Draft version May 1, 2025
Typeset using LaTeX **manuscript** style in AASTeX7

Predicting Hubble Residuals Using Host Galaxy Properties

Mykola Chernyashevskyy[1]

[1] *University of Pittsburgh, Physics and astronomy Department, 3941 O'Hara St, Pittsburgh, PA 15260. United States of America*

1. INTRODUCTION

Observations of Type Ia Supernova (SNe Ia) have yielded constraints relating to the expansion of the universe (Ex. H_0) and dark energy. Hubble residuals are the difference between observed distance moduli of SNe Ia and those predicted by cosmological models. Reducing the scatter in these residuals helps to decrease the uncertainty in cosmological parameters.

A phenomenon known as "mass step" shows that Type Ia supernovae occurring in host galaxies with stellar masses above $10^{10}, M_{\odot}$ tend to appear systematically brighter—after light-curve standardization—than those in lower-mass galaxies Scolnic et al. (2018). This suggests that residual scatter in standardized luminosities may be influenced by host galaxy properties, motivating efforts to predict or correct Hubble residuals using these features.

In this project I investigate eight separate galaxy properties in order to probe if they have a relationship with their host SNe Ia residuals using kNN, XGBoost, and RandomForest algorithms.

2. DATA

Galaxy Data was collected using the Dark Energy Spectroscopic instrument (DESI) at the Mayall 4-meter telescope at Kitt Peak National Observatory. Host galaxy data was sourced from the DR1 DESI data release DESI Collaboration (2025).

The Dark Energy Survey (DES) provides the SNe Ia data used in this research. Data was collected using the Dark Energy Camera (DECam) mounted on the Blanco 4-meter telescope at Cerro Tololo Inter-American Observatory Sánchez et al. (2024).

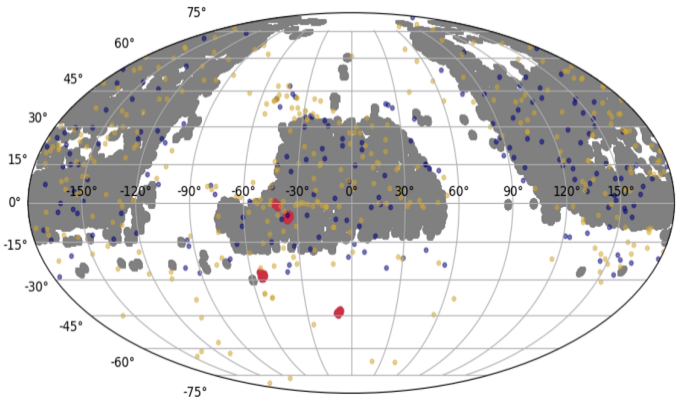

Figure 1. Mollweide projection of DESI galaxies (grey), and subsets of the DES 5YR SNIa data release. Yellow is the low-z data set, blue is the foundation data set, and red is the DES-SN5YR data set.

Galaxies were matched together within 5 half light radii of each other as well as within 0.05 difference in redshift Lunnan et al. (2016); Gupta et al. (2018). After matching, the final dataset consisted of 67 matched supernova–host galaxy pairs. Hubble residuals were computed using Flat Labmda CDM cosmology model with $H_0 = 70 \mathrm{km\,s^{-1}\ Mpc^{-1}}$ and $\Omega_m = 0.315$.

3. ANALYSIS METHODS

3.1. *Spearman Correlation Coefficient*

Spearman's rank correlation coefficient (ρ) is a measure of the strength and direction of a monotonic relationship between two sets of date. Spearman's correlation assesses how well the rank of two sets of data agree. This method is robust to outliers and can characterize non linear relationships.

$$\rho\left[\mathrm{R}[X], \mathrm{R}[Y]\right] = \frac{\mathrm{cov}\left[\mathrm{R}[X], \mathrm{R}[Y]\right]}{\sigma_{\mathrm{R}[X]}\, \sigma_{\mathrm{R}[Y]}}$$

where R[X], FR[Y] are the ranked data sets and "cov" is the covariance operator Laerd Statistics (2015).

3.2. *kNN Regression*

The k-Nearest Neighbors (KNN) algorithm is a learning method used for both classification and regression. To make a prediction, KNN calculates the distance between a test data point and all training points, then selects the k closest neighbors. For classification, it assigns the class most frequently represented among those neighbors. For regression, it returns the average value of the k nearest training points. There are various styles of kNN classification: Euclidean, Manhattan, Minkowski, or Hamming distance Kumar (2020).

