Final: Cosmic Voids Illuminate Galaxy Properties

Christina Kreisch
Department of Astrophysical Sciences, Princeton University
ckreisch@astro.princeton.edu

Abstract

This work investigates the predictive power of cosmic voids on galaxy properties. I focus on predicting galaxy mass based solely on void properties. I find that the Random Forest yields the best predictive power, with a mean squared error of $\approx 5 \times 10^{-6}$. Linear models have a more difficult time predicting the relationship between voids and galaxy mass. To alleviate this difficulty, I also investigate transforming the void properties into a 2nd order polynomial and a power law. Both of these transformations help with the predictive power of the linear models, but the random forest still provides the best performance. The relationship between galaxy mass and void properties is more complex than a 2nd order polynomial or a power law. I find that the void's density contrast, ellipticity, and radius are the most predictive features of galaxy mass.

1 Introduction & Related Work

Cosmic voids are the most underdense, largest regions of our universe. They lie within the dense components of our universe that form a pattern called the cosmic web [3]. The empty regions in the cosmic web are the voids. Despite being relatively empty, cosmic voids still contain a few galaxies. However, since the voids are mostly empty the environment of these galaxies is different than for galaxies found outside of voids. This different environment impacts the evolution of these galaxies [7], making it interesting to study how the properties of galaxies within voids are different than for those outside of voids. Previous work, such as that in [11], has shown with observations that halo[1] property distributions in voids are different than for halos outside of voids. Investigating the connection between voids and galaxies is still an active area of research, especially using simulations since not much theoretical work has been done yet. For example, [5] (in prep.) are currently studying the relationship between voids and black holes, as well as galaxy properties based on the distance of the galaxy from the center of the void, using a different simulation that what I use here. The literature involving the application of machine learning to the relationship between voids and their galaxies is even sparser[2]. This is not surprising as voids themselves are a relatively new object to study in astrophysics, and so there is a substantial amount of work still to be done in the field regarding them. Nevertheless, machine learning has already played a significant role in classification of galaxies, for example, so the future seems promising [6].

Understanding the relationship between voids and the galaxies that reside within them will have substantial impacts on designing future surveys to observe galaxies. Observing galaxy properties like star formation rate can be extremely difficult. Voids, however, are easier to observe since they are located based on the locations of galaxies. If we are able to predict galaxy properties directly from void properties, this will be invaluable as we will be able to measure galaxy properties much more easily and gain insight about galaxy evolution much more quickly.

2 Data & Methods

I use the IllustrisTNG simulation[3] for my analysis [9]. I use the simulation at redshift $z = 0$, which corresponds to present day. Choosing a fixed time for the analysis simplifies the interpretation since the galaxies are not evolving over time in my analysis. I use the largest simulation available, which has a box length of 300 Mpc, in order to have a large number of large voids within the simulation. If my simulation box was smaller, I would not have the large voids present in the larger simulation. The IllustrisTNG simulations are N-body simulations of cosmic structure formation that incorporate hydrodynamics. This is interesting for studies like mine so I can analyze features of galaxies impacted by hydrodynamics. This positions, properties, and IDs of the halos in the simulation are denoted as the halo catalog.

I locate voids in the IllustrisTNG halo catalog with the void finding algorithm VIDE [10]. This produces a void field containing void positions, properties, and IDs. With this information I can assign the void ID for a given void to the halo IDs for the galaxies that reside within that void. In

[1] Galaxies are often referred to as halos in simulations, so I will use these terms interchangeably.

[2] I checked and could actually not find any papers on this.

[3] Downloadable here: http://www.tng-project.org/data/

this way I produce a feature matrix with feature vectors for each halo ID. In addition to the features provided by the simulation and void finding algorithm, I compute moments for different feature distributions for the halo populations within voids. I compute the median, standard deviation, skew, and kurtosis for each halo feature provided by the simulation. In this way, all halos associated with a given void ID will have the same values for the feature distribution moments. I compute the median rather than the mean since the feature distributions tend to be highly skewed. The entire data set then consists of 17,625,892 galaxies and 64 features. Features and their descriptions are listed in Table 8. None of the data need to be cleaned since they are from a simulation.

When predicting features about the galaxies within voids, I filter my dataset to only contain galaxies that reside inside voids. This reduces my dataset to 888,676 galaxies. This in itself is an interesting result showing that such a small fraction of the galaxies are in voids. Since there is an entire population of galaxies within a single void, I choose to predict moments of the galaxy features from void features rather than individual features. The magnitude of the features in my dataset ranges from 0 to 10^{10}, so I standardize all my features before performing any quantitative analysis. Some of this is based on code from precept [1].

To evaluate my results and assess the generalizability of my models, I split my data into a training and testing set, in which the testing set is held out from training the models and constitutes 25% of the original data (75% is the training set). I use scikit-learn's StratifiedShuffleSplit algorithm to do this [8]. This algorithm allows us to do stratified K-fold cross-validation when splitting the data. This is beneficial so that I do not accidentally bias either my training or testing data relative to the original data. I choose to stratify the data based on the galaxy mass (this is often considered the most important feature of galaxies, as mentioned below, and so most likely maintains the stratification across multiple features) such that the training and testing data will have the same proportion of galaxies within each mass cut as exists in the original data set. I also verify this produced roughly equivalent proportions of galaxies that reside in voids for both the training and testing data. Throughout my analysis, I set the random seed so that I can reproduce my results.

