Advanced Data Analysis Project

Linking galaxies to their progenitors based on galaxy morphology

Advisors: Ann B. Lee[1] & Peter Freeman[1]

Collaborator: Gregory Snyder[2]

Manjari Das[1]

March 1, 2018

Abstract

Galaxies evolve by accreting gas from the intergalactic medium and via mergers. The details of galaxy evolution physics are complex and are linked to the global properties of the Universe. It is impossible to observe a single galaxy's evolution; we can only record its current appearance (or morphology) and estimate its physical properties. However, by studying how simulated galaxies evolve, we can begin to infer how galaxies in our own Universe are related across time. In this project, we aim to link galaxies at a given time to their progenitors. We analyze the statistics of simulated galaxy data from the Illustris Project at different time points, and we use these results to build models that predict the past mass rank of a galaxy given its current mass rank, morphology, and estimated physical properties such as star-formation rate.

1 Introduction

Astronomers have long been interested in linking galaxies across time; for instance, [2] Barro. et al (2014) studied 45 massive galaxies ($mass > 10$) to see whether they were progenitors of any quiescent galaxy. Then in the following year, [3] Papovich et al. (2015) have studied which galaxies at various time points could be the progenitors of the Milky Way or the Andromeda galaxy. However, the evolution of galaxies cannot be studied in the observable Universe. We do not have continuous information about any galaxy across time. To overcome this obstacle, a standard assumption ([2, 3]) used for the linking process is that the comoving number density of galaxies[3] is constant as a function of time. This furthermore implies that the relative rank order of galaxies, for example, in mass remains constant over time.

Observation of several galaxies continuously over time is not practically possible. Hence, instead of using real data, we have to use simulated data in this project for the analyses.

[1] Carnegie Mellon University

[2] Space Telescope Science Institute

[3]The comoving number density of galaxies is the density within a volume element whose physical size expands at the same rate that the Universe as a whole expands.

One such simulated galaxy data is the Illustris Project (Vogelsberger et al. 2014),[4]. Wellons & Torrey (2017), using Illustris data, attempt to construct models linking galaxies which emitted their light about 2.2 billion years after the Big Bang to galaxies today, 13.8 billion years after the Big Bang, and vice-versa, and they presented a probabilistic model of linking galaxies across time. They have used constant number density, evolving number density and probabilistic number density for linking the galaxies.However, the works in [2, 3 & 6] have not used information on morphological statistics of the galaxies to their full extent. In this paper we will check whether we can relax the assumption of constant comoving number density as well as improve the linking method by including morphological statistics. We begin by analyzing the simulated galaxy data from two time points followed by constructing models that link galaxies while taking into account morphological parameters, mass, and star formation rate. Ultimately, we test the performance of our method on real galaxy data.

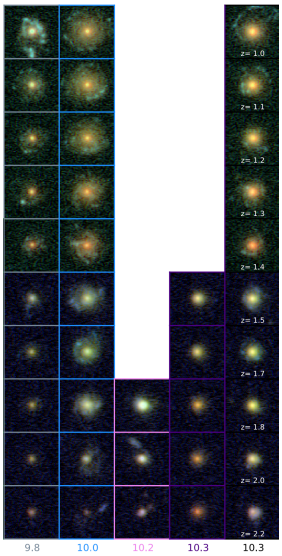
We describe the Illustris image dataset and its statistics in Section 2. In Section 3, we describe data pre-processing and introduce the rank statistic, which we use to track how galaxy mass changes with time. In Section 4, we analyze the data so as to understand the relationships among the statistics and see which are effective in explaining the mass rank of galaxies at a later time given information at a earlier time. In Section 5, we fit a random forest using cross validation to model the rank of galaxies at earlier and later times respectively, and compare them with baseline (i.e., constant-rank) models. Also we calculate the conditional density of mass rank at n earlier time given the data at a later time for each galaxy. Ultimately, in Section 6, we introduce our method of linking galaxies across time using the model from Section 5. Initially we calibrate it using Illustris data, then apply it to observed data collected by the *Hubble* Space Telescope's CANDELS program (Appendix 0).

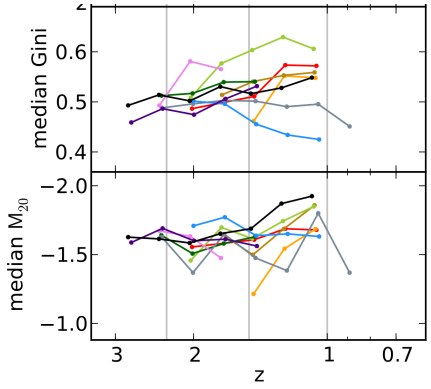
2 Data

This project will eventually feature two sets of data, one simulated and one observed (or real). In this work, we describe our analyses of simulated data; we will use the results to predict how observed galaxies evolve in time. For cosmological objects time and distance are measured in terms of the quantity called redshift[5] (denoted by z). The current time (13.8 billion years after Big Bang) corresponds to redshift $z = 0$ and the further back in time we move in, the higher the redshift gets. For distance, a greater distance corresponds to a higher redshift.

[4]www.illustris-project.org

[5]Wikipedia on redshift





plot color	M_* at $z \sim 1$ $\log_{10} M_*/M_\odot$	M_* at $z \sim 2$ $\log_{10} M_*/M_\odot$
red	9.6	9.2
orange	9.6	9.3
brown	10.0	9.4
green	9.8	9.5
dark green	9.9	9.7
gray	10.1	9.8
blue	10.3	10.0
pink	10.5	10.2
purple	10.4	10.3
black	10.4	10.3

Figure 1: Left : *A part of an image from Snyder et al. (2014). Image of simulated galaxies where the horizontal bars correspond to different mass for the galaxies and the vertical bars correspond to different redshifts.* Right: *An image from Snyder et al. showing change in median value of Gini and M_{20} as a function of redshift.*

The Illustris Project dataset consists of information about galaxies measured at many different redshifts. In this project, we begin by concentrating on two of those redshifts−z=2 and z=1−which correspond to times 3.3 and 5.9 billion years after the Big Bang, respectively. The data are pre-processed as in Snyder et al. (2014) (Figure 1) to make it appear as if they had been observed from the Hubble Space Telescope at wavelengths 105 nm (Y band), 125 nm (J band), and 160 nm (H band). Currently we work with J-band data only.

A galaxy is considered detected at a given redshift if its mass is $\geq 10^{10}$ solar masses ($M_\odot$). Since galaxies generally grow in mass, more galaxies are detected at z=1 than at z=2. Our sample consists of 2144 galaxies that are detected at each redshift.[6] The images for each galaxy are summarized via a number of statistics described in Appendix A; in addition, we have estimates of relative size (*sizes*), stellar mass (*mass*), and star-formation rate (SFR).

Table 1: Ranges of values of galaxy attributes

	M	I	D	$Gini$	M_{20}	C	A	$sizes$	$\log_{10} mass$	SFR
Min	0.00	0.00	0.00	0.00	−2.28	0.8	−0.14	0.07	9.90	0
Max	0.33	1.00	1.72	0.57	−0.43	4.8	0.87	2.77	12.05	35000

[6] CHECK: REPEATED GALAXIES AT z=1; AFFECT ON RESULTS/INTERPRETATION.

Each galaxy at each redshift is identified via a "subhalo ID." This number is not same for a given galaxy across redshifts.
In Table 2 we show a slice of the data. Each galaxy is observed from four different angles; in our analysis we randomly choose one angle for each galaxy, and use that angle at each redshift.

Table 2: Sample of the first, fifth and ninth rows of the data at redshift z=2

M	I	D	$Gini$	M_{20}	C	A	$sizes$	$\log_{10}mass$	SFR	Subhalo ID
0.000	0.000	0.096	0.488	-1.624	4.065	0.079	0.349	11.881	58.579	0
0.000	0.000	0.141	0.447	-1.482	3.054	0.087	0.403	10.106	55.132	100199
0.039	0.398	0.561	0.334	-0.762	1.603	0.140	0.877	10.373	40.219	100519

3 Methods

First we want to observe how the masses of galaxies change with time. For this we calculate the mass quantiles for the galaxies at each redshift. We call these the rank, where rank lies in (0,1]; a higher rank corresponds to a larger mass.

$$\text{rank}_{z=i} \text{ of galaxy } x : R_i(x) = P(M_{*,z=i} \leq mass_{z=i}(x))$$

The idea is to see how the rank changes for galaxies as we move from one redshift to another. Figure 2 shows how rank change from z=1 to z=2. Our interest lies in how frequently galaxies change rank, and by how much. Also, we want to see how the morphological parameters and star formation rate are related to the rank changes and the rank groups in general. We present these analyses in Section 4.

4 Exploratory Data Analysis

We want to fit a model that can estimate rank at z=2 based on the rank and other statistics at z=1. In Section 4.1 we look at the relationship between just the ranks at both redshifts, and then we look at the conditional density of rank at z=2, given the rank at z=1, in Section 4.2. In Section 4.3 we look at the relationships among the morphological statistics and rank at z=1, and finally in Section 4.4 we examine the link between morphological statistic values at z=1 and the change in mass rank from z=1 to z=2.

4.1 Mass rank at z=1 and z=2

We denote the rank at z=1 as R_1 and the rank at z=2 as R_2. The rank at redshift $z = k$ is

$$R_i(j) = \frac{1}{n_{\text{gal}}} \sum_{i=1}^{n_{gal}} \mathbb{I}[M_{*,z=k}(i) \leq M_{*,z=k}(j)] \,,$$

where $n_{\rm gal} = 2144$.

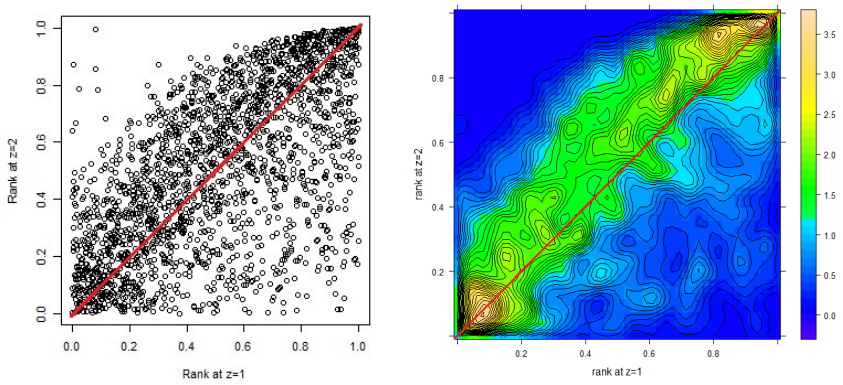


Figure 2: Left: *Scatter plot of rank at z=1 and z=2. Each point corresponds to a galaxy. There are more points above the diagonal line compared to below it. However the scatter is more pronounced below the diagonal line. Even though the actual masses at z=2 are lower than those at z=1, the relative rank value is more likely to be higher at z=2 compared to at z=1.* Right: *Joint density plot for R_1 and R_2.*

In Figure 2 we show the bivariate joint density plot for R_1 and R_2. As can be seen, the points near the diagonal have high density. For extreme rank values, the density at the diagonal is highest compared to the other points in the region.

4.2 Changes in mass rank from z=1 to z=2

In Figure 3 we show a surface plot of the conditional distribution of R_2, given R_1.

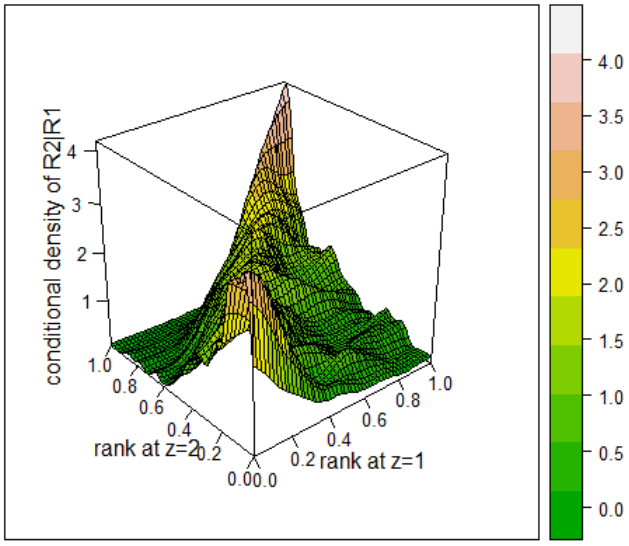


Figure 3: *Conditional density of R_2 given R_1. The distribution is negatively skewed for low values of R_1, with the skewness gradually turning positive as R_1 increases.*

We find that the rank at z=1 seems to an informative predictor of the rank at z=2.[7] Below we will see how the other variables can improve this estimation.

