Classification and clustering algorithm applied in SDSS catalogue

Xiaojing Lin,[1]★
[1]*Department of Astronomy,School of Physics, Peking University*

Finished 19 December 2019; in original form 21 November 2019

ABSTRACT
In this work, different classification and clustering algorithms are applied to classifying spectral sub-classes and identifying large scale structure in the SDSS projected map. Density estimation is also done as the first step to reconstruct cosmic web in many previous works.

Key words: Machine Learning –astronomy – classification–clustering – large-scale structure of Universe

1 INTRODUCTION

Classifying celestial object types is always an important topic in astronomy. The large amount of data sets renders manual data reduction a boring and tough task for astronomers. Machine learning offers great tools for feature extraction and precise prediction, which is useful in astronomic data processing. In fact, there are plenty of pipelines in astronomy applying machine learning as classifiers.

The network-like large-scale structure of the Universe is well-known as the cosmic web.One of the most famous filamentary structures is The Sloan Great Wall (SGW), announced in 2003 from the Sloan Digital Sky Survey.

Filamentary structures are denser regions where matter clusters compared to large empty voids in the universe.Filamentary structures at different redshifts help to probe cosmological models.Properties of filaments and their interaction with nearby galaxies offer new insights into universe and galaxy evolution.(e.g. Y. Chen,et al 2015)

Different approaches have been applied to construct a catalogue for filaments.(e.g. Y. Chen,et al 2016) As Machine Learning technique are developing quickly, multiple clustering algorithms of unsupervised machine learning can be used in identifying the filamentary structures.

In this project,different classification and clustering measurements are applied to identifying different spectral sub-classes and the large-scale structures as well. These algorithms are all fundamental methods in machine learning,taught by prof. LiDou in her class.

2 DATA AND TOOLS

2.1 The SDSS Data

The data used in this project come from The Sloan Digital Sky Survey Data Release (SDSS DR8 and DR12). SDSS is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States.Its Eighth Data Release (DR8) contains all of the imaging data taken by the SDSS imaging camera (now totalling over 14,000 square degrees of sky), as well as new spectra taken by the SDSS spectrograph during its last year of operations for the SEGUE-2 project. More details about SDSS can be reached in `http://www.sdss3.org/`.

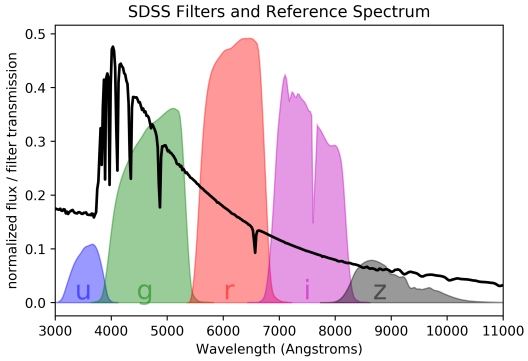


Figure 1. five SDSS filter bands along with a Vega spectrum.The figure is plotted using the example code in `astroML`.

SDSS photometric data are observed through five passbands, u, g, r, i,and z, using filters sensitive to photons in different wavebands(see fig.1). The differences of celestial bodies in those passbands indicate their color, revealing more information about their temperature, evolutionary phase and so on. Besides, SDSS data also comprise of spectral and redshift information of celestial bodies.The redshift z can be derived from their spectrum patterns in comparison with laboratory spectrum references.

Locations of SDSS celestial bodies are expressed in the form of right ascension (RA), declination (DEC), galactic longitude(l) and galactic latitude(b).The former pair is the standard coordinate frame in celestial coordinate system in which celestial poles are situated directly over the earth's geographical poles, and the latter

★ E-mail: linxiaojing@pku.edu.cn

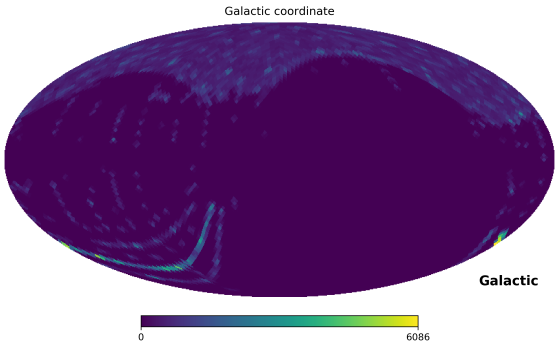


Figure 2. Map of DR12 objects in galactic coordinate.

