May 8th, 12:00 AM

Analyzing The Uncertainty of Sloan Digital Sky Survey Data

Kristen Gardner
University of Southern Maine, kristen.gardner@maine.edu

Analyzing The Uncertainty of Sloan Digital Sky Survey Data

Kristen Gardner, Julie Ziffer

Department of Physics, University of Southern Maine, Portland, ME 04104

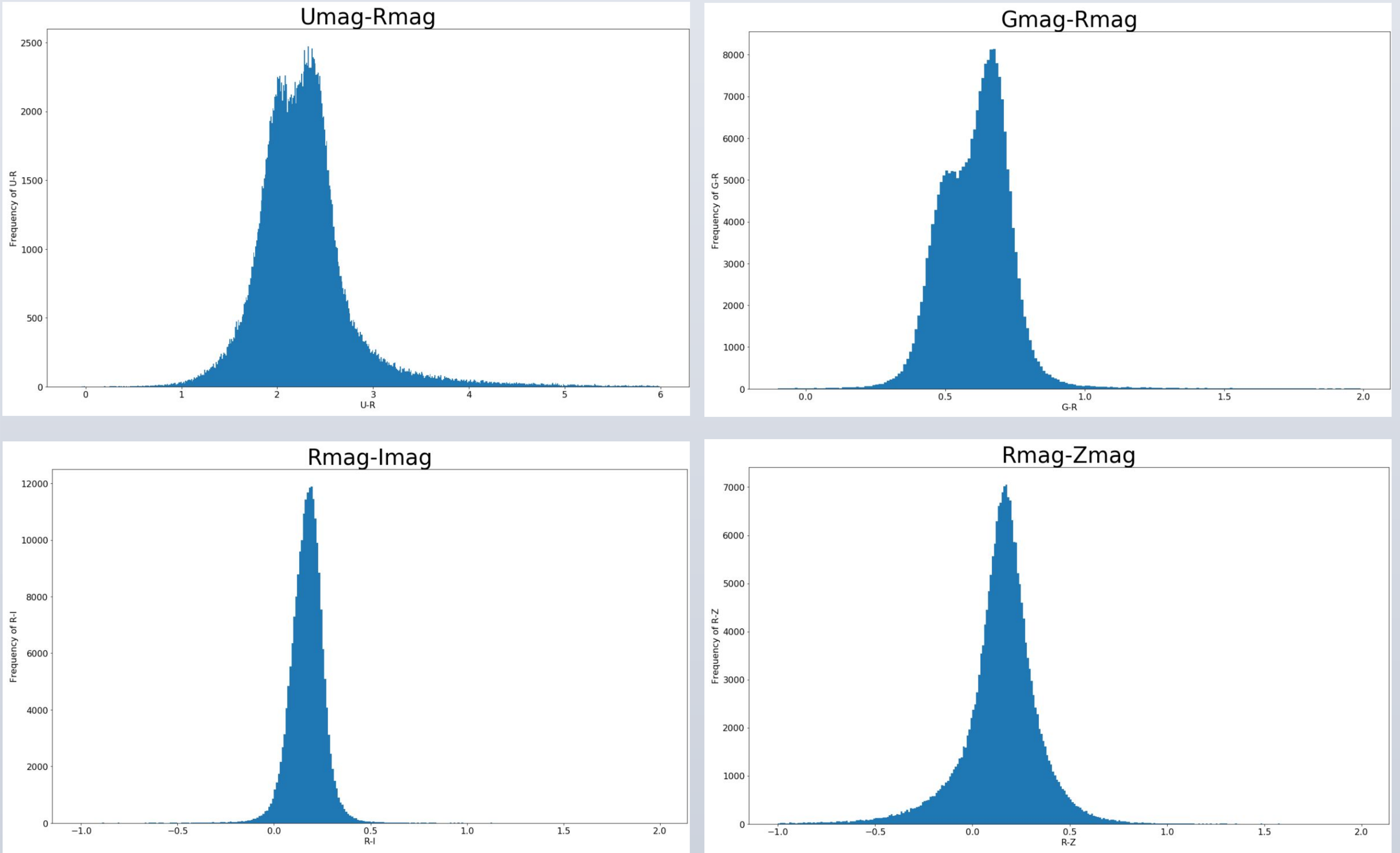
SUPPORTED BY A GENEROUS GRANT FROM MAINE SPACE GRANT CONSORTIUM

Introduction

The Sloan Digital Sky Survey (SDSS) is a spectroscopic all-sky survey. Data collected by the SDSS is used by many researchers. The success of research projects is dependent on the accuracy of the data, and the validity of the uncertainty interval associated with each data point. The purpose of this project is to assess the uncertainty of SDSS data. The data examined in this project is the reflected light of asteroids measured by the SDSS. Python was used to fit gaussian curves onto frequency histograms of the data, and analyze the relationship between the precision of data, and its distribution.