4.3 Galaxy morphology at z=1

Here we examine how the morphological statistics, mass, and star-formation rate at z=1 are related. To visualize the relationships we randomly sample 700 galaxies and create a scatter plot matrix with statistic values at z=1. See Figure 4.

[7] In Appendix B, we show the results of an analysis in which we examine rank *groups*; specifically, we group the ranks into quintiles and estimate the probability that a galaxy changes groups from one redshift to another.

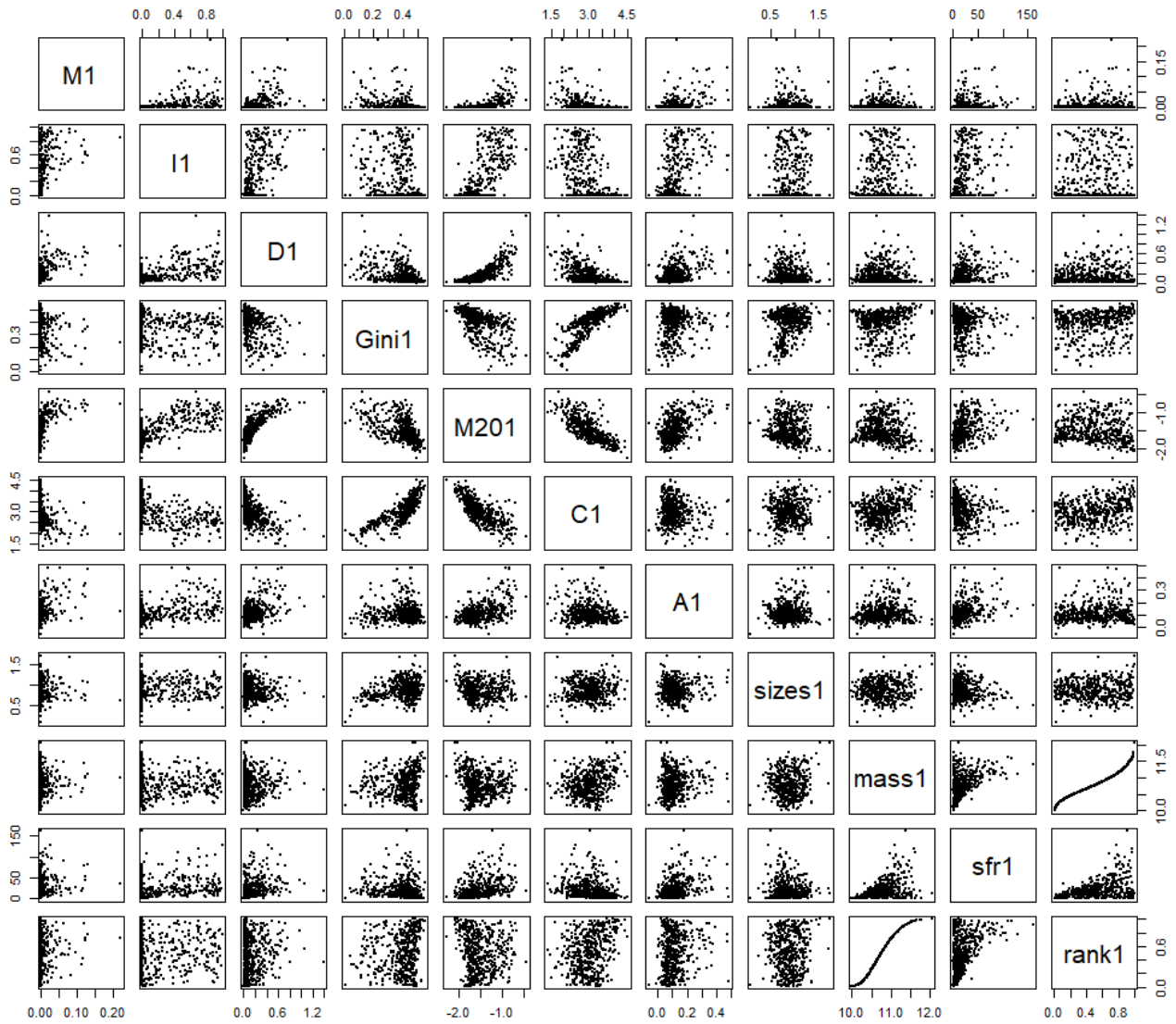


Figure 4: *Scatterplot matrix of variables at z=1.The are some linear relations among the variables. For example D, M_{20} and A have linear relation with each other with positive slope. The suffix 1 with each variable name stands for redshift* $z = 1$.

In Figure 4 we observe that the statistics I, D, M_{20}, A are positively correlated and are all negatively correlated with *Gini* and C. Secondly, *mass* and SFR are positively correlated. And lastly, *mass* and *rank* are strongly correlated, by definition. Even though there are some non zero correlations, we do not see any strong relationship between the morphological statistics and any of *mass*. SFR and *rank*. Hence, we move on to predict R_2 using all the information at $z = 1$.

5 Predicting Mass Rank at an Earlier Epoch

We want to develop a model to link galaxies at $z = 1$ with their progenitors at $z = 2$. Rank, by construction, is a bijection from the set of galaxies to the parse subset subset of (0, 1]. Hence, first we fit model to the data at $z = 1$ to predict rank at $z = 2$ (R_2) as a function of R_1 and morphological statistics, mass, and star-formation rate at $z = 1$. We divide the dataset randomly into eleven equal parts with one part left out as a test set. We apply ten-fold cross validation on the remaining parts. We then apply the models obtained from the 10 random forests separately on the test set. We use the predicted value of R_2 from the random forest as one of the estimators of R_2.

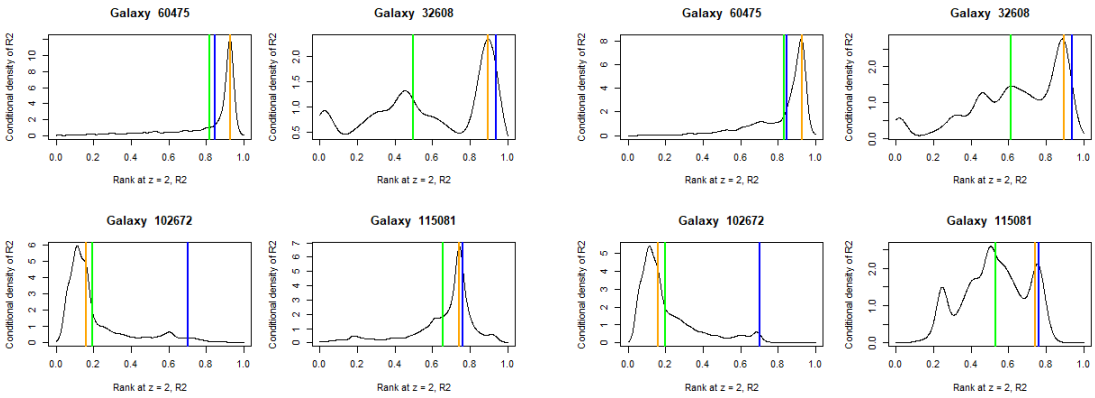
The random forest estimate is just a point estimate. If we can see a probabilistic estimate ($\mathbb{P}(estimated\, R_2 = r)\, for\, r$ in $(0, 1]$), it will give us more information about each galaxy. Hence, use the individual 500 estimates (one estimate from each tree among the 500 trees in the random forest) of R_2 for each galaxy to estimate the conditional density of R_2 for each galaxy given the information of that galaxy at $z = 1$. Step-by-step we first fit the random forest model described in the previous paragraph. Then we obtain the $\hat{R}_2$ values for the given galaxy at each tree. Then we use these individual values and calculate a kernel density with Gaussian kernel. This density is a probabilistic model for estimation of R_2. We use the mode of the density as one of the estimators for R_2. We repeat this after excluding all the morphological variables except M_*, $sizes$, and SFR.

$$f_{rank_2}(y \mid \text{morphology}, mass, SFR, R \text{ at } z = 1) = \frac{1}{n_{\text{trees}}} \sum_{j=1}^{n_{\text{trees}}} f_{N(\mu_j, \sigma^2)}(y),$$

where μ_j is a vector of length n_{trees} containing the values obtained from each tree in the random forest for R_2 of the galaxy under consideration and σ^2 fixed variance controlling the smoothness of the density.

5.1 Predicting mass rank at z=2 given data at z=1

R_2. $mass$ and R_1 are the most important predictor variables. In Figure 5, shows the variable importance for the statistics at z=1 to predict R_2. Also, it is interesting to note that morphological statistics M_{20} and C at z=1 are also effective to some extent in explaining the prediction even when the gap between the two redshifts is large ($\sim$ 2.5 Gyr).



(a) *Using all variables at* $z = 1$ *including morphologies.*

(b) *Using only* M_*, SFR *and* R_1

Figure 6: *Predictive density of 9 galaxies. The blue, the red and the green lines correspond to* R_1, R_2 *and the aggregated estimate of* R_2 *from the random forest.*

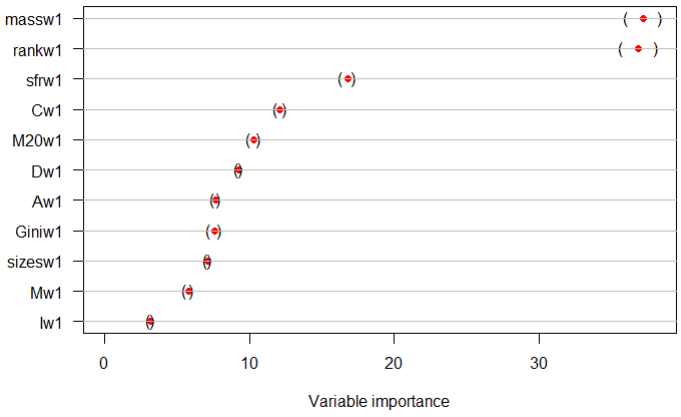


Figure 5: *Variable importance plot for* R_2 *prediction. The parentheses show the extent of variability calculated by combining the ten random forests. The most important variables in the rank estimation are mass, rank and* SFR *at* $z = 1$.

We want to check whether the inclusion of the morphological statistics improves the model. So we repeat the random forest model fitting and estimation method as earlier but without the morphological statistics. Morphologies improve the prediction model. In Figure 6, we plot the density estimated using the data at $z = 1$ ((a) with morphologies and (b) without morphologies) for finding R_2 for a sample of nine galaxies. It can be seen that the mode can be used as a predictor of R_2. Also, the model with morphologies seems to be a better fit for R_2 in general. thus we conclude that morphologies help to some extent in estimation.

We now compare our model against the model that assumes that galaxies have the same

rank at $z = 2$ as they do at $z = 1$ while using all the information at $z = 1$. We refer to it as the "fixed rank" model. We used mean square error to compare the different methods. In Table 3, we show how allowing the rank to vary leads to a significant decrease in the mean-squared error for predicting the mass rank at $z = 2$. In varying rank without morphologies, we fit the random forest using only $mass$, R_1 and SFR. Also for either case, we also find the conditional density of R_2 from the random forest model. Then the mode of the density is also used as an estimator for R_2.

Table 3: *Errors of different methods estimating rank at z=2. Including morphologies decreases the error. Mode of the conditional density estimate has the least error.*

Method	**Error**
Fixed rank	0.063 ± 0.252
Varying rank (with morphologies)	0.009 ± 0.098
Varying rank (without morphologies)	0.016 ± 0.125
Mode of cond. density (with morphologies)	0.003 ± 0.050
Mode of cond. density (without morphologies)	0.015 ± 0.122

In order to assess the consistency of the model, we also apply it to predict rank at z=1.5 using statistics at z=1 (see Appendix C). The results are consistent with the results in this section. In Appendix D, we reverse the direction of the prediction, i.e., we estimate rank at a later epoch using data at an earlier epoch. Our results are consistent with those here, with the exception that SFR seems to be the most important variable.

