Name: Soyoung Lee, Group: 3
Date: November 5, 2018

II. Method

For this specific section, the focus is on how galaxies with similar redshifts cluster in photometric space. The data used in this report is Data Release 14 (DR 14) from the website, https://www.sdss.org/dr/14, Slogan Digital Sky Survey SkyServer. The SQL query used to extract the data is written below.

```
SELECT TOP 2000
  p.objid,  p.ra, p.dec, p.u, p.g, p.r, p.i, p.z,
  p.cModelMag_u as c_u, p.cModelMag_g as c_g, cModelMag_r c_r,
  p.cModelMag_i as c_i, p.cModelMag_z as c_z,
  p.deVAB_u, p.deVAB_r, p.deVAB_i, p.deVAB_z, p.deVAB_g,
  p.expAB_u, p.expAB_g, p.expAB_r, p.expAB_i,  p.expAB_z,
  s.specobjid, s.class, s.z as redshift
FROM PhotoObj AS p
  JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
  s.z BETWEEN 0 AND 10
  and s.class = "GALAXY"
```

For the purpose of small scale data experiment, we selected the first 2000 galaxies whose redshift is between 0 and 10. This can be extended later when we find the clustering model to use. The SQL query has some extra variables that are potentially useful as well. The details of those variables can be found here, http://skyserver.sdss.org/dr14/en/help/browser/browser.aspx. The variables that are used in our modeling are summarized in the below Table 1.

Table Name	**Name in Table**	**Name in Data**	**Description**
PhotoObj	objID	objid	Unique SDSS identifier composed from [skyVersion,rerun,run,camcol,field,obj].
SpecObj	z	redshift	redshift
PhotoObj	cModelMag_u	c_u	Composite model magnitude for ultraviolet light
PhotoObj	cModelMag_g	c_g	Composite model magnitude for green light
PhotoObj	cModelMag_r	c_r	Composite model magnitude for red light
PhotoObj	cModelMag_i	c_i	Composite model magnitude for near infrared light
PhotoObj	cModelMag_z	c_z	Composite model magnitude for far infrared light

PhotoObj	deVAB_u	deVAB_u	The axis ratio of the de Vaucouleurs fit for ultraviolet filter band
PhotoObj	deVAB_g	deVAB_g	The axis ratio of the de Vaucouleurs fit for green filter band
PhotoObj	deVAB_r	deVAB_r	The axis ratio of the de Vaucouleurs fit for red filter band
PhotoObj	deVAB_i	deVAB_i	The axis ratio of the de Vaucouleurs fit for near infrared filter band
PhotoObj	deVAB_z	deVAB_z	The axis ratio of the de Vaucouleurs fit for infrared filter band

Table1. Descriptions of Variables

To understand cModelMag better, we need to have some background. As mentioned in the introduction, we will find sosie pairs using photometric observations of galaxies. That is to look at the intensity of light across the wavelength. The light flux can be used to represent the brightness of light emitted from a galaxy. Astronomers fit a model to acquire a galaxy flux and SDSS has two models, de Vaucoluleur's and the exponential models. They give two different fluxes, F_{deV} and $F_{exp,}$ respectively$_{.}$ $F_{composite}$ is a linear combination of these two fluxes as written in the formula below.

$F_{composite}$ = fracDeV F_{deV} + (1 – fracDeV) F_{exp}
where F_{deV} and F_{exp} are the de Vaucouleurs and exponential fluxes.

Then, composite model magnitude is calculated by taking a log inverse of $F_{composite.}$ This means that higher magnitudes correspond to deemer light and the lower magnitudes correspond to brighter light. There two types of AB ratios, one from each model. We chose to use AB ratio from de Vaucoleur's fit but which AB ratio to use can be further discussed if necessary.

We are interested in finding sosie galaxies within similar redshifts. Thus, we first start by grouping galaxies with similar redshifts. The rolling windows system allow us to overlap the windows so that we can avoid cutting out sosie galaxies by putting them into different redshifts groups. Then, we apply our clustering method within this redshifts groups.

Our clustering method is Density-based spatial clustering of applications with noise, DBSCAN, which is also a built in function in R. We chose DBSCAN because it works well to find similarities between observations with high dimensional feature spaces. We have 5 model magnitudes and 10 AB ratios so the space dimension can get quite large depending on how many of these variables we are using. DBSCAN is good at handling noise and our data have noise because we are measuring light from galaxies which are very far from us. The idea of this clustering method is to find high density in data space and cluster them together. This method requires two parameters, minPts and ε. The minPts is the minimum number of points that a cluster should have. In our case, we are looking for sosie pairs so we want to have at least two galaxies in each pair. Thus, we set minPts to be 2. The other parameter ε specifies the radius of neighbor points. In other words, if two points are within ε euclidean distance, then they are considered to be neighbors so DBSCAN cluster them together. When ε is big, we will find more clusters which might find clusters that are not actually sosie pairs. When ε is small, we will find less clusters which might cut off our sosie pairs because they are not close enough based on mischosen small ε criteria. The point is that ε should be sufficiently small enough to find actual sosie pairs but not too small. This is hard to pick unless we have a general idea of how far our observations are in the feature space. We choose ε by trial-failure methods to adjust for a given data and we will let astronomers choose what would be the best ε to use based on diagnostic plots. In this paper, 0.08 is used as ε to illustrate how DBSCAN find sosie pairs but this is adjustable.