Analysis

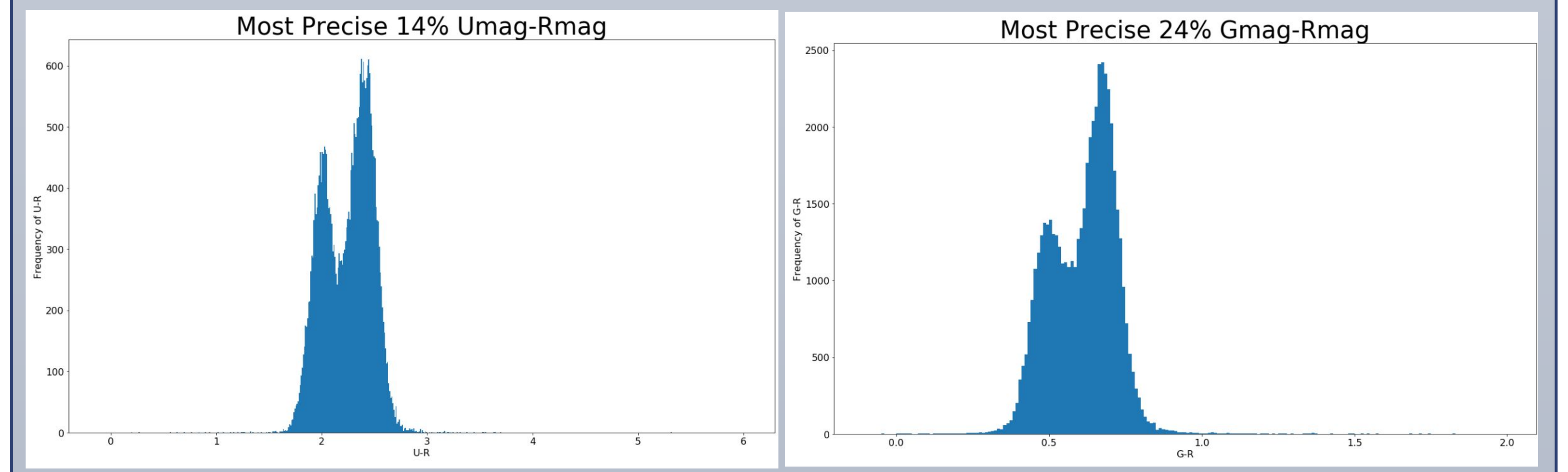
A dataset consisting of the light reflectance of different asteroids, at different wavelengths of light, was chosen for analysis. Because the ratios of the intensity of the different wavelengths being reflected reveals information about the composition of asteroids, the data was examined as a set of four different wavelength intensity ratios. The reflectance ratios appear to be nearly gaussian in their distributions. However, two of the datasets appear significantly misshapen.



Histogram plots of asteroid reflectance data: the frequency of occurrence of each wavelength ratio measured by the SDSS. The histograms each contain over 219000 data points.