3.3. *Random Forest Regression*

Random Forest is a machine learning algorithm that constructs an ensemble of decision trees where each tree is trained on a random subset of bootstrapped data. The model aggregates predictions from all trees to produce a final result. In layman's terms a decision tree is a model that learns any relationships between the data and the values we want to predict. The decision tree forms a flow-chart structure, calculating the best questions to ask in order to make the most accurate estimates possible Koehrsen (2018).

3.4. *XGBoost*

XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm that, like Random Forest, builds an ensemble of decision trees. However, XGBoost is different because it constructs trees sequentially using gradient boosting. Each new tree is trained to reduce the loss of the previous model. A loss function is defined that measures how far the trained model is from the true value. The gradient of this function is computed and is used as a guide to find the direction of smallest loss. XGBoost also offers many hyperparameter tuning options to further optimize performance.Sonawane (2019).

3.5. *k-Fold Cross Validation*

k-Fold Cross-Validation is a resampling technique used to evaluate the performance of a machine learning model on limited data. The dataset is divided into k equally sized folds; the model is trained on $k-1$ folds and tested on the remaining fold, repeating this process k times. Brownlee (2018).

4. RESULTS

The Spearman Correlation test was run between the residuals and: Stellar velocity dispersion, logarithm of stellar mass, the 4000 Angstrom break strengh (DN4000), star formation rate (SFR), SDSS redband color magnitude, log of the specific star formation rate, stellar mass and the Sloan Digital Sky Survey SDSS g-r color index (COLOR G-R).

The parameters that exhibited Spearman correlation with the lowest p-values were DN4000 (p = 0.0592, spearman ρ -0.2317), SFR (p = 0.09966, spearman ρ = 0.2029), and COLOR G-R (p = 0.02153, spearman ρ = -0.2804). Data suggests some weak positive correlations of residuals with SFR and a weak negative correlation with DN4000 and COLOR G-R.

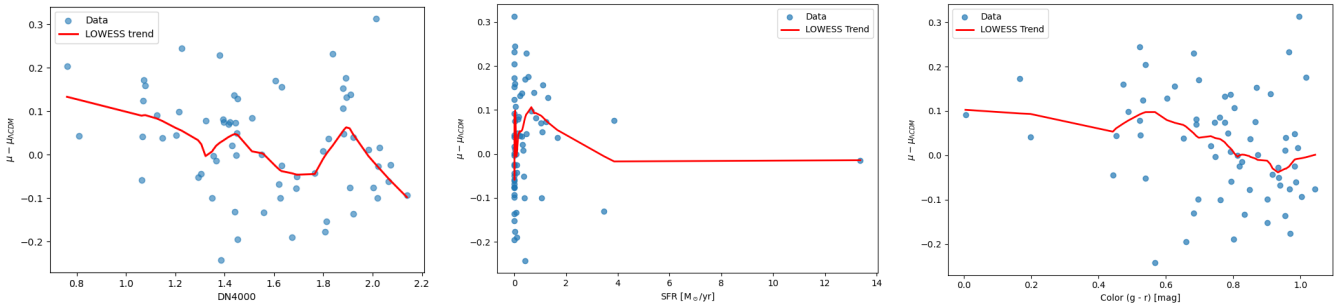

Figure 2. Trhends of Hubble residuals with promoising host galaxy properties. LOWESS curves included to more easily observe trends.

DN4000, SFR, and COLOR G-R were used to train kNN, RandomForest, and XGBoost. Normalized Median Absolute Deviation (NMAD) of the residuals was computed to be 0.1168. Outler cutoffs were determned to be +/- 0.3 mag based on information in Scolnic et al. (2018). NMAD was also used as a performance metric of the ML algorithms:

- k-Nearest Neighbors (kNN) - NMAD: 0.1012, Outlier Rate: 2.99%
- Random Forest - NMAD: 0.1187, Outlier Rate: 4.48%
- XGBoost - NMAD: 0.1043, Outlier Rate: 4.48%