6 Linking galaxies at $z = 2$ to galaxies at $z = 1$

Here we develop a model to determine the progenitor(s) of a galaxy from a given set of galaxies at an earlier (higher) redshift. For this we use the random forest model from Section 5.1.

Linking algorithm:

1. Let Set be the set of galaxies on which we want to apply our method of linking.
2. For each $galaxy_j \in Set$, estimate rank at $z = 2$, $\hat{R}_2(galaxy_j)$
3. Then we look at the $R_2(galaxy_k)$ for each $galaxy_k \in Set$
4. We choose the one that has rank closest to $\hat{R}_2(galaxy_j)$. We call it $galaxy_{j'}$.

Hence,

$$galaxy_j \longrightarrow galaxy_{j'}$$

6.1 Checking performance of linking method with simulated data

To assess model performance in Section 5, we use the mean squared error. But to check consistency for real data we do not have information of the real progenitor, so utilizing the mean squared error is not an option. Hence first we check the performance of the linking method on the simulated data.

6.1.1 Two-sample paired t-test

We use the two-sample paired t-test, assuming as a null hypothesis that the means of the true and predicted progenitor populations are the same for each morphological statistic. For each $galaxy_i$ in the test set we define

$$\delta_i = statistic_{galaxy_i, z=2} - statistic_{galaxy_{i'}, z=2}$$

for each statistic, where i' is the galaxy closest in rank to the predicted rank of $galaxy_i$.

Test setup
Assumptions : δ_i's are i.i.d.
dependent variable is continuous.
$\theta \equiv \mathbb{E}[\delta_i]$.
$H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$
Test statistic : $\frac{\bar{\delta}-0}{\hat{\sigma}_{\bar{\delta}}} \sim t_{n-1}$.

To compare the performance we repeat the linking method using the fixed rank model. In the following we assume that the rank of a galaxy at $z = 2$ is the same as that at $z = 1$ for the linking. We plot the boxplots of the δ_i's side-by-side in Figure 7.

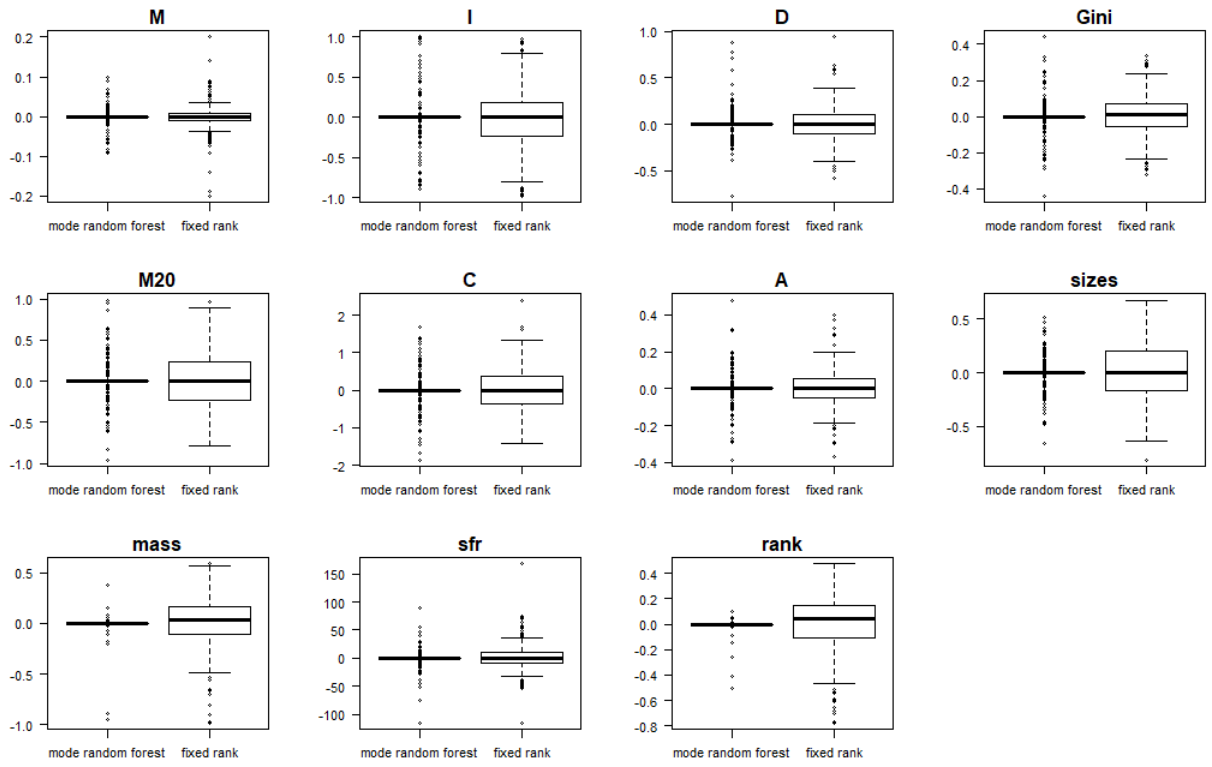


Figure 7: *Boxplot of δ_i's for different morphological statistics for mode of conditional density using random forest model vs fixed rank model.*

It can be clearly seen that the variance of the δ_i's is significantly larger for the fixed rank model. In Table 4, we present the p-value of the two-sample paired t-test for each statistic for the two methods of linking.

Table 4: *Table of p-value of the paired t-test for mode of cond. density model and for the fixed rank model.*

Method	M	I	D	$Gini$	M_{20}	C	A	$sizes$	M_*	SFR	$sizes$
Mode	0.99	0.91	0.29	0.78	0.61	0.88	0.73	0.79	0.12	0.78	0.08
Fixed rank	0.66	0.90	0.65	0.45	0.75	0.94	0.82	0.52	0.96	0.54	0.72

Thus, there is no clear evidence that $\mathbb{E}[\delta_i]$ is different from zero. Hence we do not reject the null.

6.1.2 Compare k-nearest neighbor with remaining galaxies

Next we want to know whether the morphology statistics of galaxies in the neighborhood of the predicted galaxy are more similar to the true progenitor galaxy compared to the remaining galaxies. For this we choose and fix a galaxy and we look at k-nearest neighbor in rank at $z = 2$ for the predicted galaxy. Then we calculate δ_i same as above for all the galaxies and the true progenitor for each statistic. Ultimately we test whether the k nearest neighbor galaxies have δ_i significantly lower than those of the remaining galaxies. In Figure 8 we present the boxplots of the absolute values of the δ_i's for

1. linking using mode of conditional density calculated from random forest; and
2. linking assuming the rank at $z = 2$ is the same as that at $z = 1$

Formally, we look at the following (difference between true progenitor and $galaxy_i$) for each $galaxy_i$ in the test set

$$\delta_i = |statistic_{galaxy_{true},z=2} - statistic_{galaxy_i,z=2}|.$$

Then for a given $galaxy_j$ we define the following two sets :

$Set_k = \{galaxy_i : galaxy_i$ is among the k nearest neighbors of $galaxy_j$ in terms of $R_2\}$
$Set_k^C =$ Test set $- Set_k$

Below is the formal hypothesis test for the above mentioned problem. Under the alternative H_1, we can say that the linking method is consistent:

$H_0 : E[\{\delta_i : galaxy_i \in Set_k\}] = E\left[\{\delta_i : galaxy_i \in Set_k^C\}\right]$.
$H_1 : E[\{\delta_i : galaxy_i \in Set_k\}] < E\left[\{\delta_i : galaxy_i \in Set_k^C\}\right]$.

In Figure 8, we present boxplots of $delta_i$ in Set_k and Set_k^C for all the galaxies in the test set accumulated together for different values of k. We can see in Figure 8, that there is not much difference in the boxplots for the eight morphology statistics. But for $mass$, SFR and for $rank$ the within-knn ball δ_i has lower median compared to the outside-knn ball δ_i for both the linking methods. But for the mode of conditional density linking method, the performance seems slightly better because the difference between the within-ball and outside-ball medians is a little larger compared to the fixed rank method.

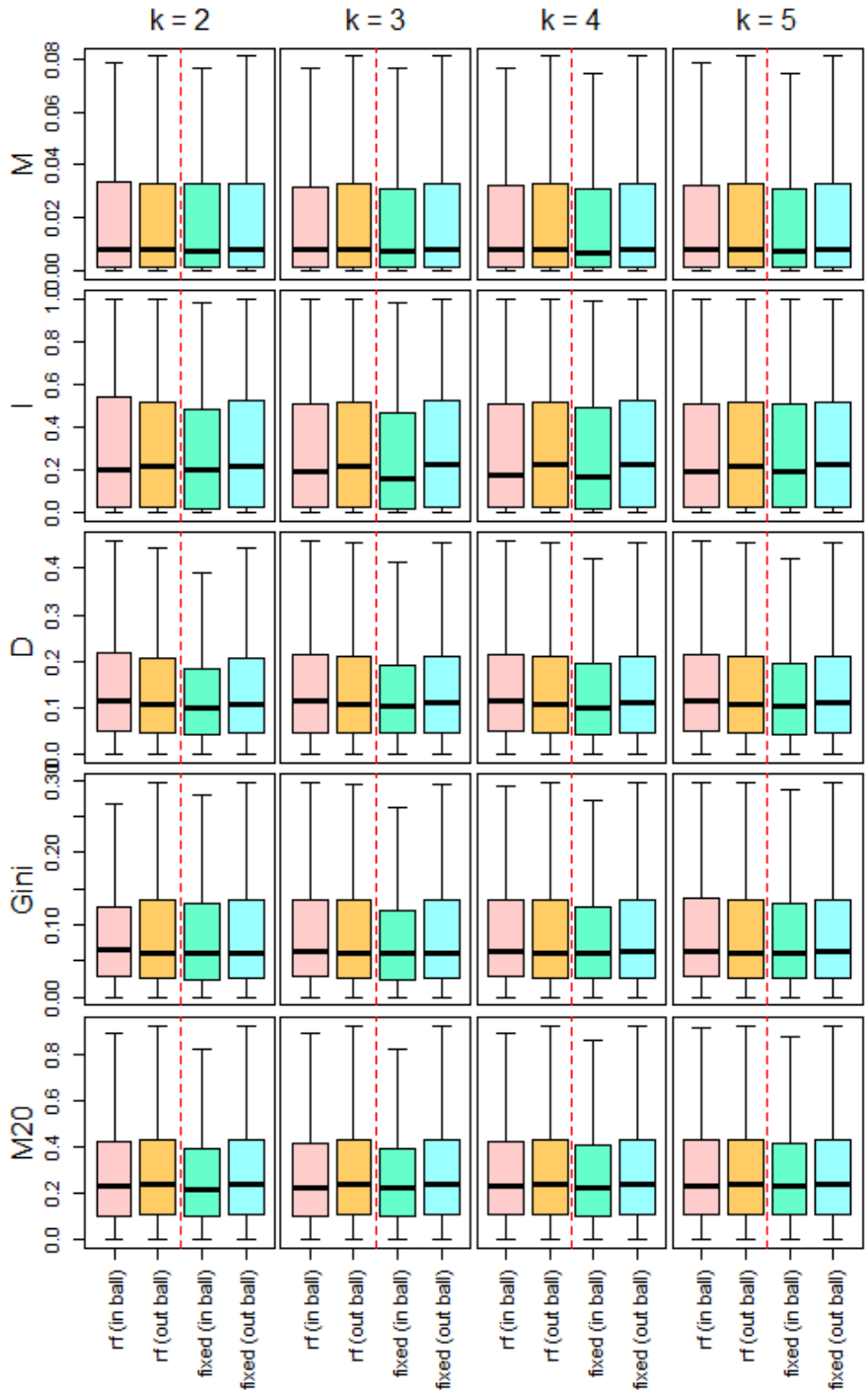


Figure 8: *Boxplot of δ_i's within the k-nearest neighbor (kNN) ball versus outside the ball, for both the RF-CDE model and the fixed-rank model. We choose k to be* 2, 3, 4 *and* 5. *The result does not seem to depend much on k except for the difference between the medians. The results are consistent across the k's.*

