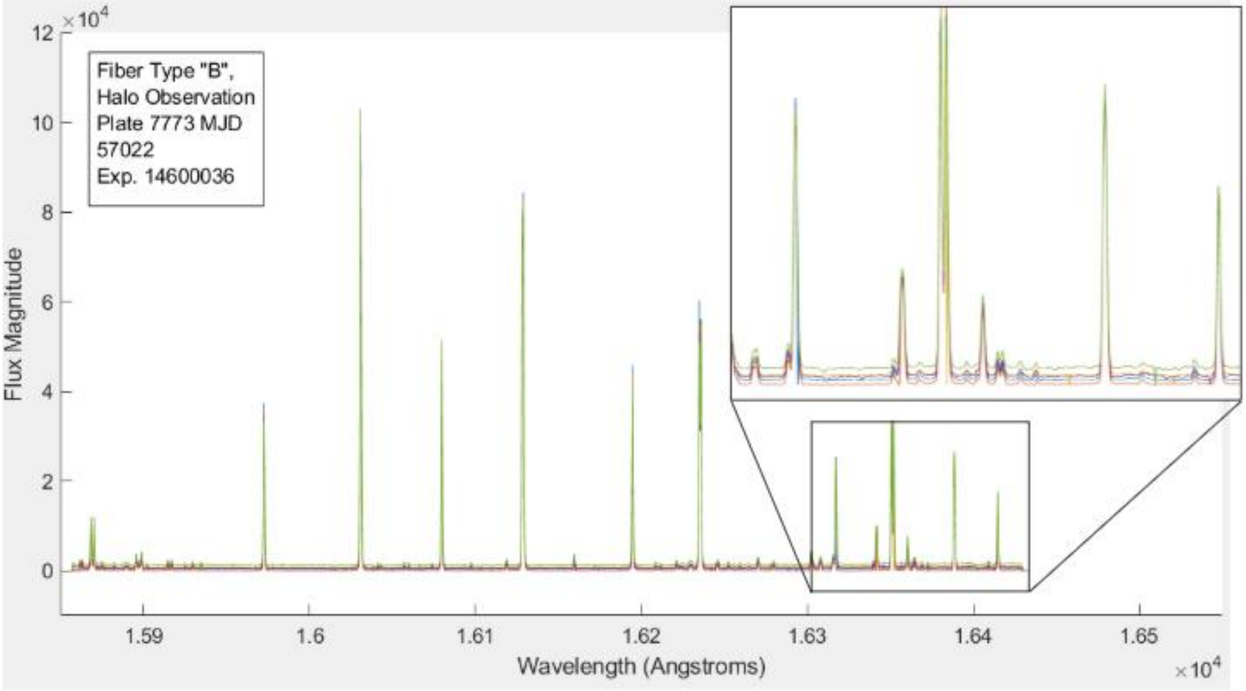


Figure 2. Example spectra of sky fibers [1].