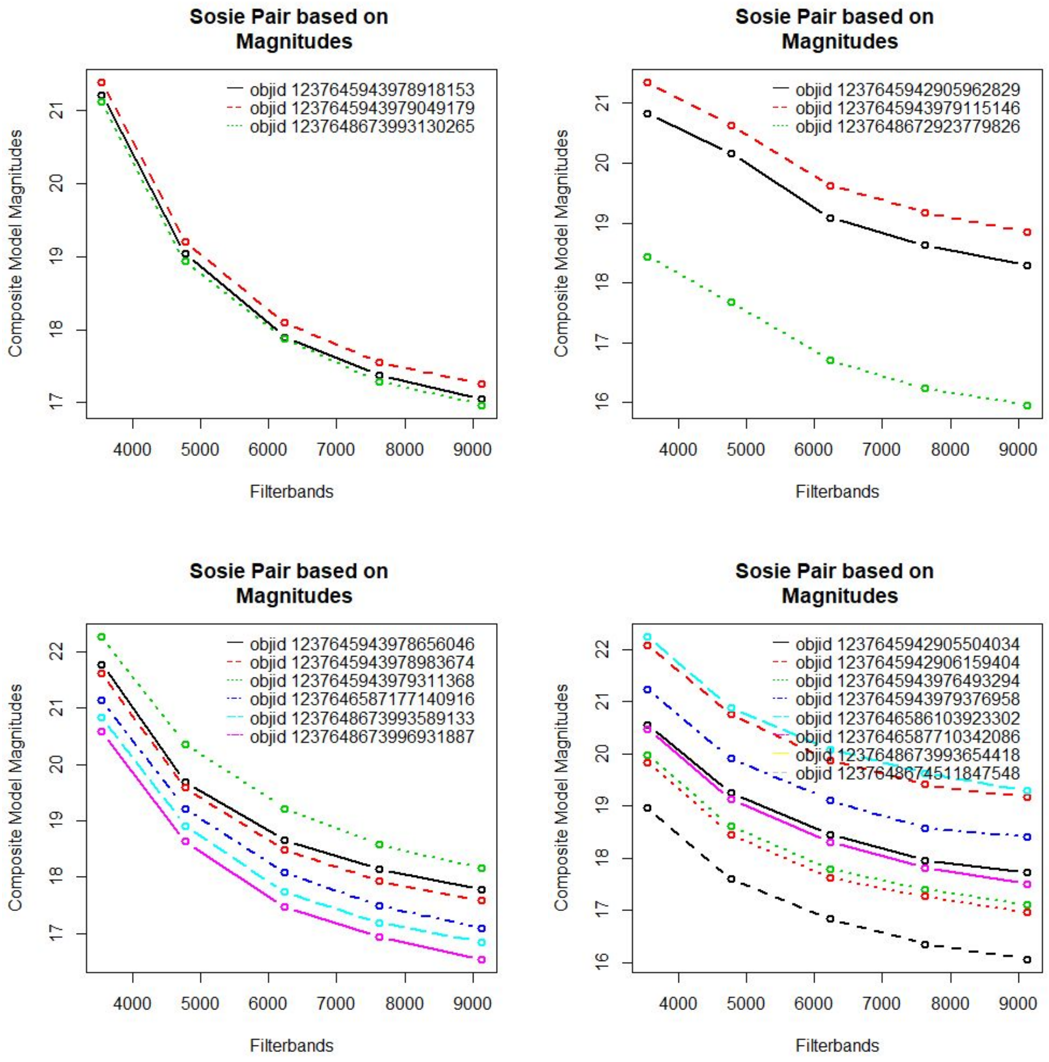

Figure 1. Sosie Groups with Varying Sizes clustered by Composite Model Magnitudes

Our first trial is to apply DBSCAN on cModelMag variables. The model magnitudes are normalized by subtracting cModelMag_u from other cModelMag variables. If the variables were fluxes, we would normalize them by dividing by u filter, but a magnitude is a log of flux. Thus, c/u becomes log(c/u) which corresponds to log(c) - log(u). This is why we subract cModelMag_u from other cModelMag variables when normalizing. After a few trials, we find 144 sosie pairs using ε = 0.08. About half of them are size of 2 and the other half sosie groups have varying sizes mostly between 3 and 6. Figure 1 shows how similar magnitudes for each filter these sosie galaxies have. Each line represents a galaxy. The x-axis are wavelength points where each filter works the best provided in Table 2.

Filter	Ultraviolet (u)	Green (g)	Red (r)	Near Infrared (i)	Infrared (z)
Wavelength (Angstroms)	3543	4770	6231	7625	9134

Table 2. X axis in Composite Model Magnitudes vs. Filterbands plots

The top two graphs look like they can be sosie triplets. The bottom two plots make us to think that we might need to add more data other than magnitudes. There are a few outliers where they have more than 100 galaxies in a sosie group. This might be because we do not have enough information or our ε is too big to further cluster them. We should be careful at providing smaller ε because we might lose actual sosie galaxies by doing that.