The goal of this work is to predict the median galaxy mass for the population of galaxies within a given void based on that void's properties (motivated further in section 3). Thus, my problem is a regression problem. Before performing any regression, I drop columns that contain only 0s. I compare 4 regression models: Linear Regression (LR), Elastic Net (EN), Bayesian Ridge Regression using an Elastic Net for feature selection (BR + EN), and a Random Forest (RF). LR, EN, and BR are all linear regression models. I use the LR as one baseline model to compare my results to. I also use the EN as a baseline regularized model. I include the BR in my analysis so that I can obtain uncertainties on my parameter estimates. To optimize the features used in the BR, I use the EN to select the optimal features. I choose to use an EN so that the balance between the L1 and L2 norms can be optimized. I want the optimal balance between sparsification and robustness. Finally, I use a RF due to their general success in prediction and classification. While I cannot obtain an equation with coefficients from the RF, this model gives us an opportunity to have a highly accurate model. A random forest is an ensemble model, whereas a decision tree is not. A single decision tree can be prone to severely overfitting the data. I chose to use the random forest rather than a single decision tree in an effort to avoid this potential overfitting. To set hyperparameters for the models, I use scikit-learn's GridSearch cross-validation [2]. Optimized hyperparameters are listed in Table 6 for the EN and Table 7 for the BR in Appendix A. All other parameters are set at their default scikit-learn values. I chose not to optimize the number of trees in the RF due to the computational time this cross-validation required. Oftentimes in machine learning I must make a choice between computational cost and optimal parameters. However, it would be interesting to probe the optimal number of trees for the RF in future work using a cluster. I use the mean squared error of the model's prediction for the testing data as my performance metric throughout my analysis.

3 Results

To get a sense of how the halo distribution spreads along their properties, I begin with a PCA analysis of only the halo features (16 features) for all halos in the simulation (see Table 8 in Appendix A for parameter descriptions). In other words, using halos that are inside voids and also halos that are outside of voids. After varying the number of principle components, I found that 4 components provided the most physically sensible number of features (see Figure 1). Indeed, with more than 4 components there was significant repetition among the components and the weightings of the features were not physically meaningful. For example, the features VX, VY, and VZ would alternate weightings for a number of these components, which is not meaningful because the directions are defined in terms of the simulation. The magnitude of the peculiar velocity vector, Vpec, is most meaningful.

PCA 0: EVR = 0.2940

	Component	Weight
7	GroupWindMass	0.361959
6	GroupStarMetallicity	0.342809
5	GroupSFR	0.334031
15	hasWind	0.317440
1	GroupStellarMass	0.299451
14	hasSF	0.299415
0	GroupBHMass	0.293638
4	GroupMass	0.286832
13	hasBH	0.285080
2	GroupBHMdot	0.231045
12	GroupStellarMassFraction	0.227405
3	GroupGasMetallicity	0.070208
11	Vpec	−0.003480
9	VY	−0.000197
8	VX	0.000161
10	VZ	0.000130

PCA 1: EVR = 0.1701

	Component	Weight
4	GroupMass	0.448101
0	GroupBHMass	0.444417
1	GroupStellarMass	0.439940
14	hasSF	−0.320800
13	hasBH	−0.314541
15	hasWind	−0.285350
6	GroupStarMetallicity	−0.258725
12	GroupStellarMassFraction	−0.200031
2	GroupBHMdot	0.076224
5	GroupSFR	0.071476
3	GroupGasMetallicity	−0.069487
7	GroupWindMass	−0.046647
11	Vpec	0.001916
10	VZ	−0.000248
8	VX	−0.000026
9	VY	0.000021

PCA 2: EVR = 0.0702

	Component	Weight
2	GroupBHMdot	−0.570992
5	GroupSFR	−0.458801
13	hasBH	0.320360
7	GroupWindMass	−0.308943
14	hasSF	0.284231
4	GroupMass	0.225279
0	GroupBHMass	0.210482
1	GroupStellarMass	0.198500
11	Vpec	−0.145379
3	GroupGasMetallicity	−0.133183
15	hasWind	0.071248
10	VZ	0.053215
6	GroupStarMetallicity	0.042378
12	GroupStellarMassFraction	0.028162
8	VX	−0.027290
9	VY	−0.005933

PCA 3: EVR = 0.0684

	Component	Weight
11	Vpec	0.701499
3	GroupGasMetallicity	0.544937
10	VZ	−0.340744
8	VX	0.145931
2	GroupBHMdot	−0.131348
5	GroupSFR	−0.115315
7	GroupWindMass	−0.113043
12	GroupStellarMassFraction	0.109012
9	VY	0.074093
0	GroupBHMass	0.053906
4	GroupMass	0.053527
1	GroupStellarMass	0.051693
6	GroupStarMetallicity	0.051074
15	hasWind	−0.040855
14	hasSF	−0.010210
13	hasBH	0.003740

Figure 1: Halo PCA.

The principle components and their weightings capture physically relevant trends in the galaxies. The first component shows that features relating to the stars and star formation activity in the galaxy explains the most variance in the data, with an Explained Variance Ration (EVR) of 0.2940. The second component with an EVR of 0.1701 predominantly describes the mass of the galaxies. The physical interpretation of these 2 dominating PCs is striking. [4] performed a PCA on galaxy spectra and found that the first 2 components described galaxies characterized by stellar properties and star formation activity. Their third component described post-starbust galaxies that are predominantly described by their mass. The fact that my first 2 components map well to the dominating components of [4] is exciting since I am performing an analysis on an N-body simulation while they analyzed galaxy spectra. My third component describes the black hole features in the galaxies, and black hole feedback directly impacts star formation so this is why star formation rate is still weighted heavily for this component. My last component predominantly describes the peculiar velocity of the galaxies and their gas metallicity.

In astrophysics, we consider the most defining characteristic of galaxies to be their mass [12]. Since this feature carries a large amount of information about each galaxy, I will begin with trying to predict the galaxy mass from the void features for my regression analysis. I show a histogram of galaxy masses in Figure 2 colored by galaxies that are inside voids and galaxies that are outside voids. While there is not a dramatically different mass distribution for the two galaxy populations, there is, indeed, a difference. Galaxies that are inside of voids tend to have lower masses than galaxies outside of voids. This suggests that using only void properties to predict galaxy mass may be successful.