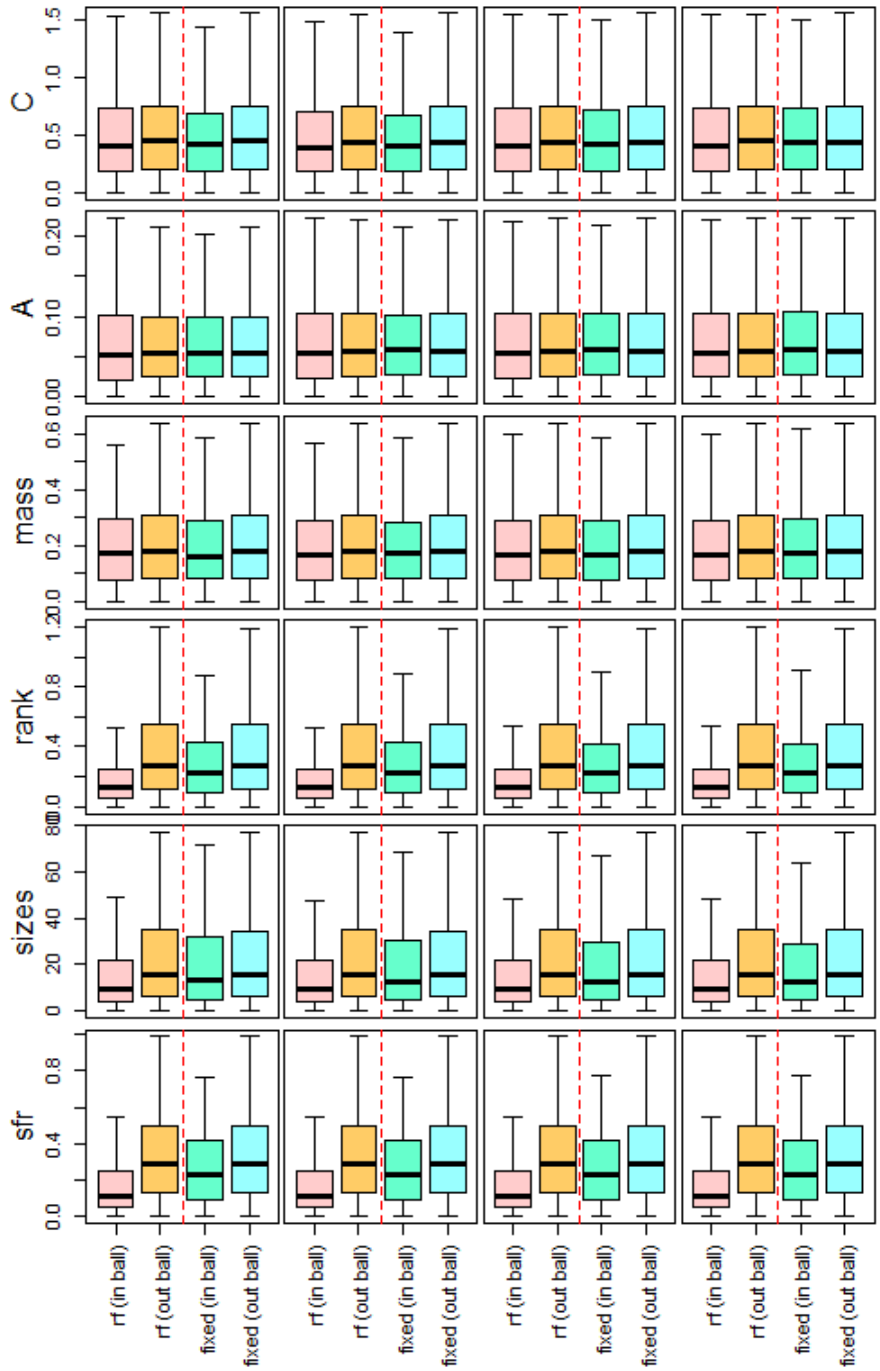


Figure 8: *Boxplot of δ_i's within the k-nearest neighbor (kNN) ball versus outside the ball, for both the RF-CDE model and the fixed-rank model. We choose k to be 2, 3, 4 and 5. The result does not seem to depend much on k except for the difference between the medians. The results are consistent across the k's.*

For each galaxy in the test set, we performed the hypothesis test described above with two-sample paired t-test. In Figure 9, we have the boxplots of the p-values of all galaxies cumulated together for each statistic for the two linking methods.

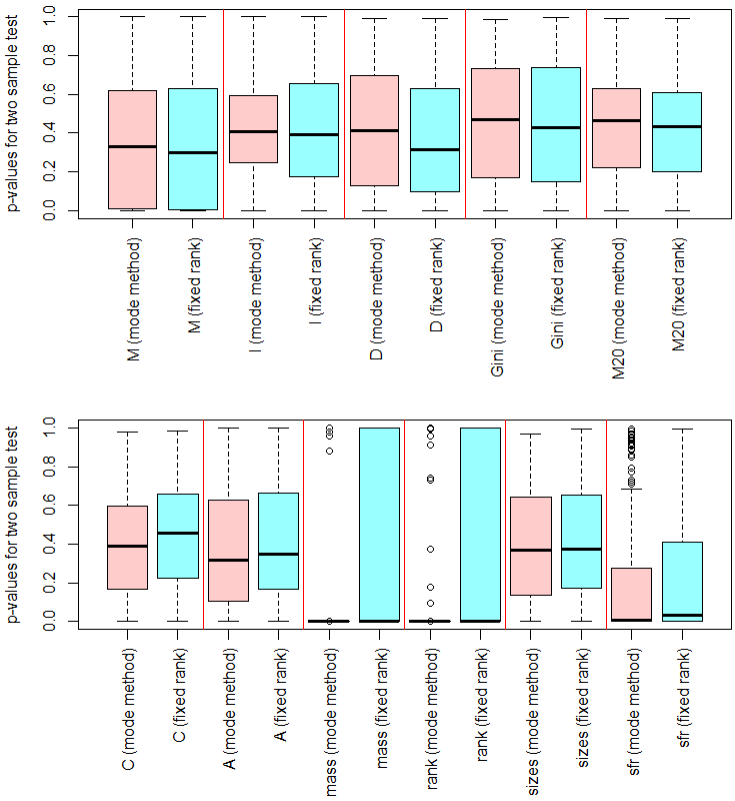


Figure 9: *Boxplot of p-values of two-sample paired t-test for each statistic for the two methods: mode of conditional density estimate and fixed rank method.*

We can see that for statistics M, I, D, $Gini$ and M_{20} there is not enough evidence to reject the null for mode of conditional density estimate compared to the fixed rank method. But for the remaining statistics the result is the opposite. We can see that for the eight morphology statistics, the median p-value is quite high for both the linking methods indicating that there is not enough evidence to reject the null.

Next, we will apply the linking method involving the conditional density on the real data in the following section.

6.2 Checking performance of linking method with real data

The real data comes from HST/CANDELS (Appendix 0) . We have galaxies from the three redshifts $z = 1, 1.5, 2$. It is in almost the same format as is the of the simulated data except

there is no camera angle here. We will apply the linking method on the galaxies from $z = 1$ and attempt to assign them a progenitor from the set of galaxies at $z = 2$.

Table 5: *Number of galaxies at each redshift in the real data set*

redshift (z)	1	2
Number of galaxies	2143	2292

We will be applying the linking method described at the beginning of section 6. Hence we will check the similarity between the real data and the simulated data at $z = 1, 2$ in Figure 10 using violin plots. The violin plots have comparable shape for all the morphological statistics and the rank. But *mass* and SFR show significant difference. We will continue with applying the linking method on the real data.

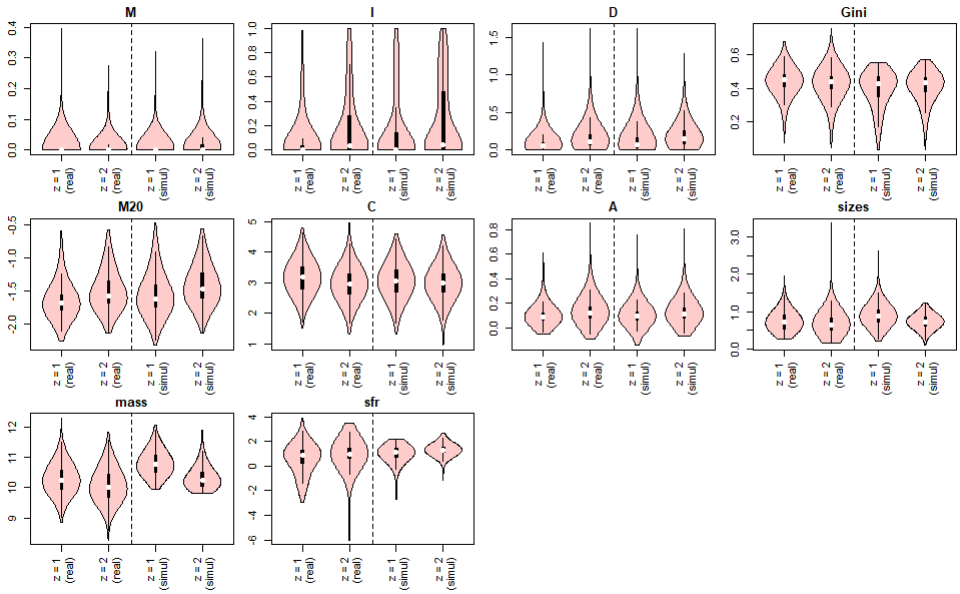


Figure 10: *Violin plots of the statistics at $z = 1, 2$ in the true real and the simulated datasets. The violin plots for rank are all the same by construction. Violin plots of mass and SFR in the real dataset seem to differ from those in the simulated dataset.*

Real galaxy data does not have the information about individual true progenitor mapping of the galaxies. Hence we cannot use common error rates to measure the performance of our linking method. Thus we will compare the whole set of estimated progenitor galaxies with the whole set of true progenitor galaxies. We use violin plots of all the statistics from the true and the estimated populations at $z = 2$ in Figure 11. It can be seen that the violin

plots for *rank* and *mass* are different for the true and estimated population. We will test the similarity between the two populations more formally.

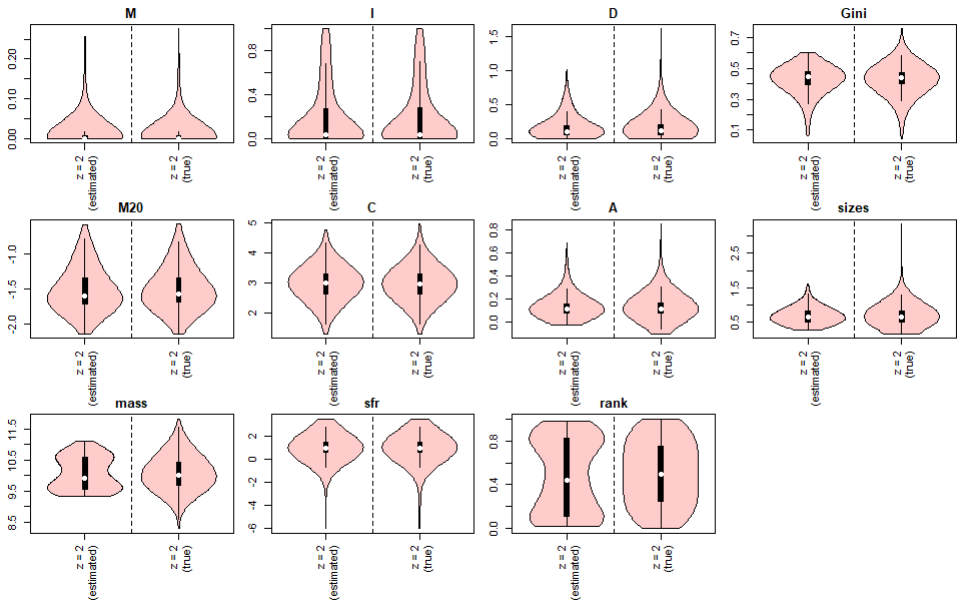


Figure 11: *Violin plots of the statistics at $z = 2$ in the true vs in the estimated progenitor population. The plots in the true and the estimated cases seem to match for most of the variables except mass and rank.*

We use Kolmogorov-Smirnov statistic to test whether the statistics in the true progenitor set and those in the estimated progenitor set come from the same distribution. Let $F_{statistic\, i,\, true}$ be the cumulative distribution function of statistic i from the true progenitor population. Similarly we define for the estimated population. Formally, we can write the following hypothesis:-

Kolmogorov-Smirnov Test

$F_{stat\, i,\, true}(x) = P_{true}(stat\, i \leq x)$

$\hat{F}_{stat\, i,\, true}(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(X_{ij,true} \leq x)$

$\hat{F}_{stat\, i,\, estimated}(x) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(X_{ij,estimated} \leq x)$

$H_0 : F_{stat\, i,\, true} = F_{stat\, i,\, estimated}$ vs $H_1 : F_{stat\, i,\, true} \neq F_{stat\, i,\, estimated}$

Test statistic : $D = sup_x |\hat{F}_{stat\, i,true}(x) - \hat{F}_{stat\, i,estimated}(x)| \sim$ TBD.