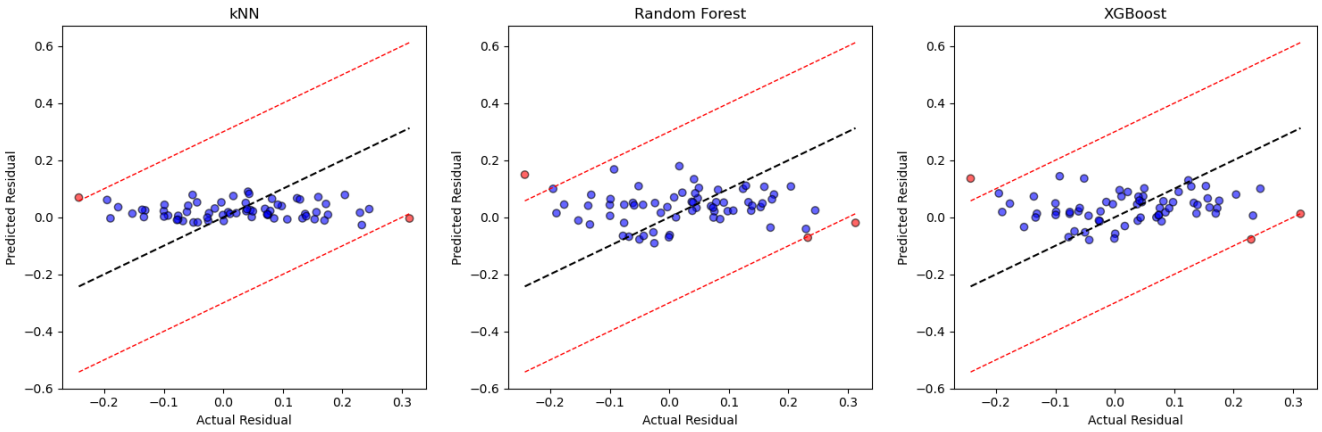

Figure 3. Predicted vs. Actual Hubble residuals of the kNN, Random Forest, and XGBoost methods. Red lines define the outlier limits, solid black line is the line of 1:1 correlation. Optimized hyperparameter for kNN are 25 neighbors, and weighed distance. For random forest, 142 estimators. For XGBoost, 11 estimators and 13 max depth.

5. DISCUSSION AND CONCLUSION

Outlier rates across all models are low: between 3% and 5%, which means models are demonstrating promising predictive power. Random forest does not improve NMAD over pure residual data and thus does not predict or improve residuals from galaxy properties.

Using the observed scatter in our residuals: $\sigma_{\text{total}} = 0.1168$ mag, and looking at intrinsic supernovae scatter: $\sigma_{\text{intrinsic}} \approx 0.10$ mag (e.g., Scolnic et al. (2018)), we can estimate the contribution from host galaxy properties as:

$$\sigma_{\text{galaxy}} = \sqrt{\sigma^2_{\text{total}} - \sigma^2_{\text{intrinsic}}} = \sqrt{(0.1168)^2 - (0.10)^2} \approx 0.06\,\text{mag}$$

This suggests that the host galaxy properties used in our models (DN4000, SFR, and COLOR G–R) account for approximately 0.06 mag variation in the Hubble residuals.

REFERENCES

Brownlee, J. 2018, A Gentle Introduction to k-Fold Cross-Validation, https://machinelearningmastery.com/k-fold-cross-validation/

DESI Collaboration. 2025, Dark Energy Spectroscopic Instrument (DESI), https://www.desi.lbl.gov/

Gupta, R. R., Sako, M., Bassett, B., & et al. 2018, Astronomy & Astrophysics, 617, A73, doi: 10.1051/0004-6361/201832644

Koehrsen, W. 2018, Random Forest: Simple Explanation, https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

Kumar, V. 2020, K-Nearest Neighbor, https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4

Laerd Statistics. 2015, Spearman's Rank-Order Correlation - A Statistical Guide, https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Lunnan, R., Chornock, R., Berger, E., & et al. 2016, The Astrophysical Journal, 817, 144, doi: 10.3847/0004-637X/817/2/144

Scolnic, D. M., Jones, D. O., Rest, A., et al. 2018, A&A, 852, L3, doi: 10.1051/0004-6361/201731425

Sonawane, P. 2019, XGBoost: How Does This Work?, https://medium.com/@prathameshsonawane/xgboost-how-does-this-work-e1cae7c5b6cb

Sánchez, B. O., Brout, D., Vincenzi, M., et al. 2024, The Astrophysical Journal, 975, 5, doi: 10.3847/1538-4357/ad739a