Since I will only use void features to predict mass, I will only be using 5 features to predict the mass. I choose to do this because galaxy properties are hard to observe and are generally correlated with each other (see Figure 10 in Appendix A). Since each void has a population of galaxies within it, I choose to predict the median galaxy mass of the galaxy population within the void. To accomplish this I utilize a RF regression and BR regression, whose features are selected from the top features from the EN. I choose to select all features that carry a weight greater than 0.01 to only eliminate features that hardly have any impact. I compare these results to baseline fits from LR and the EN. I list the mean squared error (MSE) in Table 1 for each of these models. Predicting the median galaxy mass corresponds to the 'Linear' features row. Coefficients (for LR, EN, and BR+EN) and feature importances (for RF) are listed in Table 2. Overall, the RF performs the best, followed by the LR. I believe the EN and BR+EN do not do as well because they involve regularization. Eliminating just 1 feature from 5 could be having a significant impact on the fit of the model. The RF does well compared to the LR since it can directly capture non-linearities in the data.

I compare my predictions for the median galaxy mass to the individual galaxy masses to test if predicting the median galaxy mass is, indeed, the best way to predict mass. Predicting the individual galaxy masses corresponds to the 'Linear, individual mass' row in Table 1. The models overall could not predict the individual galaxy masses as well as the median galaxy mass, as expected. The weights for the coefficients for the EN were 0 (and so no BR was performed), indicating the EN could simply not fit the data due to the regularization–evidently none of the parameters were informative for prediction. The MSE for both the LR and RF are higher than when predicting the median galaxy mass. The most striking difference between the two prediction cases occurs for the RF. When predicting the median halo mass, the RF is astonishingly accurate, but when predicting the individual halo masses, the RF is no better than the other models. I believe this speaks to the fact that voids contain an entire population of halos, so there is a large amount of scatter in the halo masses within the void, making it difficult for the single values describing the void properties to be predictive for this parameter. In contrast, the median halo mass does not have scatter and describes the population of halos within the void with a single value.

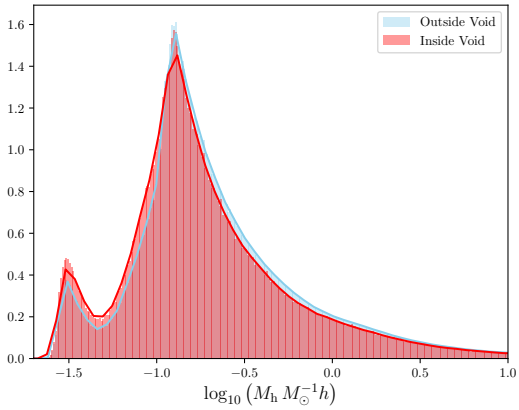

Figure 2: Halo mass distribution for halos inside voids and halos outside voids.

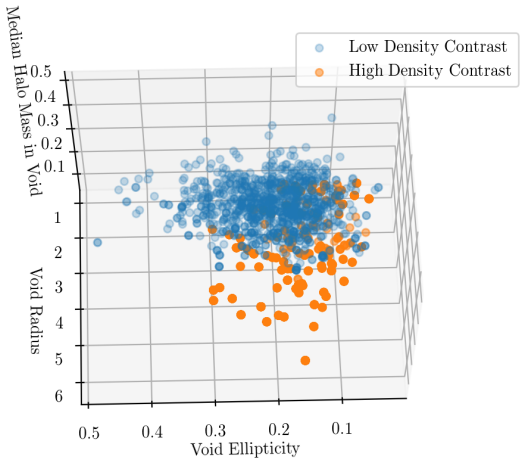

Figure 3: Relationship between void ellipticity and radius for low density contrast voids and high density contrast voids.

What information do voids add? To test this I perform the exact same analysis as above to predict the median galaxy mass, but this time I not only use the void features but also use all of the halo features (see Table 8 in Appendix A for a full description of all parameters used). This test corresponds to the 'Linear halo + void features' row in Table 1. See Figure 11 in Appendix A for the feature weights/ importances for this analysis. The MSE for the linear models improves, as expected, since the halo features are correlated with each other (as discussed earlier). However, the RF MSE degrades, which is quite interesting. These MSE values indicate that mass scales relatively linearly with the other halo features (again, this is expected since these features are correlated with each other), but that incorporating the halo features is unnecessary for models that are capable of taking into account non-linearities in the data. The relationship between mass and void features is non-linear so the RF does well for this relationship. Adding the halo features does not improve the RF's prediction ability, indicating void features are sufficient to predict mass when using non-linear models.

Why is the Random Forest model so predictive of the median halo mass? In Figure 6 and Figure 7 I show qq-plots for the RF and BR+EN models for the test data and test data predictions. I see that the RF halo mass predicted distribution is in exquisite agreement with the true values. In

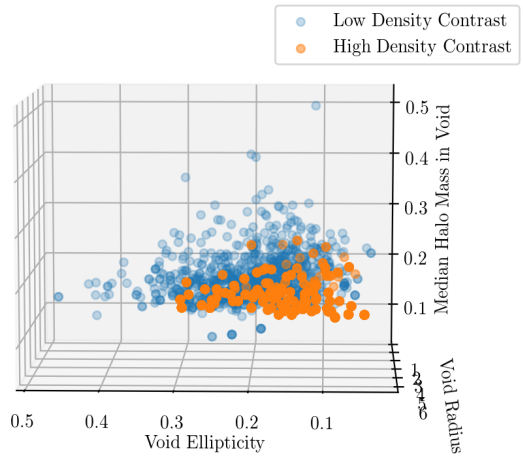

Figure 4: Illustration of how void ellipticity impacts the median galaxy mass.