ID Description	Teff effective temperature	log g gravity	FeH Fe Abundance [Fe/H]
ID Description	b-v color index	u-g color index	g-r color index
ID Description	r-i color index	i-z color index	

Table 1. Descriptive features in Classification

one is the galactic coordinate system setting its reference points in the Sun and the center of our milky way(see fig 2).

2.2 Open-source Python Tools

There are kinds of useful open-source python packages designed for machine learning as well as astronomical data reduction. In this project, invoked packages include `astroML`, `scikit-learn`, `astropy`, `numpy`, `matplotlib`, etc. Details and source codes of these packages can be found out in their homepages.[1]

3 CLASSIFICATION

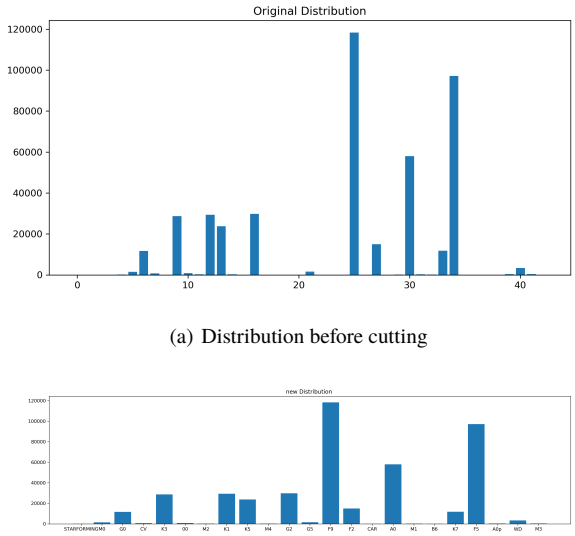
3.1 Data exploration and preparation

In Section 3 , we use the data set named *ssppoutput* in SDSS DR12, which contains 1,843,200 objects with 239 columns of features in a 1.9 GB file. It includes not only estimation for multiple physical features induced from different theories, templates and approximations, but confidence and errorbars introduced by both theoretical calculation and observation itself.

From a perspective of physics, we select 8 of them as the descriptive features in this task(see table 1). For the purpose of simplifying the problem, we just ignore errors though described in the data set.

All the objects are divided into 52 classes based on their spectral patterns.

[1] `www.astroml.org`
`scikit-learn.org`
`matplotlib.org`
`www.astropy.org`
`numpy.org`



(a) Distribution before cutting

(b) Distribution after cutting

Figure 3. Sample distribution in feature space before and after cutting

3.1.1 *Data visualization and preliminary reduction*

After visualizing the dataset as shown in figure 4, outliers are easy to identify. They come from observational fault (no effective record) and scientists use -99999 to fill the blank. We simply choose objects with effective records in all 8 dimensions as our samples, which contains 433,876 objects.

As shown in fig 3, the sizes of subsets in feature space differ a lot. Since several classes have only several samples, lacking representativeness , we just discard these parts and hold those classes with more than 100 samples.

Therefore 433,561 objects in 23 classes remain.

3.1.2 *Sampling*

Due to the ill - balanced sizes of subsets, we apply *Stratified sampling* to maintain the relative frequencies when producing training and test set: select 75% of the samples in each class as the training set and 25% as the test set. Fig 5 shows their similar distribution. The training set contains 325,161 samples and the test set 108,400 samples.

3.2 Classifiers

We apply several classifiers and calculate their class accuracy to make a comparison.