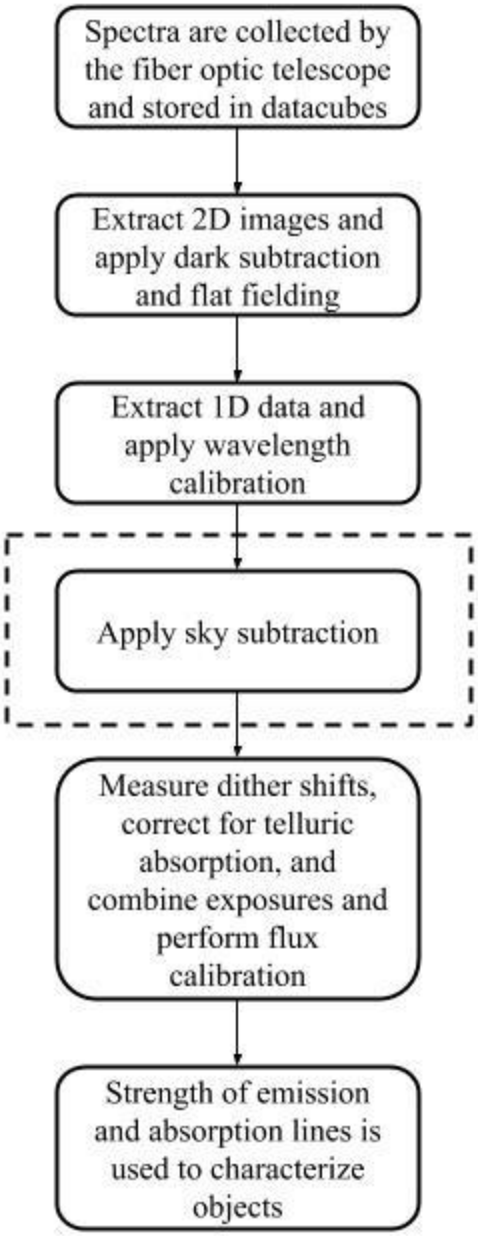

Spectra are collected by the fiber optic telescope and stored in datacubes
Extract 2D images and apply dark subtraction and flat fielding
Extract 1D data and apply wavelength calibration
Apply sky subtraction
Measure dither shifts, correct for telluric absorption, and combine exposures and perform flux calibration
Strength of emission and absorption lines is used to characterize objects

Figure 3. Diagram of the SDSS-IV data pipeline showing the subject of this project in the dotted line.

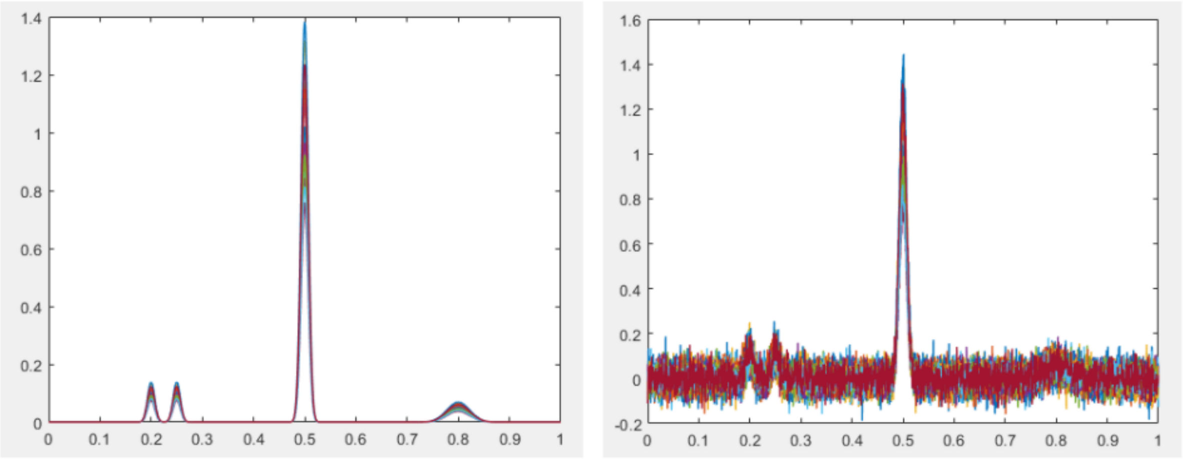


Figure 6. The true sky was simulated as having a large dynamic range with a few features. First, the sky was varied in amplitude (left) to simulate 35 instances of sky, and then white noise was added (right).