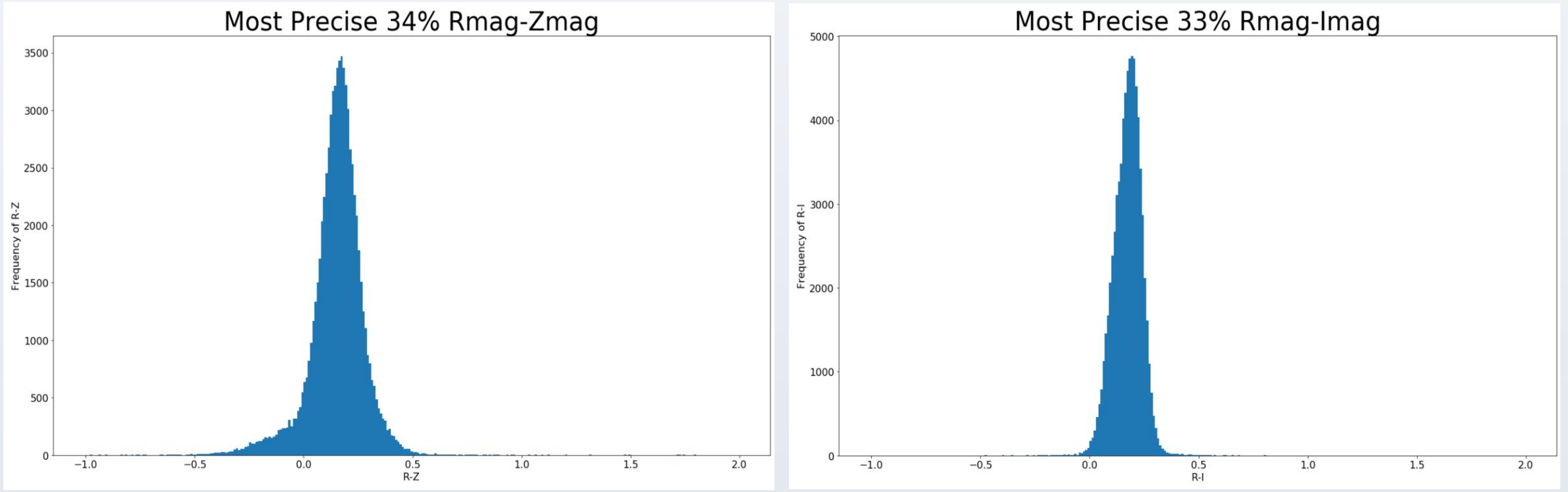
It's not clear how the shape of each gaussian is affected by the uncertainty in the data. To obtain a more accurate representation of this distribution, the datasets were partitioned by the size of the uncertainty in each data point. Many subsets were created with different maximum thresholds of uncertainty allowed in each subset. As the data was reduced, the shape of the distribution changed.

By selecting data with the least uncertainty, a more realistic representation of the data distribution was obtained. The refined datasets also reveal how the asteroids can be categorized into different groups. Two of the wavelength ratios that were studied have data distributions that appear to be wide, misshapen gaussians. As the dataset was refined, the histogram plots showed distributions that appear to be two distinct gaussians.