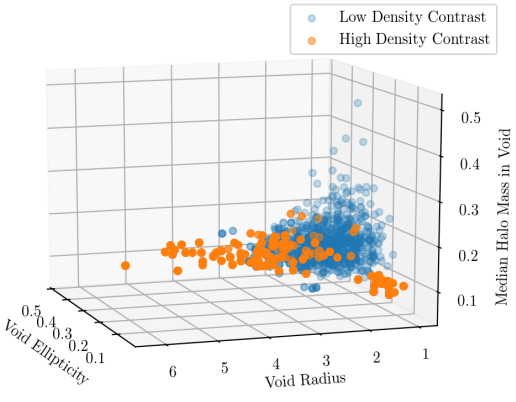

Figure 5: The relationship between void ellipticity, radius, and median galaxy mass for high density contrast voids.

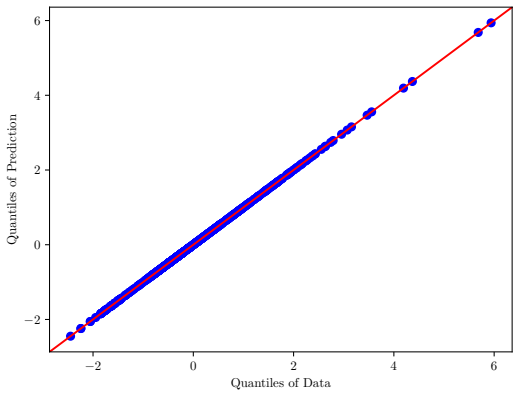

Figure 6: qq-plot for the RF regression using linear void features.

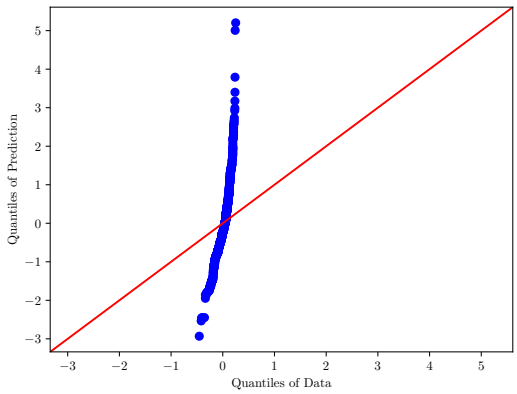

Figure 7: qq-plot for the BR + EN regression using linear void features.

contrast, the BR+EN model has heavier left and right tails than the true distribution, indicating it overpredicts high and low masses. Thus, it has a worse MSE than the RF.

I show the weights/ feature importance in Table 2 for the LR, BR+EN, and RF models for the void features used in the analysis with only void features. The top 3 important features for the RF were the void density contrast, ellipticity, and radius. The top 3 features for both the LR and EN included ellipticity, as well. The LR also ranked density contrast with the highest weight, but included the number of children in its top features rather than radius. The BR+EN included the radius, but did not included the void density contrast in its features.

To visualize how void density contrast, ellipticity, and radius impact the median galaxy mass (since these are the top features for the RF and the RF performed the best), I split the training data into two groups: one group that has density contrasts less than the mean density contrast, denoted 'Low Density Contrast,' and one group that has density contrasts larger than the mean density contrast, denoted 'High Density Contrast.' I then plot the median halo mass as a function of ellipticity and radius for these two groups in Figure 3, Figure 4, and Figure 5. Figure 3 illustrates the interplay of the void features. I see that for voids with a low density contrast, there is minimal spread in void radius. Most of these voids are small. In contrast, most of the spread in this population is along the void ellipticity. For voids with a high density contrast, void ellipticity tends to increase with void radius and has more scatter for large values of radius and ellipticity. Since the low density contrast voids tend to be small and do not have much scatter in size, it is useful to look at how the median mass changes predominantly as a function of ellipticity, shown in Figure 4. I see that the low density contrast voids tend to have higher median halo masses than the high density contrast voids. Further, the median halo mass for the low density voids tends to decrease and have less scatter as void ellipticity increases. Rounder low density contrast voids tend to have larger median halo masses. Figure 5 shows the complex relationship between high density contrast voids and the median halo mass. Small, round, high density contrast voids have low median galaxy masses. As the void radius and ellipticity increase, the median halo mass in the high density voids increases and then levels off, illustrating the highly non-linear relationship between these features.

The low density voids have a less complex relationship with the median halo mass, and the low density voids also contain the more massive median galaxy masses. Thus, it seems the LR may be predominantly fitting the low density contrast population, which would explain why void radius was not one of its top features, since the low density contrast voids tend to be clustered around small radius values.

Features	LR	EN	BR + EN	RF
Linear	0.9072	1.02758	0.98629	5.316×10^{-6}
Linear, individual mass	0.9999	1.0	—	0.9996
Linear halo + void features	0.3076	0.9802	0.4727	2.084×10^{-5}
Deg 2 Poly	0.8446	1.01779	0.88572	4.456×10^{-6}
power Law	0.82076	1.63769	0.82076 (no EN)	5.066×10^{-6}

Table 1: Mean squared error (MSE) for regression models.

Feature	LR	BR + EN	RF
voidDensityContrast	-0.288656	—	0.402250
voidEllipticity	-0.167343	-0.109672 ± 0.001221	0.342885
voidRadius	-0.005088	0.027225 ± 0.001610	0.244195
voidCentralDen	-0.002654	-0.016113 ± 0.001333	0.007460
voidNumChildren	0.044095	0.010645 ± 0.001566	0.003210

Table 2: Feature weights/ importances for linear void features.