- **ExtraTree**:ExtraTree could be seen as an advanced extension for random forest algorithm. The difference is that ExtraTree uses the whole dataset and randomly splits the features.It also estimates importance of each feature. Fig 6 and 7 shows results of ExtraTree with 10 and 100 estimators respectively.ExtraTree with 10 estimators has an average class accuracy of 0.795 and a balanced accuracy of 0.613 (weighed accuracy of each class).ExtraTree with 100 estimators has an average class accuracy of 0.813 and a balanced accuracy of 0.658.
- **K-neighbors**: In KNN algorithm,weights are set to be Eu-

clidean distance. Fig 8 shows accuracy for KNN with different neighbors. It seems that considering 10 neighbors is a better choice.

• **Naive Bayesian** In this Baysian classifier, we set the prior probability to be the frequency of each class. Fig 9 shows the accuracy of this classifier.

3.3 Conclusion

It is obvious that the class accuracy largely depends on the sizes of classes. For larger subsets, it is more possible for a precise result. For those smaller classes, manual handling is indispensable.

4 CLUSTERING

4.1 Data preparation and sample selection

In this paper,the sample uses galaxies with 14.5 $<r_{pet}$ [2]$<$ 17.6, the same as (Y. Chen,et al 2016). The $r_{pet}>$ 14.5 limit ensures that all sample galaxies have reliable SDSS photometry and the r_{pet} < 17.6 limit allows a homogeneous selection as previous works .

Only samples with z < 0.13 are considered here. A small range of redshift are set to remove the Finger-of-God effect[3], to compare filamentary structures at different redshifts, and to reduce the computational cost.

Since the SDSS data do not directly contains spacial coordinate information. It is dispensable to establish a coordinate grid and settle the projected locations manually (fig.10). See more details of the fundamental approaches of sampling and projecting in **Appendix A**.

4.2 Clustering

In unsupervised machine learning, clustering is a good way for grouping a set of similar objects other than those sharing different properties in other groups. The 'Similarity' here refers to a relative short distance in the multi-dimensional feature space.In our work,the feature space could be interpreted as the spatial coordinate system, for we aim at grouping galaxies which are close to each other in the universe space. The descriptive features contain the continuous x-value and y-value of the galaxy samples' location in the coordinate system.

When applying clustering to search for large-scale structure, the total number of clusters is unknown and also could not be predicted. Thus, those algorithms sensitive to a prset cluster number (e.g. K-means clustering, spectral clustering) are not suitable for our task.

• **Hierarchical clustering (hierarchical cluster analysis or HCA)**: HCA builds a hierarchy of clusters by merging or splitting them successively. Usually a dendrogram (tree) is used to present the results.

One of HCA methods is Agglomerative Clustering, performing a hierarchical clustering from bottom to up : it starts by each point being a single cluster, and clusters are successively merged together after calculating and determining the metric. The function `astroML.clustering.HierarchicalClustering` is trying to reach the approximate Euclidean minimum, by establishing an Approximate Euclidean Minimum Spanning Tree (MST)

• **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**: It finds core samples of high density and expands clusters from them. 'High density' refers to a relative large number of neighbors within a selected radius, and usually the neighbor numbers should be larger than a predefined threshold. Generally it's good for data which contains clusters of similar density.However, using this algorithm needs constraints on the maximum distance between two samples for one to be considered as in the neighborhood of the other.Otherwise it does't works well.

• **Ordering Points To Identify the Clustering Structure (OPTICS)**:It is closely related to DBSCAN, finding core sample of high density and expands clusters from them. Differently, OPTICS keeps cluster hierarchy for a variable neighborhood radius. Theoretically it might be more suitable than DBSCAN for data varying in density.

5 DENSITY ESTIMATION AND STRUCTURE IDENTIFICATION

In this work, **Gaussian Mixtures Algorithm** is used for density estimation. Density in each cluster is assumed as a mixed Gaussian distribution and the number of Gaussian components are decided through a Bayes-based algorithms, by adjusting the weight of each component.

5.1 Identifying the Great Wall

The most famous large-scale structure in SDSS datasets is a part of cosmic web similar to the Great Wall.(fig.11)

fig.12 13 and 14 show 3 different clustering algorithms applied in the projected map of the SDSS great wall.