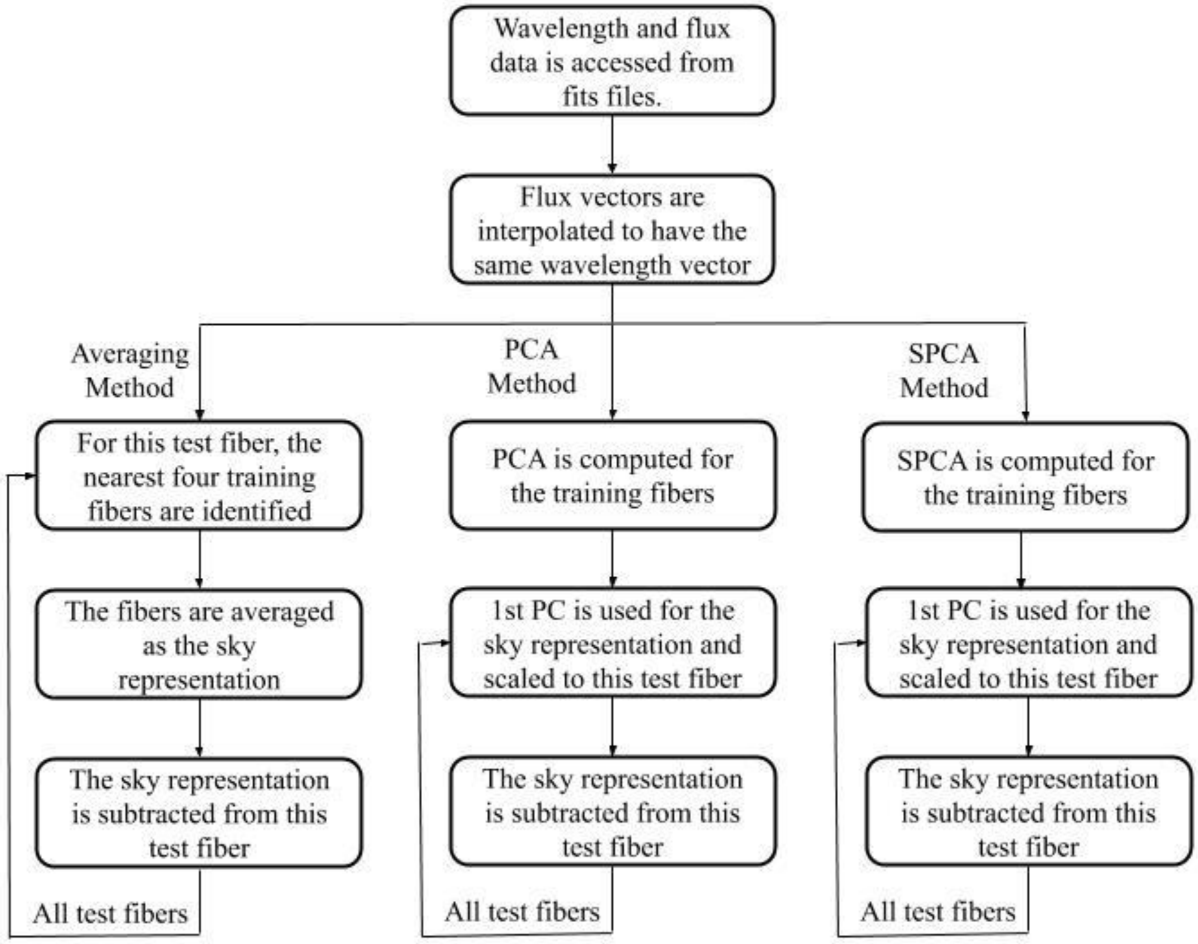