Features	Low Density Contrast	High Density Contrast
Linear	0.0062	1.1518
Deg 2 Poly	0.93384	0.84726
power Law	0.95695	0.82075

Table 3: LR MSE split by density contrast for linear void features.

To test this I split the LR MSE by void density contrast to determine the MSE for voids with a low density contrast and the MSE for voids with a high density contrast. The dramatic difference in MSE values in Table 3 for the two void populations when using the linear features confirms my observation that the LR is predominantly fitting the low density contrast voids when the linear features are used. Indeed, when incorporating non-linearities into the features (see discussion below), the MSE improves for the high density contrast voids and degrades for the low density contrast voids such that both have similar values, indicating that now both populations are being considered by the model and contributing positively towards its predictive power.

Since there are non-linear relationships visually present in the scatterplots, I now transform my void features to a 2nd order polynomial (denoted as Deg 2 Poly in tables). I choose a second order polynomial because lower order polynomials are often preferred in astrophysics (and in general). I show the MSE for the models using this data in Table 1 and the feature weights/ importances in Table 4. I bold the top 5 features for each model in Table 4. I find that the MSE improves for all models using the 2nd order polynomial, with the RF still providing the highest accuracy. Since the MSE for the linear models improve, this indicates that they are now better able to fit non-linearities present in the data. In Table 4 I find that the RF top 5 important features all still involve the density contrast, ellipticity, and radius. This trend is in agreement with what I found with using the linear void features. Interestingly, the RF top feature is the (ellipticity $\times$ radius) term– this perhaps speaks to the fact that for high density contrast voids, the ellipticity tends to increase with radius and as this occurs, the median halo mass tends to increase, as well. Further, the top features for the BR+EN model now only involve ellipticity, radius, and density contrast, as do the majority of its features. When using the linear features, the void density contrast was not selected at all to use in this model. This indicates that the models are now beginning to converge on which features are most predictive of mass.

In Figure 8 I show the qq-plot for the BR+EN using the selected 2nd order polynomial features. Again, the right tail of the predicted distribution is heavier than the true data, and in fact it is a bit heavier than it was for the linear void features. However, now the left tail oscillates from being heavier to being lighter and then approximately equal to the true distribution at the tip of the left tail. This is interesting and suggests that incorporating the non-linearities from the 2nd degree polynomial helped predict the low median galaxy masses. Perhaps with these nonlinearities present the BR+EN model is better able to predict the small cluster of voids with high density contrast, low radius, and low ellipticity with low median halo masses in Figure 5 but has a difficult time with the path there since this cluster is relatively separated from the rest of the points.

The fact that the MSE for the BR+EN model has improved with the 2nd order polynomial and that the highly weighted features are beginning to converge towards those ranked as important by the RF suggests that: (1) the data is, indeed, strongly non-linear and requires strongly non-linear models to fit it well, and (2) the most important features for predicting the median halo mass are indeed perhaps those selected by the RF.

In the spirit of the non-linear nature of the data and the most abundant equations in astrophysics, I now transform the void features to be power laws of just the void density contrast, ellipticity, and radius, which, again, are becoming preferred by the linear models as non-linearities are introduced to the features. Thus, my regression problem takes the form:

$$\tilde{M}_{\rm h} = \delta^{\alpha} e^{\beta} r_{\rm eff}^{\gamma}, \tag{1}$$

Feature	LR	BR + EN	RF
voidEllipticity voidRadius	-0.116405	**-0.235247 ± 0.004730**	**0.215216**
voidDensityContrast	**-1.317766**	**-0.875281 ± 0.005303**	**0.138871**
voidEllipticity voidDensityContrast	0.121850	**0.606563 ± 0.005240**	**0.133671**
voidDensityContrast^2	**0.637442**	—	**0.121085**
voidRadius voidDensityContrast	**0.350251**	—	**0.111713**
voidEllipticity^2	-0.014487	—	0.079694
voidEllipticity	-0.162293	-0.170988 ± 0.003660	0.078789
voidRadius	-0.131148	**0.332322 ± 0.008559**	0.058743
voidRadius^2	0.109956	**-0.259320 ± 0.009312**	0.050404
voidDensityContrast voidCentralDen	0.260904	—	0.003452
voidRadius voidCentralDen	**-0.692097**	—	0.002401
voidEllipticity voidNumChildren	0.104119	—	0.002399
voidCentralDen^2	-0.111892		0.001275
voidRadius voidNumChildren	-0.048491	0.224913 ± 0.017066	0.000968
voidDensityContrast voidNumChildren	-0.157952	—	0.000530
voidEllipticity voidCentralDen	0.140832	—	0.000401
voidNumChildren	0.110550	0.010304 ± 0.009558	0.000216
voidNumChildren^2	-0.012982	-0.143589 ± 0.007340	0.000130
voidCentralDen	**0.385748**	—	0.000032
voidNumChildren voidCentralDen	0.046160	—	0.000009

Table 4: Feature weights/ importances for 2nd order polynomial features regression.

where $\tilde{M}_{\rm h}$ is the median halo mass, δ is the void density contrast, e is the void ellipticity, $r_{\rm eff}$ is the void radius, and α, β, and γ are exponents to be fit. I can then transform my equation into a linear equation:

$$\log_{10}\tilde{M}_{\rm h} = \alpha\log_{10}\delta + \beta\log_{10}e + \gamma\log_{10}r_{\rm eff}, \quad (2)$$

which I can fit in the same manner as before. I note, however, that I do not perform feature selection this time since I have restricted the analysis to only 3 variables, which were selected from the performance of the previous tests.