The Kolmogorov-Smirnov test has been repeated 100 times with random subsets of the galaxies from both the true and the estimated progenitor population of sizes $75 - 90\%$ independently. The p-value of the tests has been shown in Figure 12.

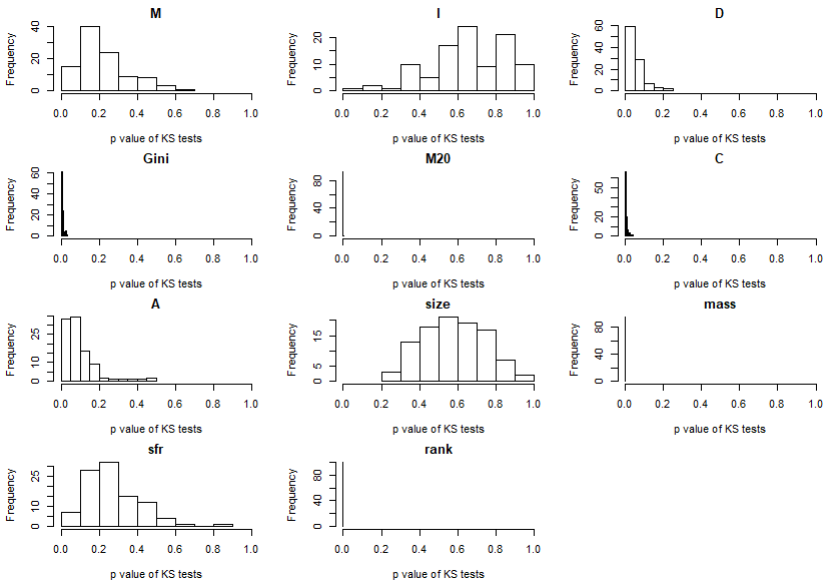


Figure 12: *Histogram of the p-values of Kolmogorov-Smirnov test performed 100 times for the eleven statistics.*

In Figure 12, we can see that the p-values have a left skewed distribution mainly for all the statistics except *intensity*, *size*, M_{20}, *mass* and *rank*, where the later three follow degenerate distribution at 0. For a clearer picture, we present the same result in boxplot below in Figure 13. It can be seen that the median value lies above the 0.05 line only for M, I, A, *sizes* and SFR. The rest of the variables except D have median less than 0.01. Thus the true and estimated progenitor population have similar distribution for M, I, A, *sizes* and SFR.

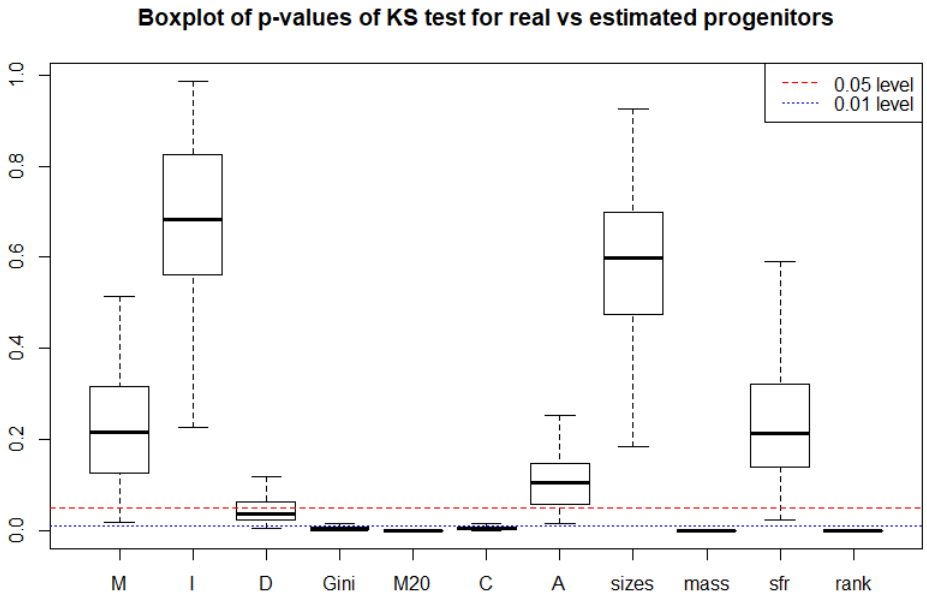


Figure 13: *Boxplot of the p-values from the Kolmogorov-Smirnov tests performed 100 times for the different statistics.*

We cannot compare the linking method of using fixed rank model with the mode of conditional density estimate by using Kolmogorov-Smirnov test. Because linking with fixed rank model does not differentiate between the true and the estimated progenitor population as a whole. Hence to check consistency with performance of simulated data, we look at the change in summary statistics between $z = 2$ and $z = 1$ for individual galaxies using their predicted progenitors at $z = 2$. We compute $statistic_{z=2,\,i\,(predicted)} - statistic_{z=1,\,i}$ for each $galaxy\,i$ for the eleven summary statistics. In Figure 14, we plot the mean of the differences computed using the mode of conditional density method and the fixed rank method along with the mean of the difference computed using true progenitors in simulated data. We see that the estimated differences for the real and the simulated data are similar for most of the variables except $mass$, SFR and $rank$. Moreover all these four differences vary from the true difference for C, $sizes$, $mass$, SFR and $rank$. For real data difference in fixed rank method is lower then the mode of conditional density method for $mass$, SFR and $rank$.

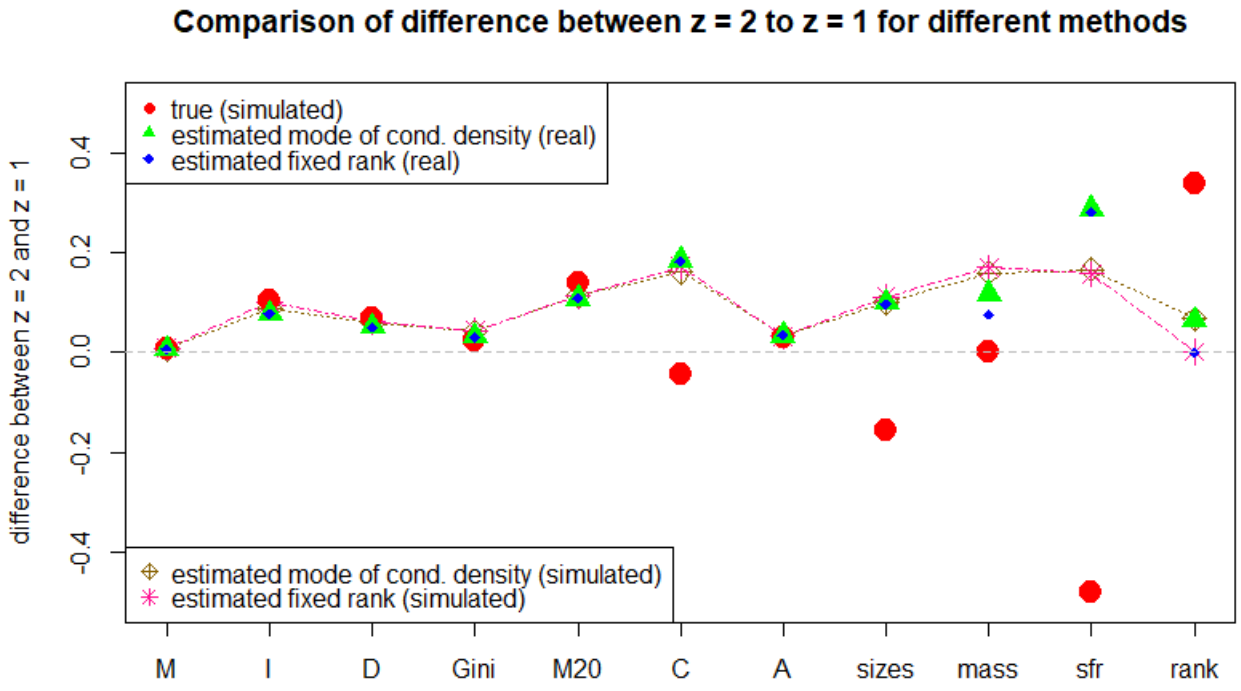


Figure 14: *The mean of differences for mode of conditional density method (green) and for fixed rank (blue) method* ($statistic_{z=2,\,i\,(predicted)} - statistic_{z=1,\,i}$) *where the progenitor is predicted using one of the two afore mentioned methods. The points in red are the mean differences for the true progenitors calculated from the simulated data. For checking consistency I also plot the same for the simulated data in gark golden and pink color for the two prediction methods.*

7 Summary and Conclusions

We wanted to test whether linking galaxies without assuming fixed rank gives better results compared to when we assume fixed rank for galaxies at $z = 1$ and $z = 2$. Summarizing what we learned from the analyses in the preceding sections, first we saw that galaxies are most likely to maintain their rank value or change by just a fraction as seen in Figure 2. The morphological statistics I, D, M_{20} and A are positively correlated and these four attributes are negatively correlated with $Gini$ and C. Mass is weakly positively correlated with I, D, M_{20} and SFR and negatively correlated with $Gini$ and C. In Appendix B.2, we determine via visual measures that M_* and SFR at z=1 are most informative variables for predicting R_2, while morphological statistics are less informative. In Section 5.1, we fit a random forest model to predict R_2 using data at z=1 whose results echo those of Appendix B.2: *mass*,

R_1, and SFR for $z = 1$ are most important in estimation of R_2. The model has better error rates compared to baseline model where the mass rank is fixed across redshift. Morphologies help in the estimation to some extent. Mode of conditional density can also be used as an estimator for R_2, but the error is higher than the situation when we used the aggregated predicted value from the random forest. In Section C in the appendix, we repeat the steps of Section 5 to predict $R_{1.5}$ using data at $z = 1$. The result is consistent with that in Section 5. In Section D in the appendix, we perform two predictions with the direction reversed. The results are consistent with that of Section 5 with some differences. Star formation rate becomes the most important when we estimate R_1 using $z = 2$.In Section 6 we describe the linking method using random forest model and compare it with linking assuming fixed rank. The linking with the random forest method is slightly better compared to the linking method using fixed rank assumption. Also, we apply the two methods on the real data from CANDELS. The performance of random forest on real data is not much different compared to the method of fixed rank.

Thus, we conclude that relaxing the assumption of constant comoving number density (fixed rank) and including morphological statistics in the analyses gives better linking results for the simulated data to link galaxies at $z = 1$ with their progenitors at $z = 2$. But we do not see such improvement when we apply our model on the real galaxy data. This may be a consequence of the fact that the real and the observed data do not come from the same distribution. The simulated data is an approximation of the real data but there are still systematic errors. Our linking method used the estimation model of just rank; i.e. we do not use the information available at $z = 2$ except the *rank* (or *mass* equivalently). Including the other statistics may help in building a robust model to nullify the effect of systematic errors.

Appendix

0 HST/CANDELs data

A part of our project is based on observations taken by the CANDELS[8] Multi-Cycle Treasury Program with the NASA/ESA HST, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. For each galaxy, we have a conditional density estimate for redshift (z) (dubbed $f(z|x)$, where x are the predictor variables entering into the estimation, such as galaxy magnitudes and colors). These conditional density estimates are summed over (calling the sum $p(z)$) some range of redshifts from which x is (are) used and to see if the sum attains some given threshold ($p(z) > threshold$). The threshold is used to ensure that the sample size is comparable across redshifts and matches that of the simulated data. For $z = 2$, the domain of summation is 1.813 to 2.187 and the threshold is 0.6. For $z = 1$, the domain is 0.85 to 1.15 and the threshold is 0.75.