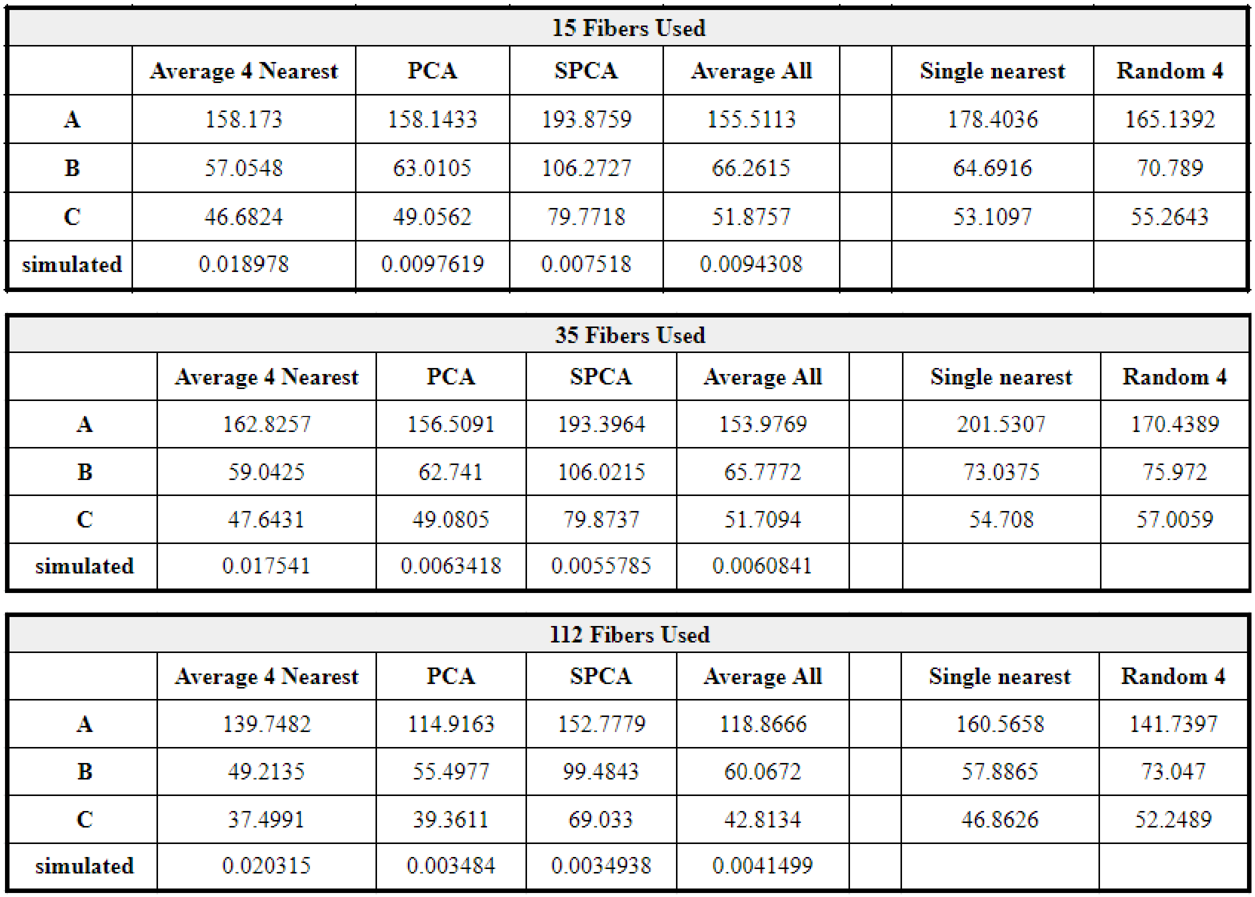