I list the weights for the power law analysis in Table 5. The values for the BR analysis are reasonable based on the patterns seen in Figure 3, Figure 4, and Figure 5: lower density contrast is associated with higher mass, lower ellipticity is associated with higher mass (seen in the pattern for the low density contrast voids), and higher mass is associated with higher radius (seen in the pattern for the high density contrast voids). The low density contrast voids have a simpler relationship with mass than the high density contrast voids, so they dominate the constraints on α and β, explaining why both values are negative. The low density contrast voids do not have much scatter in radius, so the high density contrast voids dominate the constraints on γ. Overall, the uncertainties for all the BR constraints in this work are encouraging since they are relatively small.

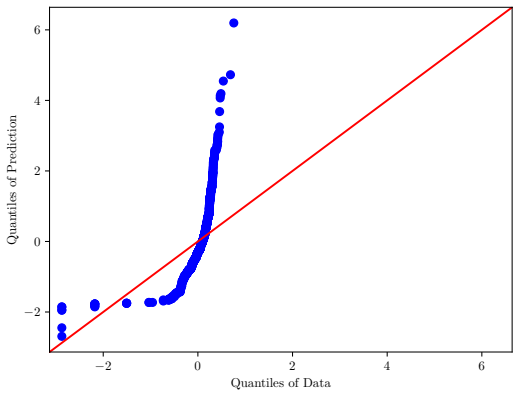

Figure 8: qq-plot for the BR + EN model using the Deg 2 Poly features.

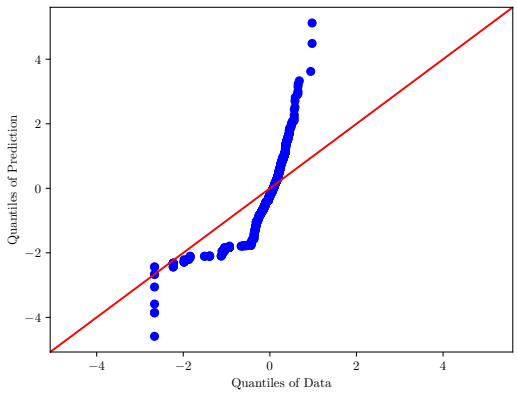

Figure 9: qq-plot for the BR model using the power law features.

I list the MSE for each of my models in Table 1. The MSE improves for both the LR and BR, but becomes worse for the EN and RF. The EN MSE likely decreases because there are so few features that the EN cannot select features and optimize its performance. The fact that the MSE improves for both of the linear models further indicates that data is extremely non-linear and that perhaps a combination of both power laws and polynomials is required to find a good fit with linear models. In Figure 9 I plot the qq-plot for the BR power law model. This time I see that even though the right tail of the predicted distribution is still heavier than the true distribution, it is not has heavy as in previous analysis. So, this model does not overpredict the massive median halo masses as severely as the other models. However, at the same time there is some tradeoff at the least massive end where I find that the predicted distribution returns to being heavier than the true distribution at its tip. The MSE for the RF decreases because I have limited the RF to only 3 features. The RF cannot try to predict all the non-linearities present in the data since I have limited its flexibility. This indicates that (1) the RF is the only model in my analaysis that is capable of capturing the complex non-linearities in the data and (2) the relationship in the data between the void features and the median halo mass is, again, far more complicated than the simple power law I have fit here. This warrants further analysis if one hopes to achieve at least an approximate equation for the trends present in the data.

4 Discussion and Conclusion

In this work I have investigated predicting galaxy mass using only void features. I used galaxies in the IllustrisTNG simulation and found voids in this simulation using VIDE. I performed a regression analysis to predict the median galaxy mass of the galaxy population within voids based only on the properties of the voids. I find that the Random Forest yields the best results with an MSE of order 10^{-6} due to its ability to capture the non-linearities in the data. One potential drawback of my analysis is that I did not have sufficient computational time to perform an optimization on the number of trees in my random forest. Since I only use the default value, perhaps the number of trees is not large enough and is, then, leading to the model overfitting the data still. This should be investigated in future work.

The performance of the linear models improves when introducing non-linearities into the void features, such as transforming them to a 2nd order polynomial or a power law. This indicates that a more complicated transformation of the features is needed if one hopes to obtain an analytic expression for the relationship between the void parameters and the median galaxy mass. It would be interesting to pursue this in future work and investigate the performance of the models when using a combination of a polynomial and power law transformation of the features.

Despite the correlations among halo features and that all them tend to be predictive of halo mass, we find that void features are independently able to make strong predictions for halo mass when using a nonlinear regression model.

This analysis has focused on the role voids play on galaxy properties. There are several take-away points from this work that were unclear at the onset: (1) the relationship between void properties and halo mass is extremely nonlinear, (2) void ellipticity seems to have an important influence on galaxy properties, (3) halo features seem to be unnecessary for predicting galaxy mass if non-linear models are used. In terms of point (2), it would be interesting to investigate void ellipticity in further detail. Perhaps this property is revealing information about the evolution history of the void, for example if it just underwent a merger with another void that, consequently, impacted the evolution of the galaxies within those voids. This could be investigated by analysing simulations at multiple redshifts. Additional work that will be interesting is investigating the predictive power of void properties for other galaxy properties, such as star formation rate and black hole mass. To investigate star formation rate, it will be best to use the simulation at redshift $z = 2$, which is when the star formation rate peaks in the universe. Overall, this work yielded interesting results on how voids can illuminate galaxy properties.

Feature	LR	BR	RF
voidDensityContrast, α	-0.430270	-0.430262 $\pm$ 0.001195	0.430152
voidEllipticity, β	0.207581	-0.177645 $\pm$ 0.001175	0.331069
voidRadius, γ	-0.177650	0.207577 $\pm$ 0.001129	0.238779

Table 5: Feature weights/ importances for power law features regression.