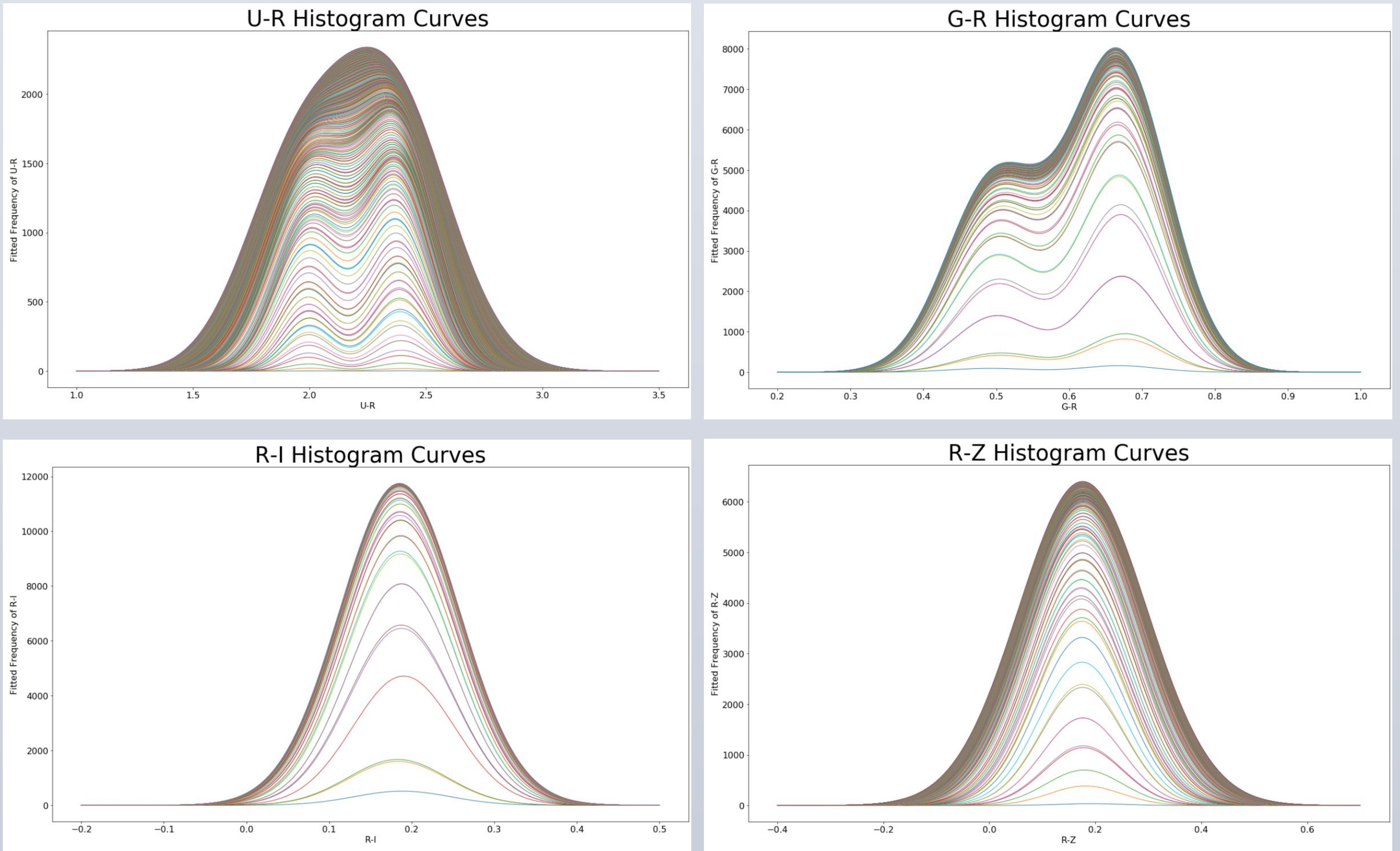
Histograms of the data points with the least uncertainty.

The same refining process shows just one distinct peak in the distributions of the other two ratios. These wavelength ratios were modeled as single gaussian curves.



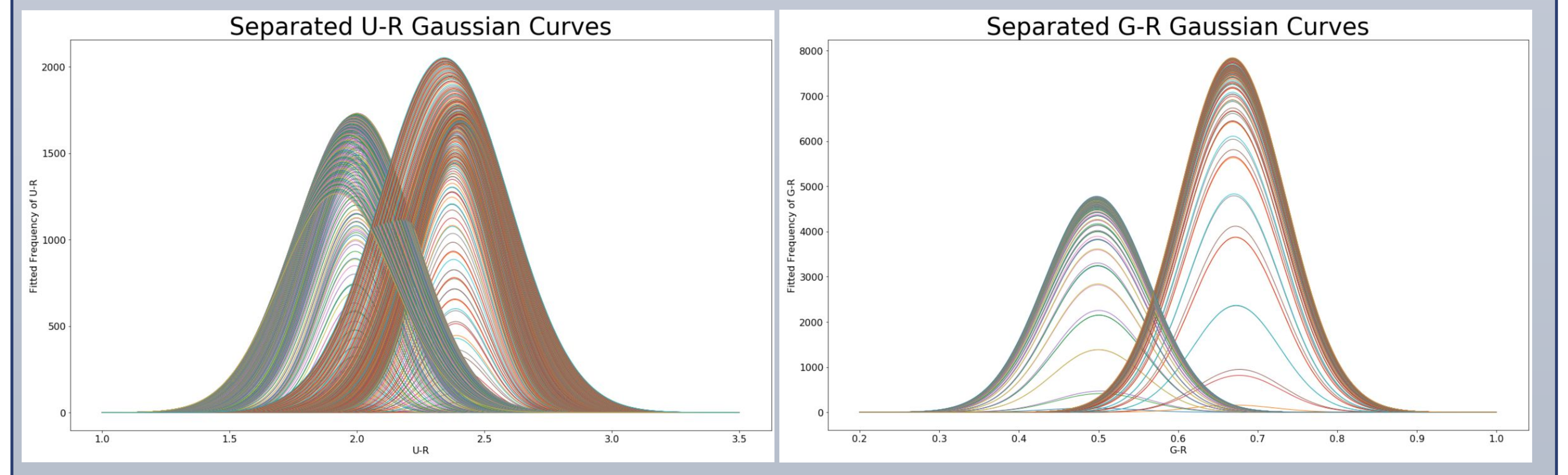
The histograms of each partial dataset were fitted to gaussian curves. The U-R and G-R histograms were each fitted to a sum of gaussian curves.

