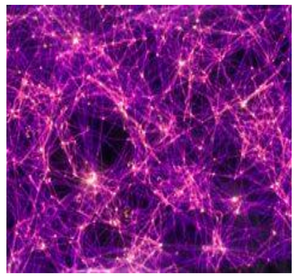
Predicting BCG Mass from Brightness and Shape

By: Megha Raavicharla, Athena Dai, Bin Zheng

Project Supervisor: Peter Freeman

INTRODUCTION

Biggest Cluster Galaxies (BCGs), located at the centers of galaxy clusters, are important objects of study in cosmology because they distort spacetime sufficiently to affect observations of all the galaxies surrounding them in space. The larger their masses, the more distortion they cause. **In this project, our goal is to determine the relationship between brightness and shapes of BCGs and their mass.** Finding a direct relationship would allow us to skip intensive computations that have previously been used to estimate BCG masses.



DATA

Predictors: We analyze data of 390 BCGs from the Spectroscopic Identification of eROSITA Sources program (SPIDERS; Clerc et al. 2020). The processed data contains 66 predictor variables including measurements of brightness and shape in different wavelength bands (dubbed `g`, `r`, `i`) and models of how light varies across the BCGs (dubbed `S`, `V`, `SX`). The variables are grouped as follows:

Shape (30 variables)	Brightness (21 variables)	Others (15 variables)
• Radial extent (`RE`) • Shape parameter (`N`) • Roundness (`AR`)	• Spectral fit metric (`CHI2NU`) • Logarithmic brightness (`MAG`)	• Redshift (`CLUZSPEC`) • Celestial longitude and latitude (`RA, DEC`) • Rotation (`PA`)

Response: the log-base-10 mass of BCGs (`logMass`).

Exploratory Data Analysis: We perform log transformations on predictor variables related to `CHI2NU` and `RE`, which we found yielded better predictions and reduced skewness. We also find an overall pattern in the data where many of the predictors, when grouped by a specific property (e.g. `MAG`) across different models and wavelengths were all heavily linearly correlated. This suggests that many of the predictor variables may contain largely redundant information. (Indeed, when we apply PCA, we find that 35 principal components are sufficient to explain 95% of the cumulative variance in the data.)

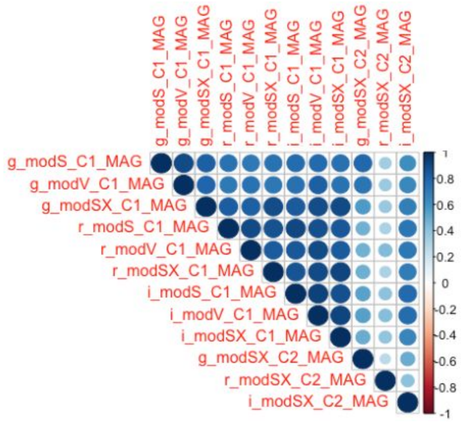


Figure 1: Correlation plot for magnitudes measured for different combinations of wavelength bands and shape models

METHODS

- We train our model on 80% of the data and test on 20% of the data.
- The predictor variables exhibit multicollinearity, but as our project goal is prediction and not inference, we do not remove those variables with high variance inflation factors.
- We apply several models of statistical learning to the data: Linear Regression, Lasso Regression, and Backward Stepwise Selection, Random Forest and K-nearest neighbors

ANALYSIS AND RESULTS

1) Out of all the methods we try, **<u>lasso regression</u>** performs best in predicting the estimated mass of the BCGs in terms of test-set MSEs. In lasso regression, variable selection is performed by using a penalty that shrinks some predictor coefficients to 0. A tuning parameter, lamda, controls the amount of shrinkage.

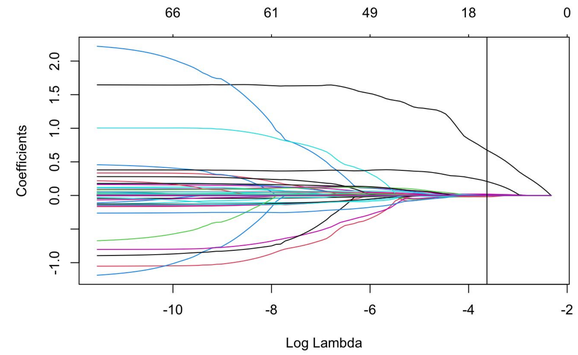


Figure 2: The log of lambda plotted against the coefficients of the predictor variables. The optimal value for log lambda is -3.628 and 11 variables are retained.

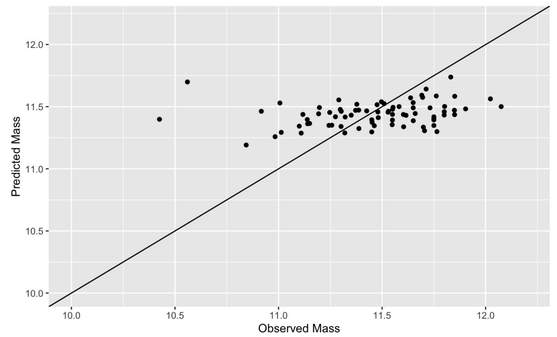


Figure 3: The Predicted log(Mass) vs. Observed log(Mass) for the best the lasso regression model, which was the best model that we obtained

2) **<u>Random Forest</u>** is the second best overall, in predicting the estimated mass of the BCGs in terms of test-set MSEs. In random forest, a group of individual decision trees are used to determine the model's prediction based on the average prediction of the individual trees.

Predictors	Increase in MSE
`CLUZSPEC`	0.035
`g_modSX_C1_MAG`	0.012
`i_modSX_C1_MAG`	0.008
`i_modV_C1_RE`	0.006
`r_modV_C1_MAG`	0.006

Table 1: Top 5 most important variables determined by random forest. Larger increases in MSE means the variable is more important.

Model	Test-Set MSE
Lasso Regression	0.095
Random Forest	0.100
Backward Stepwise Selection	0.103
Linear Regression	0.113
K-nearest neighbors	0.124

Table 2: This table contains the test-set MSE's for statistical models analyzed. Lasso regression had the lowest test-set MSE, followed by the random forest model.

CONCLUSION

Overall, we identify a linear relationship between the brightness and shape measurements of BCGs and its mass. However, the relationship is very weak and even our best model (lasso regression) does not effectively explain the data. Further analysis with more flexible nonlinear models performs at around the same level or worse in terms of prediction ability. Our second best model, random forest, determines that redshift (i.e., the distance from the Earth to the BCG) is by far the most important variables in predicting BCG mass. For the future, it may be beneficial to expand the dataset to include more observations and other predictor variables to identify if other properties of BCG can even more effectively explain BCG mass.

REFERENCES

N Clerc, C C Kirkpatrick, A Finoguenov, R Capasso, J Comparat, S Damsted, K Furnell, A E Kukkola, J Ider Chitham, A Merloni, M Salvato, A Gueguen, T Dwelly, C Collins, A Saro, G Erfanianfar, D P Schneider, J Brownstein, G A Mamon, N Padilla, E Jullo, D Bizyaev, SPIDERS: overview of the X-ray galaxy cluster follow-up and the final spectroscopic data release, *Monthly Notices of the Royal Astronomical Society*, Volume 497, Issue 3, September 2020, Pages 3976–3992, https://doi.org/10.1093/mnras/staa2066