A Supplementary Material

This section contains the correlation between halo features (Figure 10), tables for selected hyperparameters for the regression models (see Table 6 and Table 7), descriptions of the data features (see Table 8), and weights for the regressions when using all halo and void features (described in section 3, see Figure 11).

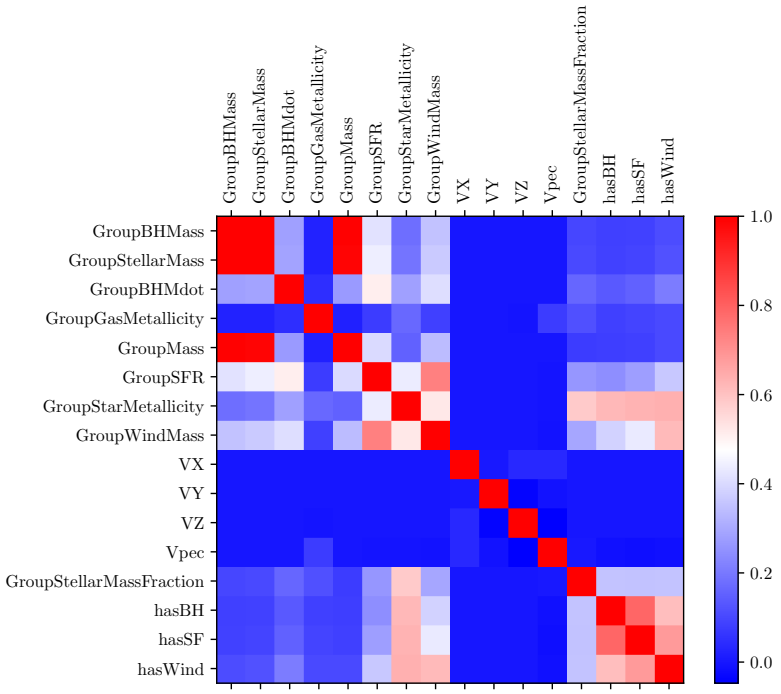

Figure 10: Halo feature correlations.

Features	α	L1 ratio
Linear	0.001	0.1
Linear, individual mass	100	0.1
Linear halo + void features	0.001	0.1
Deg 2 Poly	0.001	0.1
power Law	0.001	0.1

Table 6: Hyperparameter values for Elastic Net. α was varied over 10^x, where $x \in [-3, 3]$. The L1 ratio was varied between 0.1 and 1 in steps of 0.1.

Features	α_1	α_2	λ_1	λ_2
Linear	10^{-10}	10^{-7}	10^{-13}	10^{-7}
Linear halo + void features	10^{-8}	10^{-7}	10^{-6}	10^{-13}
Deg 2 Poly	10^{-10}	10^{-7}	10^{-13}	10^{-7}
power Law	10^{-10}	10^{-7}	10^{-13}	10^{-7}

Table 7: Hyperparameter values for Bayesian Ridge Regression. α_1, α_2, λ_1, and λ_2 were all varied over 10^x, where $x \in [-13, -6]$. These values were chosen to take their max at the default model value and probe smaller values in hopes of improving accuracy.

Feature	Description
GroupBHMass	sum of black hole masses within galaxy
GroupBHMdot	sum of black hole accretion rates within galaxy
GroupGasMetallicity	gas metallicity within galaxy
GroupMass	galaxy total mass
GroupSFR	galaxy star formation rate
GroupStarMetallicity	galaxy stellar metallicity
GroupStellarMass	galaxy stellar mass (included within total mass)
GroupWindMass	galaxy wind mass
VX	galaxy peculiar velocity X component
VY	galaxy peculiar velocity Y component
VZ	galaxy peculiar velocity Z component
Vpec	galaxy peculiar velocity magnitude
hasBH	if galaxy has a black hole or not
hasSF	if galaxy has a nonzero star formation rate
hasWind	if wind present in galaxy
GroupStellarMassFraction	fraction of galaxy mass belonging to stars
voidRadius	average void radius
voidDensityContrast	void density contrast: contrast between inner and outer densities
voidEllipticity	ellipticity of void
voidNumChildren	number of sub voids within the void
voidCentralDen	density within 1/4 radius of void center
voidGalDen	density of galaxies within void: number of galaxies / void volume
voidGalLowMassFrac	fraction of galaxies within void lower than median galaxy mass for all galaxies within voids
voidGalMedian*	median of the distribution for each galaxy feature; * = galaxy feature
voidGalStd*	standard deviation of the distribution for each galaxy feature
voidGalSkew*	skew of the distribution for each galaxy feature
voidGalKurtosis*	kurtosis of the distribution for each galaxy feature