A Summary Statistics

1. **Multimode** (M) statistic (Freeman et al. 2013)
   The M statistic identifies galaxies with disturbed morphologies. Consider an intensity quantile q_l such that a proportion l of the pixel intensities i_{mn} are smaller than q_l. (Here mn denotes pixel coordinates.) For a given q_l, define a new indicator variable j_{mn} such that

$$j_{mn} = \begin{cases} 1 & \text{if } i_{mn} \geq q_l \\ 0 & \text{otherwise} \end{cases} .$$

   Within the image j_{mn}, we obtain the areas of largest and second-largest groups of contiguous pixels, which we denote $A_{l,(1)}$ and $A_{l,(2)}$ respectively. We define the area ratio as

$$R_l = \frac{A_{l,(2)}^2}{A_{l,(1)} n_{seg}} ,$$

   where n_{seg} is the number of pixels in the segmentation map, i.e., the mask used to define the extent of the galaxy within the image. This formulation imposes a strict upper limit on R_1 of 1/2 that is achieved if $A_{l,(1)} = A_{l,(2)} = n_{seg}/2$. The M statistic is the maximum observed value of R_l over all quantiles l:

$$M = \max_l R_l .$$

[8] HST/CANDELs

2. **Intensity** (I) statistic (Freeman et al. 2013)
   One of the shortcomings of the M statistic is that it does not consider the summed intensity within contiguous pixel groups. For instance, a contiguous group with a large number of pixels may have a smaller summed intensity than other, smaller groups of pixels. To mitigate this shortcoming we utilize the I statistic. We associate each pixel mn with a local maximum $mn_{\max}$ by following the maximum gradient ascent path. All pixels that are associated with a given local maximum $mn_{\max}$ are grouped together, and for each group, we sum the pixel intensities i_{mn}. (Note that in a data pre-processing step, we smooth the image data with a symmetric Gaussian kernel with $\sigma \sim 1$ pixel, to decrease the effect that pixel noise has on the construction of pixel groups.) We rank the summed intensities in descending order and use the first and second sorted values to compute the I statistic:

$$I = \frac{I_{(2)}}{I_{(1)}} \, .$$

3. **Deviation** (D) statistic (Freeman et al. 2013)
   The deviation D statistic is used to capture evidence of galaxy asymmetry. It is the distance from the local maximum associated with $I_{(1)}$ to the galaxy's center of mass:

$$(m_{\rm cen}, n_{\rm cen}) = \left(\frac{1}{n_{seg}} \sum_m \sum_n m i_{mn}, \frac{1}{n_{seg}} \sum_m \sum_n n i_{mn} \right) \, , \tag{1}$$

   where the summation is over the n_{seg} pixels within the segmentation map. The D statistic is:

$$D = \sqrt{(m_{\rm cen} - m_{I_{(1)}})^2 + (n_{\rm cen} - n_{I_{(1)}})^2} / \sqrt{n_{seg}/\pi} \, .$$

   where the normalizing factor $\sqrt{n_{seg}/\pi}$ is a galaxy radius estimate achieved by assuming that the segmentation map is circular.

4. **Gini** ($Gini$) statistic (Lotz et al. 2004)
   The Gini coefficient measures the relative distribution of pixel intensities within the segmentation map: $G = 0$ means that the intensities are uniform across the galaxy, while $G = 1$ means that all of a galaxy's light falls into a single pixel. The $Gini$ statistic is defined as

$$G = \frac{1}{\bar{i} n_{\rm seg}(n_{\rm seg} - 1)} \sum_k (2k - n_{\rm seg} - 1) i_{m_{(k)} n_{(k)}} \, ,$$

   where $\bar{i}$ is the sample mean of all intensities within the segmentation map and $m_{(k)} n_{(k)}$ denotes the coordinates of the pixel with the $k^{\rm th}$-smallest intensity value.

5. $\mathbf{M_{20}}$ statistic (Lotz et al. 2004)
   M_{20} describes the spatial distribution of pixel intensities. First, we compute a total second-order moment:

$$M_{\rm tot} = \sum_m \sum_n i_{mn} \left[(m - m_{\rm cen})^2 + (n - n_{\rm cen})^2 \right]$$

   where $m_{\rm cen}$ and $n_{\rm cen}$ are the coordinates of the galaxy's center of mass (equation 1) and the summation in done over all pixels mn within the segmentation map. We then repeat the summation done above using only the brightest 20% of the pixels; we call this sum $M_{\rm bright}$. Then M_{20} is

$$M_{20} = \log_{10} \left(\frac{M_{\rm bright}}{M_{\rm tot}} \right) .$$

6. **Concentration** (C) statistic (Conselice 2003)
   The concentration statistic encapsulates the area over which the bulk of a galaxy's summed intensity lies. Its calculation assumes circular symmetry. At a given radius *sizes* from the galaxy's center, we define two quantities: the summed intensity within the annulus defined by *sizes* and $r + dr$, and the overall average summed intensity:

$$\begin{aligned} \mu(r) &= \frac{\int_0^{2\pi} \int_{r-\delta r}^{r+\delta r} i(r', \theta) r' dr' d\theta}{\int_0^{2\pi} \int_{r-\delta r}^{r+\delta r} r' dr' d\theta} \\ \bar{\mu}(r) &= \frac{\int_0^{2\pi} \int_0^{r+\delta r} i(r', \theta) r' dr' d\theta}{\int_0^{2\pi} \int_0^{r+\delta r} r' dr' d\theta}. \end{aligned}$$

   (We show the calculations as integrals for conceptual clarity, but the actual calculations are done as sums over image pixels.) *sizes* is the solution of the equation $\mu(r)/\bar{\mu}(r) = \epsilon$, where ϵ is commonly chosen to be 0.2. We compute the total summed intensity within the radius *sizes*, then determine the smaller radii within which there are 20% and 80% of that total summed intensity. The C statistic is:

$$C = 5 \times \log \left(r_{80\%} / r_{20\%} \right) .$$

   The smaller $r_{20\%}$ is relative to $r_{80\%}$, the higher the value of C, as the galaxy will appear "more concentrated."

7. **Asymmetry** (A) statistic (Conselice 2003)
   The A statistic is a measure of how asymmetric a galaxy is after its image is rotated 180° the central pixel and then subtracted from the original image. For an asymmetric galaxy, the difference image will exhibit significant residual structures, leading the A statistic to differ significantly from zero. The A statistic is defined as

$$A = \frac{\sum_m \sum_n |i_{mn} - i_{180,mn}|}{\sum_m \sum_n |i_{mn}|} - B_{180} ,$$

where i and i_{180} are the pixel intensities in the original and rotated images respectively and B_{180} is the average background asymmetry, defined using the intensities of pixels lying outside the segmentation map.

B Rank Group Analysis

B.1 Exploratory Data Analysis

Here, we divide the galaxies into five quintile groups (see Figure 15). In Figure 16, we show the transition from one rank group to another over redshift. Those with values $\tilde{R}_1$=2-4 have nearly equal probability (≈ 0.3) to remain in their own rank group. And for all values of $\tilde{R}_1$ the probability of transition decreases as the distance between the rank groups at z=2 increases.

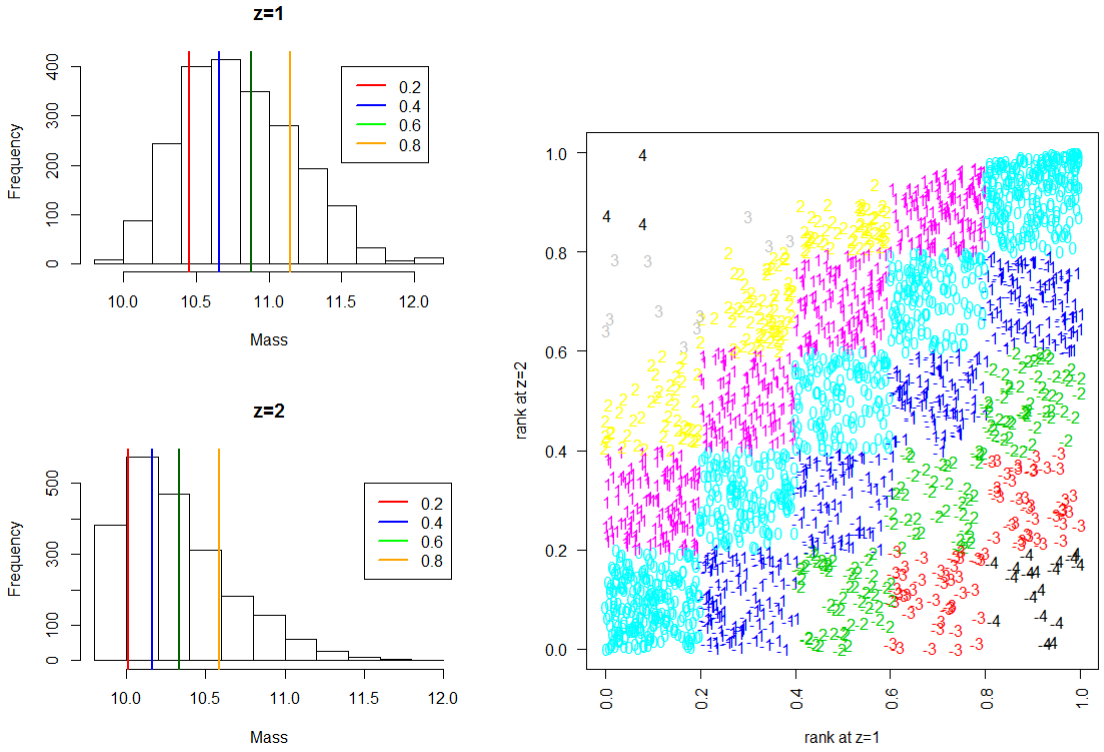


Figure 15: Left: *Division of data based on quantiles, as shown using histograms of mass M_* at z=1 and 2. The vertical lines mark the quantiles 20%, 40%, 60% & 80% and partition the masses into five groups (groups 1 through 5 from left to right).* Right: *The scatter plot in Figure 2, showing the rank change for each galaxy.*

Table 6: *Change of rank from z=1 to z=2*

rank	z=2					sum
	1	**2**	**3**	**4**	**5**	
1	232	138	48	8	3	429
2	89	145	124	68	3	429
3	49	70	125	132	52	428
4	41	38	79	131	140	429
5	18	38	52	90	231	429

(a) *Numbers of galaxies*

rank	z=2					sum
	1	**2**	**3**	**4**	**5**	
1	0.54	0.32	0.11	0.02	0.01	1.00
2	0.21	0.34	0.29	0.16	0.01	1.00
3	0.11	0.16	0.29	0.31	0.12	1.00
4	0.10	0.09	0.18	0.31	0.33	1.00
5	0.04	0.09	0.12	0.21	0.54	1.00

(b) *Transition probabilities*

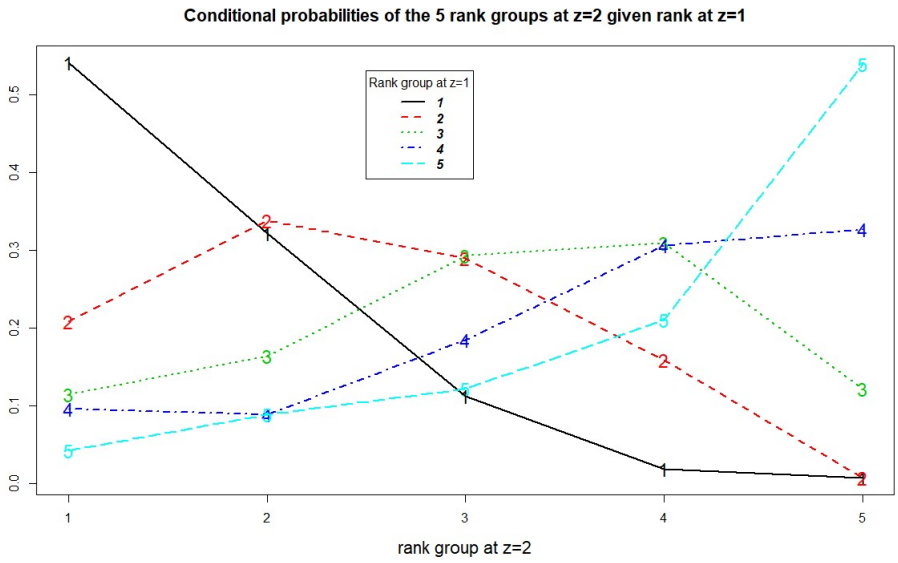


Figure 16: *Probability for each rank group at z=2 conditioned on the rank group at z=1.*

To the left in Table 6 we show the number of galaxies for each pair of ranks. The numbers in the diagonal boxes are the galaxies that stay in the same rank group. To the right, we take the total number of observations and divide them the total number of galaxies in each row to determine the proportion of galaxies that change rank as we move from z=1 to z=2. This part of the table indicates that galaxies are most likely to be in the same rank group at both redshifts.

