
Photometric Redshifts and Random Forests

Keywords: methods: statistical, galaxies: distances and redshifts, evolution, photometry

BACKGROUND: The identification of redshifts spectroscopically is a time, labor, and resource intensive process. Alternatively, photometric redshifts based on broadband photometry and other available data can provide a viable low-cost alternative when applied to large samples of galaxies. They have been successfully used to measure cosmological parameters¹, to study galaxy luminosity functions to great depth^{2,3,4}, and to identify high-redshift galaxies⁵ and clusters⁶. Moreover, photometric redshifts will begin to play a larger role in the era of large photometric survey telescopes like the Large Synoptic Survey Telescope (LSST). As a result, the astronomical community is in need of fast, accurate, portable and freely available photometric redshift software that is free of the problems associated with existing spectral energy distribution template fitting codes (eg. the need for potentially incomplete synthetic galaxy models). Nevertheless, there are no examples of photometric redshifts produced with 21st century statistical techniques yet in the literature, let alone those produced with a free publicly available code.

The Random Forests⁷ (RF) algorithm is the current state-of-the-art in machine learning. RF has yet to be adopted by astronomers despite the pressing need for efficient, accurate classifications in countless contexts, and its popularization will be beneficial to the community. The algorithm has the following desirable qualities: its accuracy is unparalleled compared to other common algorithms, it runs quickly on large databases, like all tree-based classifiers it can handle thousands of (possibly uninformative) predictors without the need for variable selection, it can identify the most important predictors and estimate their importance, it can handle missing data and maintains its accuracy even if a significant fraction of the data is not available, it generates proximities which can be fed to other classification algorithms for cross-validation and unsupervised learning, it provides a way to experimentally detect variable interactions, and it can be used for nonlinear regression. In addition, an excellent implementation of the RF algorithm is available as part of the freely available *R* statistical analysis package⁸, the standard statistical analysis tool of statisticians the world over.

RESEARCH: I propose to use the RF algorithm to compute photometric redshifts for DEEP2 galaxies that lie in the Extended Groth Strip (EGS) for which spectroscopic redshift identification has either failed or been rendered impossible by faintness. I will use DEEP2 galaxies with measured spectroscopic redshifts and panchromatic EGS data as a training set to build an RF classifier with the ability compute photometric redshifts for other EGS galaxies with no available spectra. As a proof of concept, I have already used RF to compute photometric redshifts based on the publicly available DEEP2 Data Release 1 redshift catalog. The catalog contains 3850 galaxies, each with a high-quality spectroscopically determined redshift and B, R, and I apparent magnitudes. I find that the standard deviation in my redshift residuals is less than 0.2. The fact that RF returned such a small residual over a wide-range of redshifts with only

three predictors indicates that its performance will be unmatched once the rest of the panchromatic EGS photometry is included. Beyond the classifications, the decision trees output by RF can be used to study the dependence of broadband galaxy colors and luminosities on redshift – a relationship of fundamental importance in galaxy evolution.

Research questions:

- **Utility of Random Forests:** How do photometric redshifts produced by RF compare to those in the literature? How does the performance compare (which code is faster)?
- **Galaxy Evolution:** How do broadband colors and luminosities change as a function of redshift, and what does this mean for galaxy evolution?

Hypotheses:

- **Utility of Random Forests:** The photometric redshifts produced by RF will be superior to those produced with other methods in terms of accuracy and computational cost.
- **Galaxy Evolution:** The results of this analysis will confirm published galactic luminosity functions, though I also expect to identify a handful of previously unobserved relations.

Methods:

I will use the randomForest package freely available with the R statistics software environment.

- **Utility of Random Forests:** I will duplicate photometric redshifts produced from publicly available catalogs and compare the standard deviation of RF redshifts residuals with other published results. I will also measure the computational performance of the algorithm.
- **Galaxy Evolution:** I will create galaxy luminosities functions for many redshifts and interpret the decision trees produced by the RF algorithm.

PROJECT DELIVERABLES: I will publish the results of my work in peer-reviewed journals and present it at a meeting of the American Astronomical Society. Furthermore, I will make my astronomy specific R code and documentation publicly available from my website so that other astronomers interested in using will have access to it. I will also strive to integrate this project into the K-12 educational programs of UCSC's Center for Adaptive Optics. I will identify the results of this research that are important for the Milky Way's formation and evolution and I will work with the campus public relations staff to ensure that these implications are effectively communicated to the campus, the city of Santa Cruz, the state of California, and nation as a whole. Finally, there are opportunities for undergraduate participation in this project, and I plan to advertise those opportunities at the UCSC Astronomy & Astrophysics Department's annual undergraduate research fair.

CONCLUSION: I plan to develop and publicly release an astronomy-specific implementation of the Random Forests algorithm optimized for use with photometric redshifts but flexible enough for use in other astronomical situations. I will use it to generate photometric redshifts for DEEP2 galaxies with EGS photometry, and will use the relationships generated in the process to study the redshift evolution of broadband galaxy colors and luminosities.

- 1 Loh, E.D., & Spillar, E.J. 1986, ApJ, 303, 154
- 2 Subbarao, M.U., Connolly, A.J., Szalay, A.S., & Koo, D.C. 1996, AJ, 112, 929
- 3 Liu, C.T., Green, R.F., Hall, P.B., & Osmer, P.S. 1998, AJ, 116, 1082
- 4 Gwyn, S.D.J., & Hartwick, F.D.A. 1996, ApJL, 468, L77
- 5 Steidel, C.C., Giavalisco, M., Dickinson, M., & Adelberger, K.L. 1996, AJ, 112, 352
- 6 Giallongo, E., D'Odorico, S., Fontana, A., Cristiani, S., Egami, E., Hu, E., & McMahon, R.G. 1998, AJ, 115, 2169
- 7 Breiman, L. 2001, Machine Learning, 45, 5
- 8 <http://www.r-project.org/>