Each set of curves allow trends in the data distributions to be identified. The sets of curves illustrate the significant change in the U-R and G-R distributions as the allowed error is reduced. The two kinds of asteroids are more distinguishable.



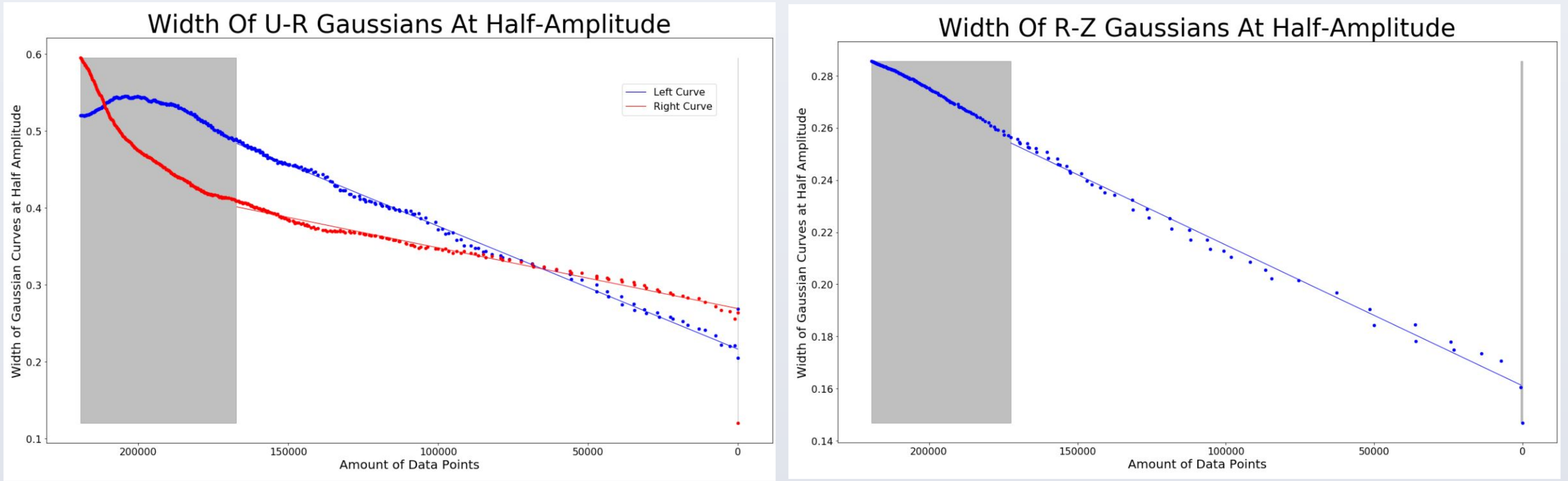
Gaussian curves fitted onto histogram plots.

The coefficients describing the summed gaussians that fit U-R and G-R were used to plot discrete gaussian curves, representing the groups of asteroids. Two separate mean values can be clearly recognized, as well as a standard distribution of data for each group. The changes in each distribution, as the allowed uncertainty is reduced, can be calculated using the coefficients of these gaussian functions.