It seems OPTICS focuses on larger clusters compared with the previous two. And it is obviously that OPTICS costs much more time than the other two.

5.2 Extension to cosmic web

Change the training dataset into the total cosmic web projected map,then we can identify filamentary structures from the whole cosmic web.

fig.15 fig.16 show the targetted filamentary structures extracted from the background. Due to limitted computing power, it is hard to run OPTICS program in such a big dataset.

5.3 Conclusions

Different clustering algorithms are applied to identify large scale structure. Density estimation highlights denser regions in our universe where contain more matter.

Identifying large scale structure by clustering is very helpful in exploring galaxy and large-scale structure evolution ,if applied in scientific research.

ACKNOWLEDGEMENTS

The idea of this project comes from the excellent work by Y.Chen,etc. in 2016 (Y. Chen,et al 2016), which focused on establishing a filament catalogue through density ridges using a novel approach called **SCMS**.

[2] r_{pet} is the extinction- corrected r-band Petrosian magnitude.
[3] The small peculiar velocities of galaxies make galaxies stretched out along the line of sight in redshift space

At the very beginning, I followed every step in the textbook *Statistics, Data Mining, and Machine Learning in Astronomy* and reproduced related example codes in www.astroml.org. Instructions in this textbook helped me quite a lot.

REFERENCES

Chen Yen-chi, Ho Shirley,Brinkmann Jon, et al, 2016, MNRAS, 461, 3896–3909

Chen Yen-chi, Ho Shirley, Tenneti Ananth, et al, 2015, MNRAS,454, 3341–3350

APPENDIX A: ABSOLUTE MAGNITUDE AND DISTANCE MODULUS

The raw data fetched from SDSS are about galaxies with spectral information. Some related calculations are needed before getting a projection map of cosmic web (fig.10).

To select samples meeting our requirements, the first step is to derive absolute magnitude through distance modulus.

$$M_{absolute} = M_{obs} - DM$$

$$M_{obs} = -2.5log(Flux) + M_0$$

DM is the distance modulus.M_0 is the calibration zero point dependent on the photometry system. For SDSS system,$M_0 = 0$

In cosmology, Hubble constant H_0 could be written as

$$H_0 = 100h\mathrm{kms}^{-1}\mathrm{Mpc}^{-1}$$

here h is a dimensionless number and we can assume $h \approx 0.7$, in agreement with the most recent observations.

Dimensionless density parameters Ω_X indicates the density of component X relative to the overdensity of the universe.

$$\Omega_M \equiv \frac{8\pi G\rho_0}{3H_0^2}$$
$$\Omega_\Lambda \equiv \frac{\Lambda c^2}{3H_0^2}$$

$$\Omega_M + \Omega_\Lambda + \Omega_k = 1$$

M, Λ, k refer to the matter, cosmological constant and the curvature of space.

Thus the comoving distance of the universe can be derived as followings:

$$D_H \equiv \frac{c}{H_0} = 3000h^{-1}\mathrm{Mpc} = 9.26 \times 10^{25}h^{-1}\mathrm{m}$$

$$E(z) \equiv \sqrt{\Omega_M(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}$$

$$D_C = D_H \int_0^z \frac{dz'}{E(z')}$$

$$D_M = \begin{cases} D_H \frac{1}{\sqrt{\Omega_k}} \sinh\left[\sqrt{\Omega_k} D_C/D_H\right] & \text{for } \Omega_k > 0 \\ D_C & \text{for } \Omega_k = 0 \\ D_H \frac{1}{\sqrt{|\Omega_k|}} \sin\left[\sqrt{|\Omega_k|} D_C/D_H\right] & \text{for } \Omega_k < 0 \end{cases}$$

D_M is the transverse comoving distance and D_C is the line-of-sight comoving distance.

Thus the luminosity distance can be defined by

$$D_L = (1+z)D_M$$

and distance modulus DM is

$$DM = 5log(\frac{D_L}{10pc})$$

According to the standard universe model and recent observations, it is not bad to assume that our universe is a flat universe with $H_0 = 73.2$ and $\Omega_M = 0.274$.