We present the same information as is to the right in Table 6 pictorially in Figure 16. Galaxies which have $\tilde{R}_1$=1 and 5 have similar, though mirrored, transitional probabilities.

B.2 Morphology comparison at z=1 across $\tilde{R}_1$ and $\tilde{R}_2$

In Figure17, we look at the galaxy morphologies at the five rank groups at z=1 and their change at z=2. However, here we group the galaxies into three rank groups as defined in Table 7. The galaxies in $\tilde{R}_1$=1 can only increase in rank group, while the galaxies in $\tilde{R}_1$=5 can only decrease in rank group. So we look at galaxies with $\tilde{R}_1$=1 and 5 separately from each other and from those with $\tilde{R}_1$=2-4.

Table 7: *Rank groups at z=1 for morphology comparison*

$\tilde{R}'_1$	$\tilde{R}_1$	$\tilde{R}_2 - \tilde{R}_1$
1	1	0, 1,..., 4
2	2, 3, 4	-3, -2, ..., 3
3	5	-4, -3,..., 0

M and I are close to zero for galaxies with stable morphology. So here we look at the indicator variable $\mathbb{I}_M$ and $\mathbb{I}_I$, which are one if M and I at z=1 are close to zero (**DEFINE CLOSE**) and zero otherwise.

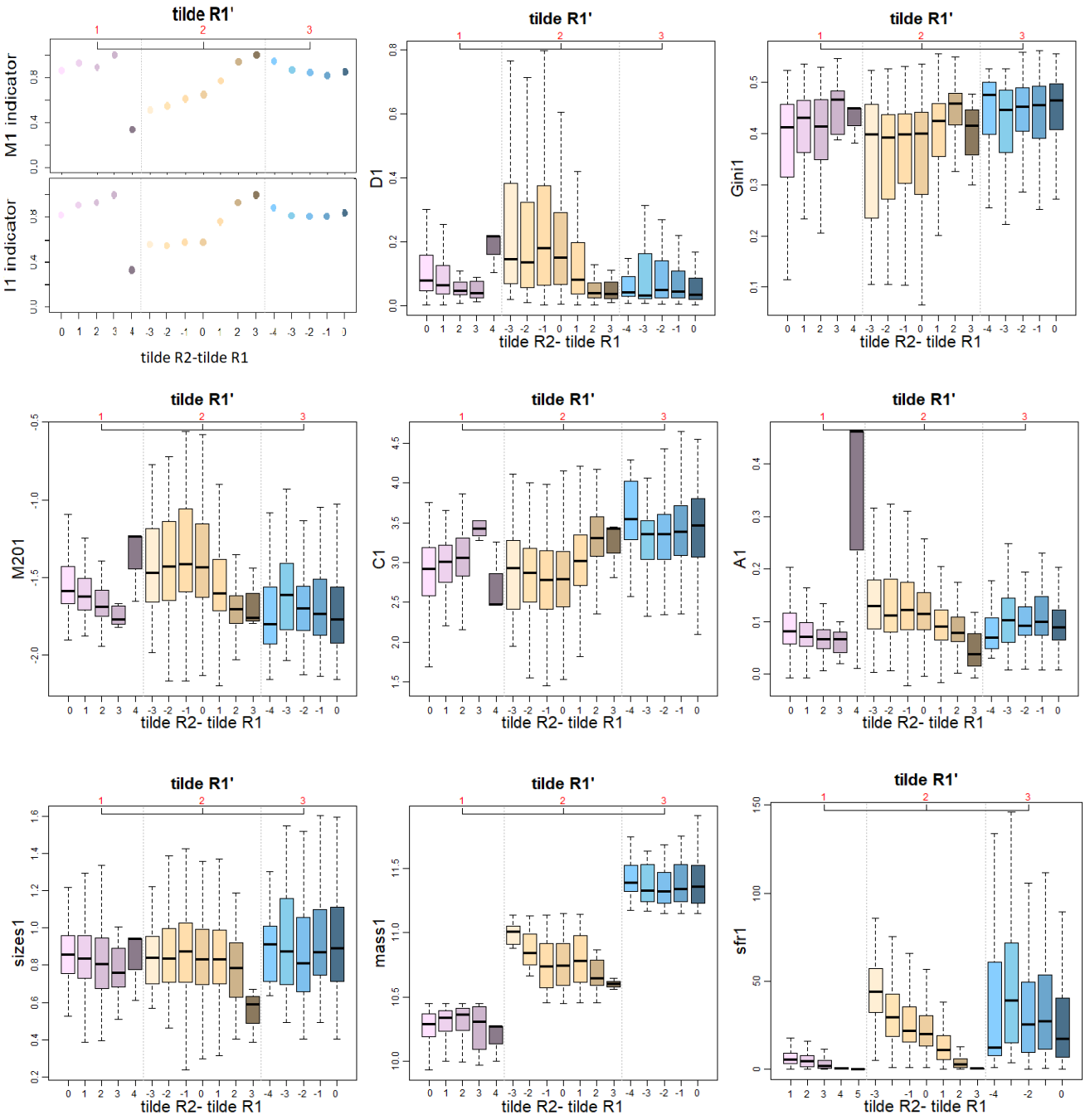


Figure 17: Upper Left: *Proportion of galaxies that have stable morphologies at z=1 for each rank change* $\tilde{R}_2 - \tilde{R}_1$ *at each* $\tilde{R}'_1$. *The three partitions correspond to the three* $\tilde{R}'_1$*'s and the columns within each box correspond to rank group changes* $\tilde{R}_2 - \tilde{R}_1$. Other Panels: *Boxplots of the other variables at z=1.*

We observe in the upper left panel of Figure 17 that the proportion of stable galaxies increases

across $\tilde{R}_2 - \tilde{R}_1$ for $\tilde{R}'_1$=2. This is consistent with the idea that galaxies that have undergone mergers, which would lie towards the left, are less disturbed. (We note that the proportion is very low for galaxies that have $\tilde{R}_1 = 1$ and $\tilde{R}_2 = 5$. These galaxies were noted above as anomalous and thus one should refrain from over-interpreting this result.)

As for the morphological statistics, we observe that D, M_{20} and A at first increase with $\tilde{R}_2 - \tilde{R}_1$ and then decrease for $\tilde{R}'_1$ =2 and 3. The exact opposite relation is seen for $Gini$ and C. However M_{20} and C have more prominent negative and positive correlation respectively with $\tilde{R}_2 - \tilde{R}_1$ for $\tilde{R}'_1$ =1. Hence the information from D and A are contained in M_{20}. And C accounts for $Gini$.

The stellar masses M_* increase with $\tilde{R}_1$ as expected, and we see that for the galaxies in $\tilde{R}'_1$=2, M_* is negatively correlated with $\tilde{R}_2 - \tilde{R}_1$, which would be consistent with merger activity.

The star-formation rates SFR show a rough decrease with $\tilde{R}_2$ for all groups of $\tilde{R}'_1$. As seen before, SFR is correlated with M_* and R_1. But as seen in Appendix A.2, it contains additional information about the pairwise relationship between galaxy ranks at z=1 and z=2.

In the next section we move on to prediction and check for consistency.

B.3 Discretized rank group in prediction of rank at $z = 2$ using data at $z = 1$

We compute $\hat{\tilde{R}}_2$ for the galaxies in the test set using the $\hat{R}_2$ for each of the five random forests in the cross validation. In Table 8 we display the confusion matrix. This is just for interpretation of the model fitting.

Table 8: *Confusion matrix for prediction of* $\tilde{R}_2$

	True $\tilde{R}_2$				
Predicted $\tilde{R}_2$	**1**	**2**	**3**	**4**	**5**
1	164	60	6	0	0
2	156	205	131	21	6
3	73	113	203	200	43
4	31	43	70	184	296
5	5	8	18	24	84

To quantify the performance for $\tilde{R}_2$ we look at the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), which are defined in Table 9. Note that "Group 0" is the group under consideration, while all other groups are combined into "Others."

Table 9: *Definition of sensitivity, specificity, PPV and NPV*

		Actual quintile group			
		Group 0	Others	Total	$Sensitivity = \frac{TP}{TP+FN}$
Predicted quintile group	Group 0	TP	FP	$TP+FP$	$Specificity = \frac{TN}{FP+TN}$
	Others	FN	TN	$FN+TN$	$PPV = \frac{TP}{TP+FP}$
	Total	$TP+FN$	$FP+TN$	N	$NPV = \frac{TN}{FN+TN}$

In Table 10 we list the results for the five values of $\tilde{R}_2$. The quintile classes $\tilde{R}_2$=1 and 5 have similar measures, as do classes $\tilde{R}_2$=2-4.

Table 10: *Performance statistics for all the five rank groups in the random forest model*

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.38	0.48	0.47	0.43	0.20
Specificity	0.96	0.82	0.75	0.74	0.97
Pos Pred Value	0.71	0.39	0.32	0.29	0.60
Neg Pred Value	0.86	0.86	0.85	0.84	0.83

C Predicting mass rank at z=1.5 given data at z=1

In Figure 18 we display the variable importance.

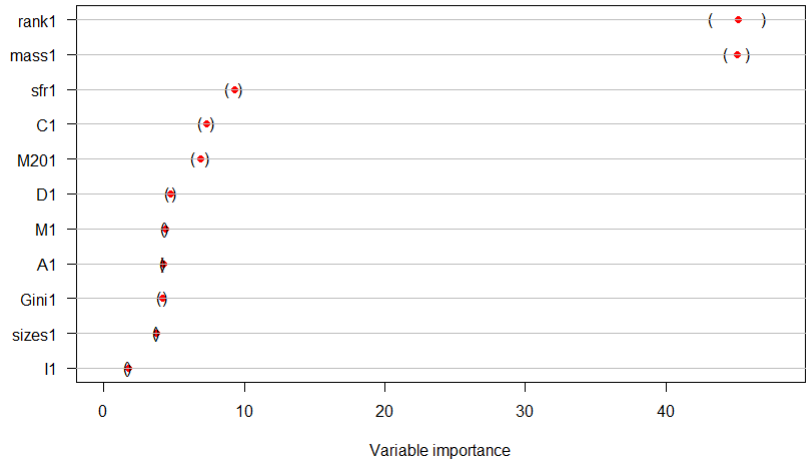
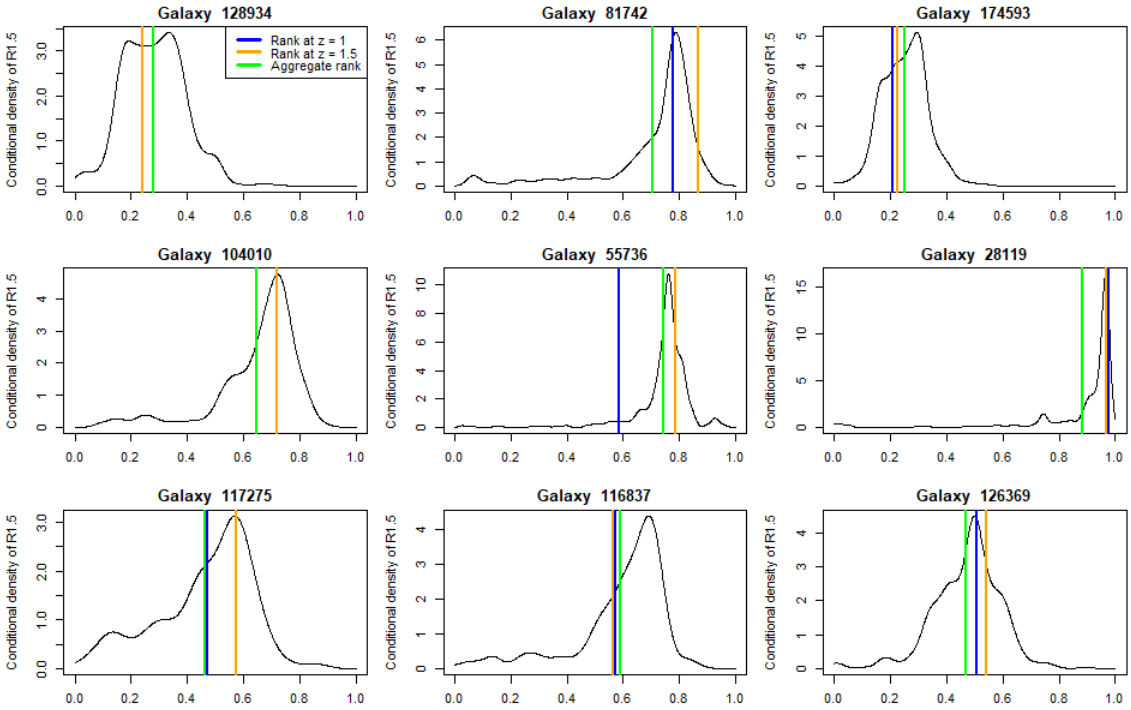