Wavelength and flux data is accessed from fits files.
Flux vectors are interpolated to have the same wavelength vector
Averaging Method
PCA Method
SPCA Method
For this test fiber, the nearest four training fibers are identified
PCA is computed for the training fibers
SPCA is computed for the training fibers
The fibers are averaged as the sky representation
1st PC is used for the sky representation and scaled to this test fiber
1st PC is used for the sky representation and scaled to this test fiber
The sky representation is subtracted from this test fiber
The sky representation is subtracted from this test fiber
The sky representation is subtracted from this test fiber
All test fibers
All test fibers
All test fibers

Figure 10. Diagram of Matlab code to test each sky subtraction method.

Table 3. The RMS error for the simulated results and each fiber type according to method.



15 Fibers Used							
	Average 4 Nearest	**PCA**	**SPCA**	**Average All**		**Single nearest**	**Random 4**
A	158.173	158.1433	193.8759	155.5113		178.4036	165.1392
B	57.0548	63.0105	106.2727	66.2615		64.6916	70.789
C	46.6824	49.0562	79.7718	51.8757		53.1097	55.2643
simulated	0.018978	0.0097619	0.007518	0.0094308			

35 Fibers Used							
	Average 4 Nearest	**PCA**	**SPCA**	**Average All**		**Single nearest**	**Random 4**
A	162.8257	156.5091	193.3964	153.9769		201.5307	170.4389
B	59.0425	62.741	106.0215	65.7772		73.0375	75.972
C	47.6431	49.0805	79.8737	51.7094		54.708	57.0059
simulated	0.017541	0.0063418	0.0055785	0.0060841			

112 Fibers Used							
	Average 4 Nearest	**PCA**	**SPCA**	**Average All**		**Single nearest**	**Random 4**
A	139.7482	114.9163	152.7779	118.8666		160.5658	141.7397
B	49.2135	55.4977	99.4843	60.0672		57.8865	73.047
C	37.4991	39.3611	69.033	42.8134		46.8626	52.2489
simulated	0.020315	0.003484	0.0034938	0.0041499			

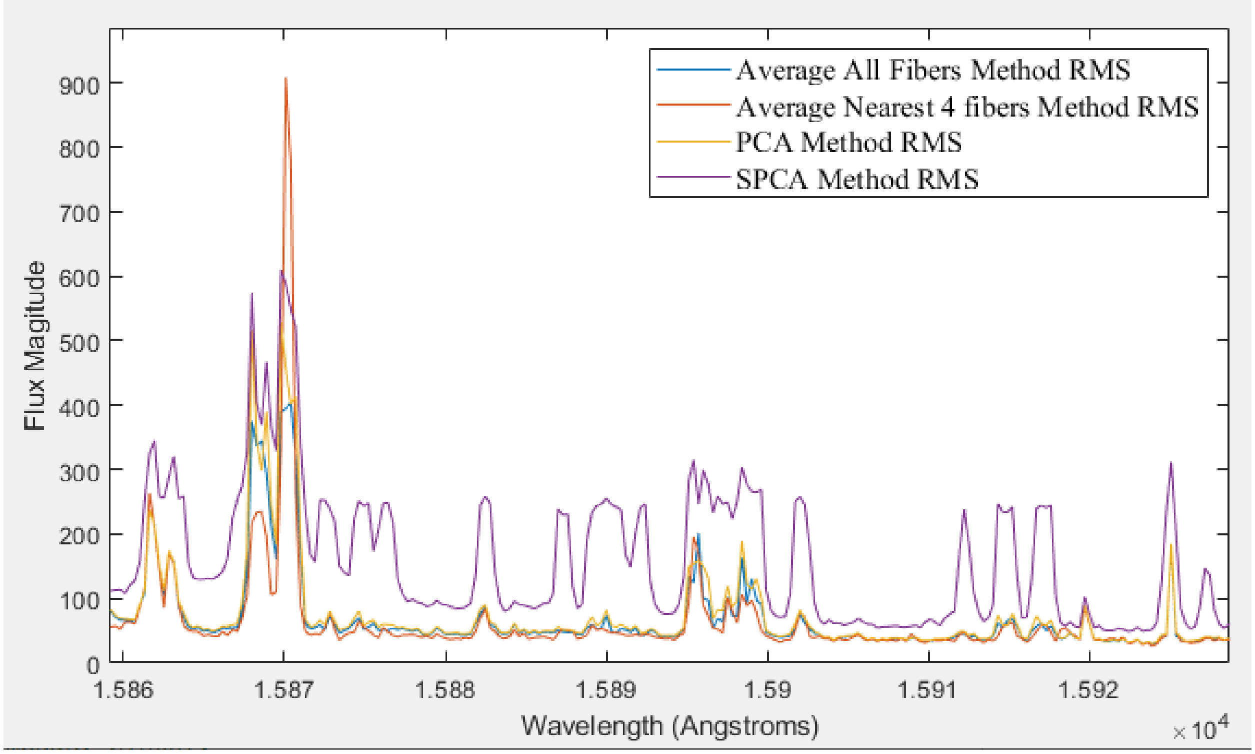


Figure 11. Snapshot of the per-pixel error for different sky subtraction methods.