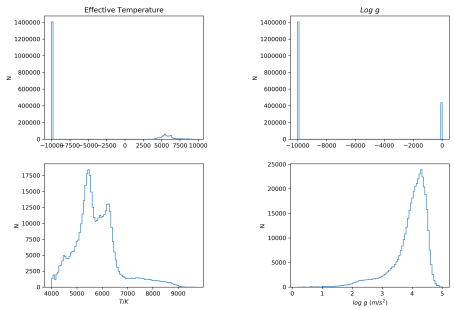
Therefore, using a standard universe model and redshift z we can derive the distance modulus DM, from which absolute magnitude can be easily obtained once compared with apparant maginitude.

Using comoving distance we can easily get the projected positions galaxies have in our celestial sphere.

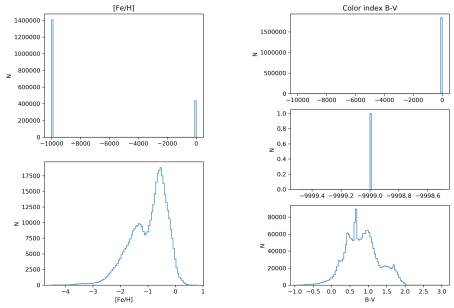
$$x = D_M \cdot sin(RA)$$

$$y = D_M \cdot cos(RA)$$

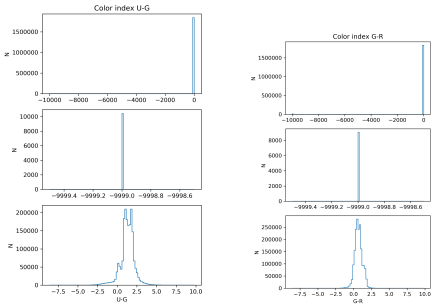
This paper has been typeset from a TEX/LATEX file prepared by the author.



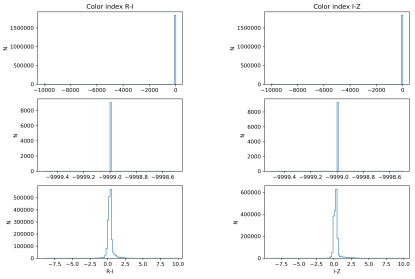
(a) Teff (b) logg



(c) FeH (d) B-V



(e) U-G (f) G-R



(g) R-I (h) I-Z

Figure 4. Data visualization and selection

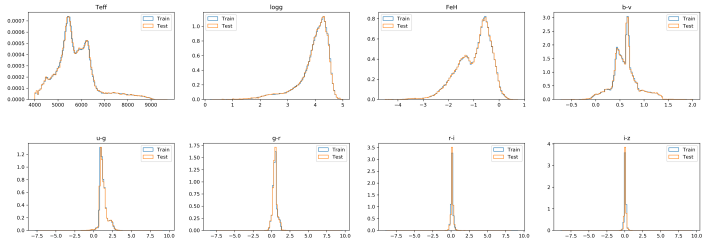
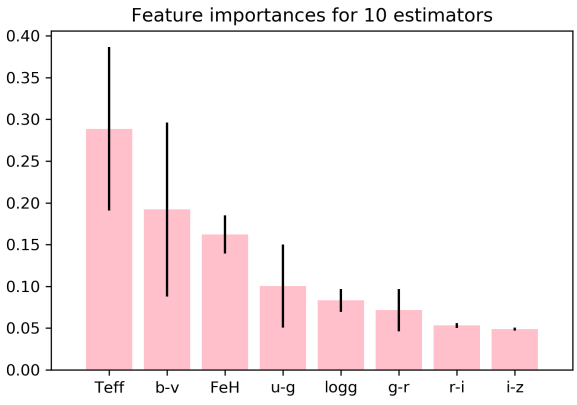
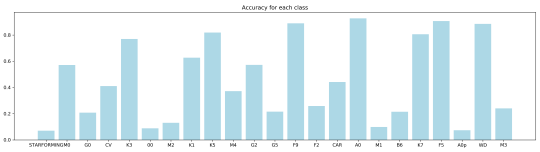