Table 8: Feature Descriptions

Linear Regression

	Component	Weight
24	voidGalMedianStarMetallicity	1.987820e+11
7	voidGalMedianStellarMass	-1.987820e+11
8	voidGalStdStellarMass	4.380353e-01
11	voidGalStdBHMass	-3.699179e-01
19	voidGalSkewGasMetallicity	3.643468e-01
17	voidGalMedianGasMetallicity	-3.136154e-01
9	voidGalSkewStellarMass	-2.866951e-01
10	voidGalKurtosisStellarMass	2.863548e-01
20	voidGalKurtosisGasMetallicity	-2.346246e-01
6	voidGalLowMassFrac	1.790105e-01
0	voidRadius	1.685188e-01
18	voidGalStdGasMetallicity	-1.684708e-01
5	voidGalDen	-9.401533e-02
26	voidGalSkewStarMetallicity	-7.620537e-02
32	voidGalStdVpec	-7.484883e-02
30	voidGalKurtosisWindMass	-6.942152e-02
1	voidDensityContrast	5.829150e-02
13	voidGalKurtosisBHMass	-5.335375e-02
3	voidNumChildren	-4.753127e-02
21	voidGalStdSFR	-4.290419e-02
15	voidGalSkewBHMdot	-4.234315e-02
25	voidGalStdStarMetallicity	4.128931e-02
16	voidGalKurtosisBHMdot	-3.864893e-02
2	voidEllipticity	-3.564278e-02
31	voidGalMedianVpec	2.563914e-02
27	voidGalKurtosisStarMetallicity	2.403145e-02
28	voidGalStdWindMass	2.001598e-02
4	voidCentralDen	-1.293319e-02
22	voidGalSkewSFR	1.289192e-02
14	voidGalStdBHMdot	-1.132908e-02
35	hasBH	9.010057e-03
33	voidGalSkewVpec	-8.496075e-03
34	voidGalKurtosisVpec	-8.137774e-03
23	voidGalKurtosisSFR	3.751486e-03
37	hasWind	-2.835206e-03
36	hasSF	2.492507e-03
38	GroupStellarMassFraction	-1.586570e-03
29	voidGalSkewWindMass	-9.387008e-04
12	voidGalSkewBHMass	7.949499e-04

Bayesian Ridge Regression + Elastic Net

	Component	Weight	sigma
3	voidGalLowMassFrac	0.386265	0.001095
5	voidGalSkewStarMetallicity	-0.346075	0.001012
1	voidGalSkewGasMetallicity	0.264941	0.000897
2	voidRadius	0.231718	0.001217
0	voidEllipticity	-0.095882	0.000857
4	voidNumChildren	-0.081038	0.001096

Random Forest

	Component	Weight
17	voidGalMedianGasMetallicity	0.640499
5	voidGalDen	0.044141
18	voidGalStdGasMetallicity	0.034881
32	voidGalStdVpec	0.020086
6	voidGalLowMassFrac	0.017942
33	voidGalSkewVpec	0.017247
31	voidGalMedianVpec	0.017239
19	voidGalSkewGasMetallicity	0.016664
26	voidGalSkewStarMetallicity	0.015147
20	voidGalKurtosisGasMetallicity	0.013343
2	voidEllipticity	0.013282
34	voidGalKurtosisVpec	0.013082
1	voidDensityContrast	0.012134
25	voidGalStdStarMetallicity	0.011962
9	voidGalSkewStellarMass	0.011603
27	voidGalKurtosisStarMetallicity	0.010504
11	voidGalStdBHMass	0.010137
0	voidRadius	0.009281
8	voidGalStdStellarMass	0.008126
14	voidGalStdBHMdot	0.007833
21	voidGalStdSFR	0.007341
28	voidGalStdWindMass	0.006066
22	voidGalSkewSFR	0.005484
12	voidGalSkewBHMass	0.005483
10	voidGalKurtosisStellarMass	0.004823
23	voidGalKurtosisSFR	0.004743
15	voidGalSkewBHMdot	0.004732
13	voidGalKurtosisBHMass	0.004573
16	voidGalKurtosisBHMdot	0.004158
29	voidGalSkewWindMass	0.003481
30	voidGalKurtosisWindMass	0.003045
24	voidGalMedianStarMetallicity	0.000899
3	voidNumChildren	0.000036
4	voidCentralDen	0.000001
7	voidGalMedianStellarMass	0.000000
35	hasBH	0.000000
36	hasSF	0.000000
37	hasWind	0.000000
38	GroupStellarMassFraction	0.000000

Figure 11: Regression weights/ feature importances for the linear halo + void feature analysis.

References

[1] Cos 424/sml302: Cos 424: Fundamentals of machine learning.

[2] Towards data science. https://towardsdatascience.com. Accessed: 2019-03-01.

[3] J. Richard Bond, Lev Kofman, and Dmitry Pogosyan. How filaments of galaxies are woven into the cosmic web. *Nature*, 380(6575):603–606, 1996.

[4] A. J. Connolly, A. S. Szalay, M. A. Bershady, A. L. Kinney, and D. Calzetti. Spectral Classification of Galaxies: an Orthogonal Approach. , 110:1071, Sep 1995.

[5] M. Habouzit and A. Pisani. Properties of galaxies and supermassive black holes in cosmic voids with the simulation Horizon-AGN.

[6] Jianan Hui, Miguel Aragon, Xinping Cui, and James M. Flegal. A machine learning approach to galaxy-LSS classification - I. Imprints on halo merger trees. , 475(4):4494–4503, Apr 2018.

[7] Sadegh Khochfar and Jeremiah P. Ostriker. Adding Environmental Gas Physics to the Semi-analytic Method for Galaxy Formation: Gravitational Heating. , 680(1):54–69, Jun 2008.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Volker Springel, Rüdiger Pakmor, Annalisa Pillepich, Rainer Weinberger, Dylan Nelson, Lars Hernquist, Mark Vogelsberger, Shy Genel, Paul Torrey, Federico Marinacci, and Jill Naiman. First results from the IllustrisTNG simulations: matter and galaxy clustering. , 475(1):676–698, Mar 2018.

[10] P. M. Sutter, G. Lavaux, N. Hamaus, A. Pisani, B. D. Wandelt, M. Warren, F. Villaescusa-Navarro, P. Zivick, Q. Mao, and B. B. Thompson. VIDE: The Void IDentification and Examination toolkit. *Astronomy and Computing*, 9:1–9, Mar 2015.

[11] S. Tavasoli, H. Rahmani, H. G. Khosroshahi, K. Vasei, and M. D. Lehnert. The Galaxy Population in Voids: Are All Voids the Same? , 803(1):L13, Apr 2015.

[12] Risa H. Wechsler and Jeremy L. Tinker. The Connection Between Galaxies and Their Dark Matter Halos. , 56:435–487, Sep 2018.