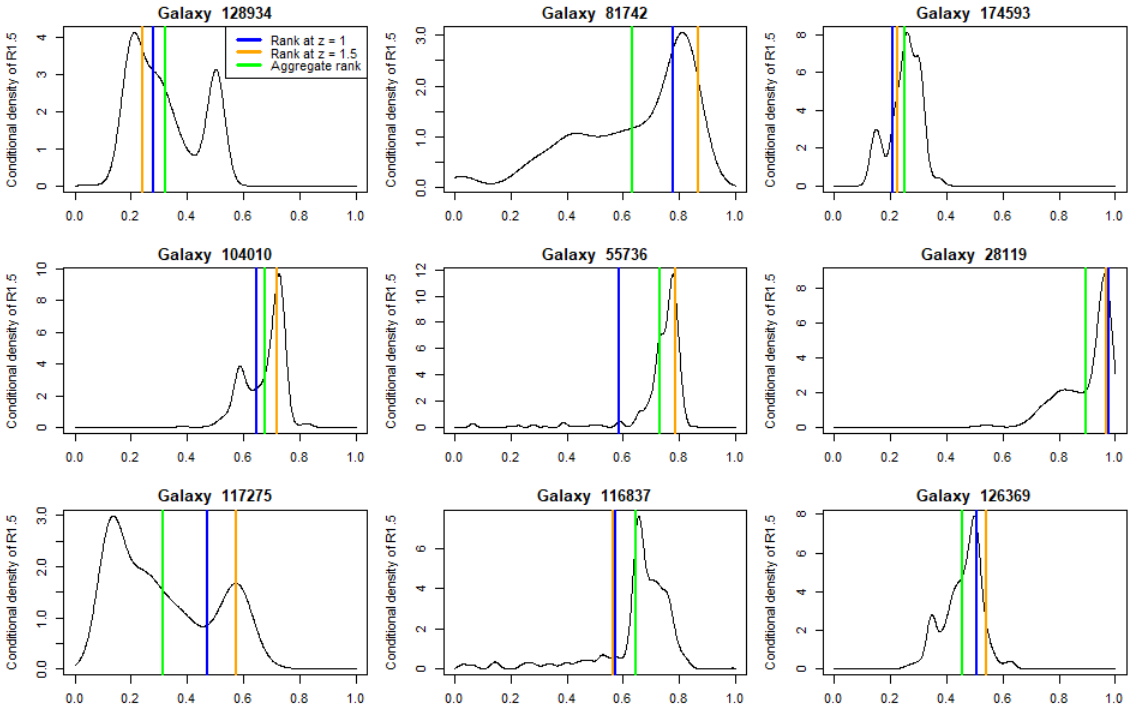
Figure 18: *Variable importance plot for $R_{1.5}$ prediction using statistics at z=1. The parentheses show the extent of variability calculated by combining the ten random forests.*

For this prediction problem, R_1 and M_* are the most important variables but the importance of SFR is reduced compared to the original prediction problem in Figure 5.

In Figure 19, we have the density estimated using all the statistics at $z = 1$ for finding $R_{1.5}$. Same as in Figure 6, it can be seen that the mode coincides with $R_{1.5}$ in Figure 19 (a) and coincides with $R_{1.5}$ in most cases in (b).



(a) *Using all variables at $z = 1$ including morphologies.*



(b) *Using only mass, SFR and R_1*

Figure 19: *Predictive density of nine galaxies. The blue, the red and the green lines correspond to R_1, $R_{1.5}$ and the aggregated estimate of $R_{1.5}$ from the random forest.*

Table 11: *Errors of estimating rank $R_{1.5}$ using statistics at z=1*

Method	**Error**
Fixed rank	0.0310 ± 0.0001
Varying rank (with morphologies)	0.0259 ± 0.0001
Varying rank (without morphologies)	0.0281 ± 0.0002
Mode of cond. density (with morphologies)	0.0314 ± 0.0001
Mode of cond. density (without morphologies)	0.0344 ± 0.0002

The random forest model performs better compared to the baseline model in Table 11. As expected the error for this prediction problem is less than the original prediction problem because the gap between the two redshifts is smaller.

Looking at the performance of the fitted model, we can conclude that M_*, R_1 and SFR are the most important for rank prediction. The random forest model with cross validation performs better than the baseline model. Hence the result of the original prediction problem is consistent with that of the second prediction problem.

D Predicting Mass Rank at a Later Epoch

In this section we will reverse the direction of the prediction and compare with the original problem. We will consider two cases. First we predict R_1 using statistics at z=2, and then we predict $R_{1.5}$ using statistics at z=2.

D.1 Predicting mass rank at z=1 given data at z=2

In Figure 20, SFR seems most important. This is different from what is seen previously in Section 5: if the order of the prediction is reversed, then SFR is most important.

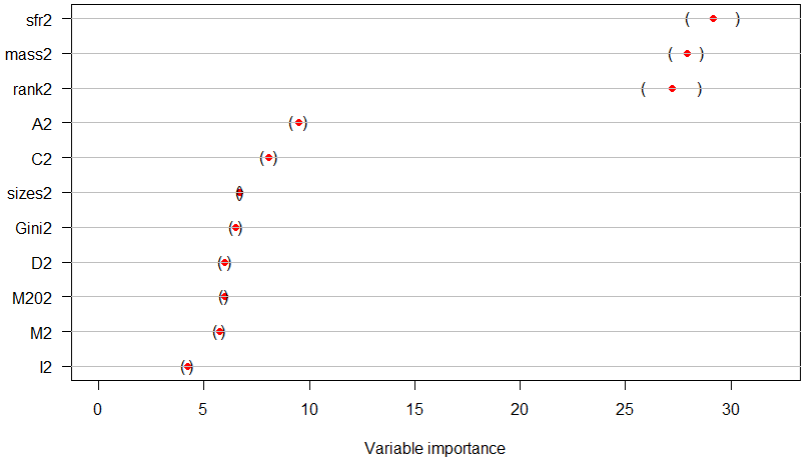


Figure 20: *Variable importance plot for R_1 prediction using statistics at z=2. The parentheses show the extent of variability calculated by combining the five random forests.*

Also the importance of M_{20} has reduced.

Table 12: *Estimating R_1 using statistics at z=2*

Method	**Error**
Fixed rank	0.0592 ± 0.0003
Varying rank (with morphologies)	0.0425 ± 0.0002
Varying rank (without morphologies)	0.0490 ± 0.0002
Mode of cond. density (with morphologies)	0.0574 ± 0.0002
Mode of cond. density (without morphologies)	0.0663 ± 0.0002

The error of the random forest model is better than that of the fixed rank model. Also the prediction error is comparable to that of the prediction problem in Section 5.1.
In Table 13 we display the confusion matrix of the prediction of R_1 using statistics at z=2.

Table 13: *$\tilde{R}_1$ using statistics at z=2*

	True $\tilde{R}_1$				
Predicted $\tilde{R}_1$	**1**	**2**	**3**	**4**	**5**
1	107	16	7	8	0
2	256	169	81	51	34
3	57	219	206	115	99
4	8	25	134	224	146
5	1	0	0	31	150

Table 14: *Performance statistics for all the five rank groups in the random forest model for estimating* $\tilde{R}_1$ *using* $z = 2$

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.25	0.39	0.48	0.52	0.35
Specificity	0.98	0.75	0.71	0.82	0.98
Pos Pred Value	0.78	0.29	0.30	0.42	0.82
Neg Pred Value	0.84	0.83	0.85	0.87	0.86

D.2 Predicting mass rank at z=1.5 given data at z=2

In Figure 21 we display the importance plot for prediction of $R_{1.5}$. This plot is very similar to the that in Figure 5.

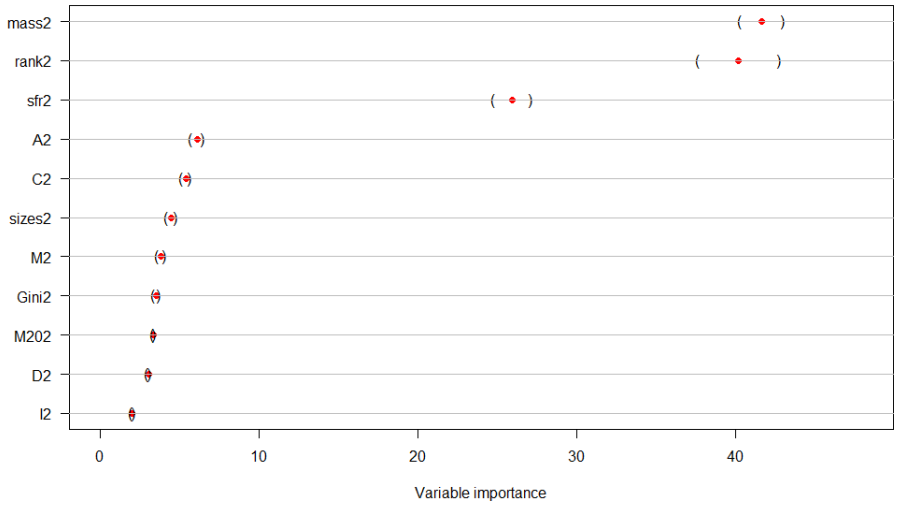


Figure 21: *Variable importance plot for* $R_{1.5}$ *prediction using statistics at z=2. The parentheses show the extent of variability calculated by combining the five random forests.*

In Table 15, the random forest model performed better than the fixed rank model. And the error is less than that of the original model in Table 3 and the model in Table 12 because the distance between the redshifts is less.

Table 15: *Errors of different methods estimating rank at z=1.5*

Method	**Error**
Fixed rank	0.0302 ± 0.0003
Varying rank (with morphologies)	0.0235 ± 0.0002
Varying rank (without morphologies)	0.0267 ± 0.0002
Mode of cond. density (with morphologies)	0.0280 ± 0.0002
Mode of cond. density (without morphologies)	0.0314 ± 0.0002

In Table 16 we display the confusion matrix and the prediction performance table for the model. The results seem similar to those of the previous models.

Table 16: *Confusion matrix for prediction of $\tilde{R}_{1.5}$ using statistics at $z = 2$*

	True $\tilde{R}_{1.5}$				
Predicted $\tilde{R}_{1.5}$	**1**	**2**	**3**	**4**	**5**
1	221	36	6	5	1
2	186	261	72	23	23
3	18	125	260	80	49
4	3	7	88	289	77
5	1	0	2	32	279

Table 17: *Performance statistics for all the five rank groups in the random forest model for $R_{1.5}$ using data at $z = 2$*

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.52	0.61	0.61	0.67	0.65
Specificity	0.97	0.82	0.84	0.90	0.98
Pos Pred Value	0.82	0.46	0.49	0.62	0.89
Neg Pred Value	0.89	0.89	0.90	0.92	0.92

E References

[1]. **Morphological statistics references to be added.**

[2]. **Barro, G.**, et. al. 2014, *The Astrophysical Journal*, 791:52 (23pp)

[3]. **Papovich C.**, et al., 2015, *The Astrophysical Journal* vol. 803, p. 26

[4]. **Snyder, G.**, et al., 2014, *Monthly Notices of the Royal Astronomical Society*, vol. 454, p. 1886

[5]. **Vogelsberger, M.**, et al., 2014, *Monthly Notices of the Royal Astronomical Society*, vol. 444, p. 1518

[6]. **Wellons S. & Torrey P.**, 2016, arXiv:1606.07815v1