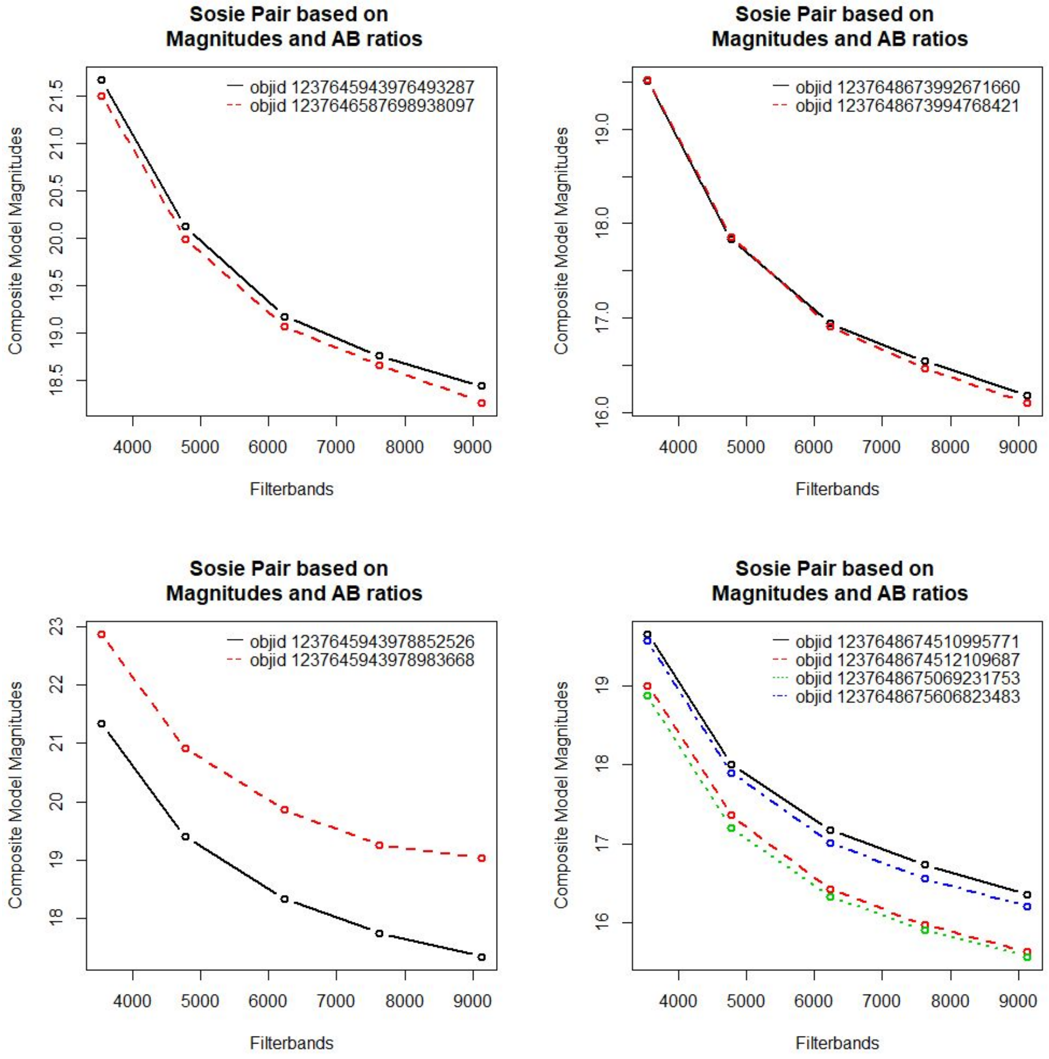

Figure 3. Sosie Groups with Varying Sizes clustered by Composite Model Magnitudes and AB ratios

We add AB ratios as features this time and apply DBSCAN with the same ε, 0.08. Now that we have more features to cluster on there will be less sosie groups. We indeed find only 9 sosie pairs and 8 of them have exactly two galaxies. The checking method works the same here. We look at magnitudes vs. filterbands. The four plots are provided above in Figure 2 for astronomers to check whether they are actual sosies or not. We can also increase ε to 0.09 which actually find 24 sosie groups with a size between 2 and 4. Increasing ε to 0.1 finds 35 sosie groups with a size between 2 and 6. As mentioned earlier, ε can be adjusted and this needs to be discussed with astronomers. Overall, plots are validating that DBSCAN works and we can say that we found 9 certain sosie pairs and there are opportunities to find more sosies by adjusting parameters or adding other variables.

References

36-611/612 Professional Skills for Statisticians. (n.d.). Retrieved November 5, 2018, from http://www.stat.cmu.edu/~hseltman/611

SDSS-III. (n.d.). Retrieved November 5, 2018, from http://www.sdss3.org/

SDSS SkyServer DR14. (n.d.). Retrieved November 5, 2018, from http://skyserver.sdss.org/