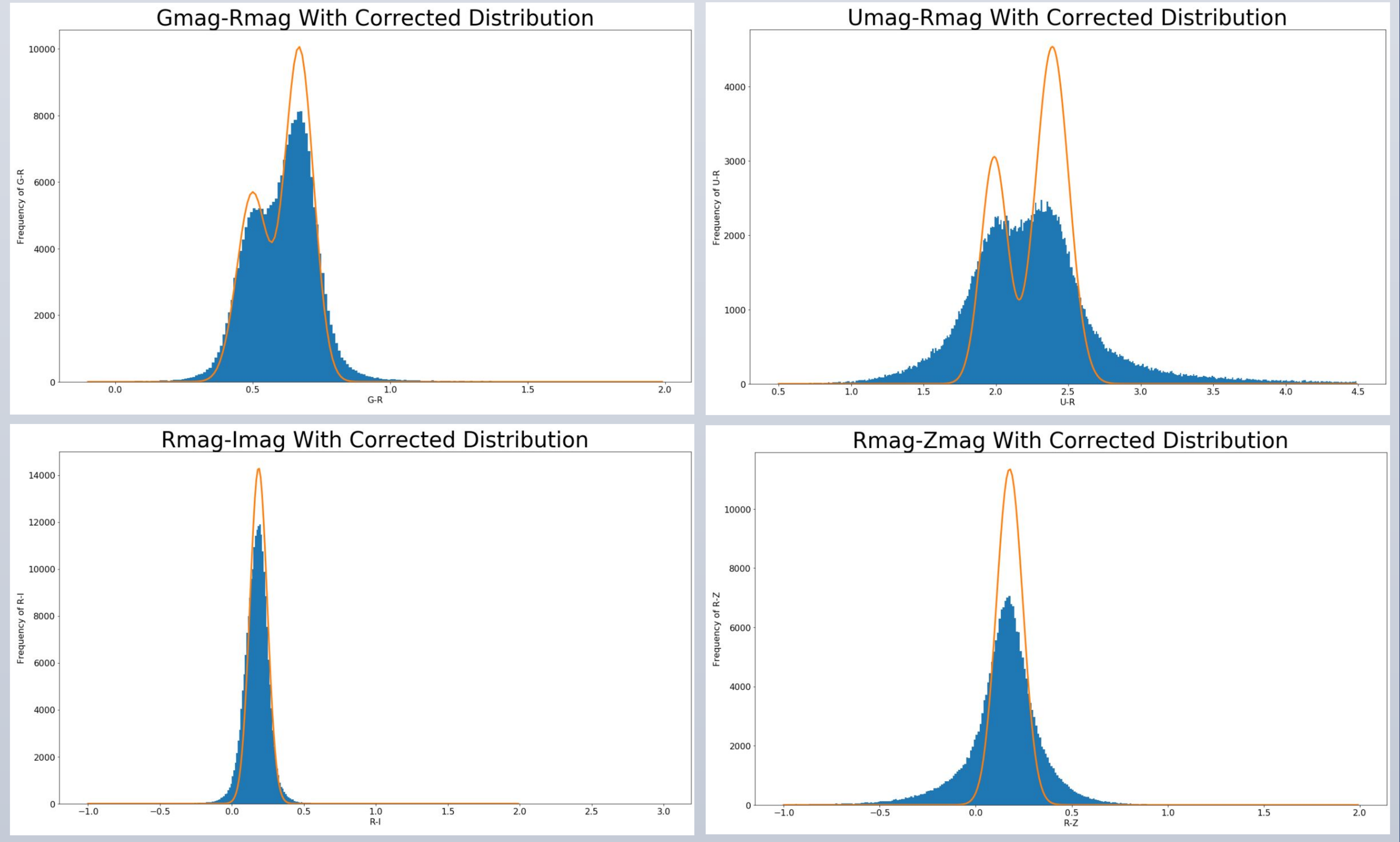
Separated gaussian curves representing two different kinds of asteroids.

As the data is reduced, the width of the gaussian fitted-curve at half-amplitude decreases linearly. This relationship was used to extrapolate the shape of the curves as the size of the dataset approaches zero. The full width at half amplitude was used to calculate the standard deviation of such a curve.



Results

The calculated standard deviation for a dataset with no uncertainty was used to model what the original datasets might look like if each measurement had no uncertainty. Each gaussian appears much more narrow and tall. Sets that contain two gaussian distributions appear more discrete, with more clearly defined median values.



Histograms of each wavelength reflectance ratio, plotted against the fitted curve of a theoretical distribution that has no uncertainty in its data points.

Future Research

It is important to investigate other variables that could affect the distribution of data. Asteroids with very high or low reflectance are more difficult to measure with accuracy. This method of refining data sets could cause the population to be misrepresented if a disproportionate amount of asteroids of a particular reflectance are being omitted from the model.

The results from this analysis can be used to generate a simulated data set. Given the newly presumed distribution of data values, and the distribution of uncertainties given in the SDSS data, a simulation can be made that is intended to replicate the SDSS data distributions. If the distributions are significantly different, then the uncertainty associated with the SDSS data points don't entirely account for the variation in the data.