Figure 5. Distribution of training and test samples in feature space

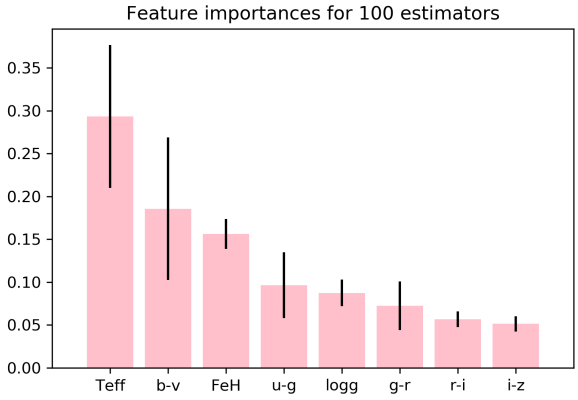


(a) feature ranking

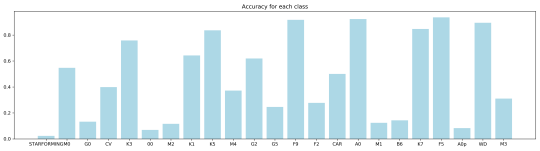


(b) class accuracy

Figure 6. ExtraTree with 10 estimators

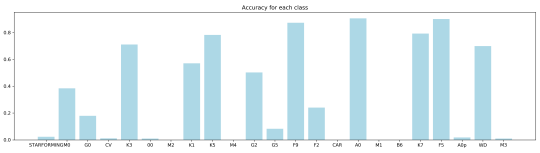


(a) feature ranking

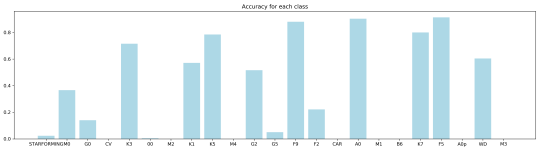


(b) class accuracy

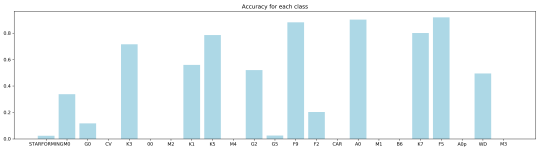
Figure 7. ExtraTree with 100 estimators



(a) 5-KNN: average accuracy 0.768, balanced accuracy 0.597



(b) 10-KNN:average accuracy 0.770, balanced accuracy 0.548



(c) 15-KNN: average accuracy 0.764, balanced accuracy 0.465

Figure 8. Accuracy of KNN

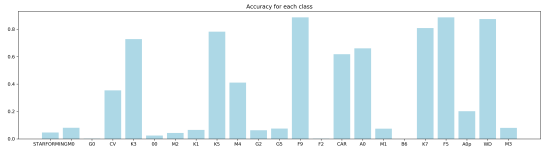


Figure 9. Naive Bayesian classifier: average accuracy 0.670 balanced accuracy 0.310

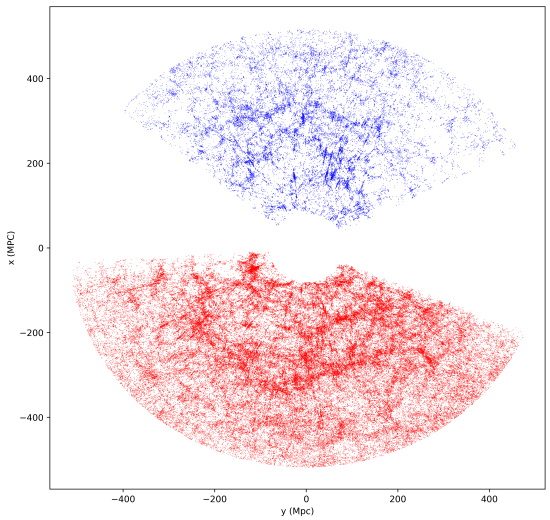


Figure 10. Projection map of the large-scale structure of the universe.Sample galaxies with $z < 0.13$ are plotted.

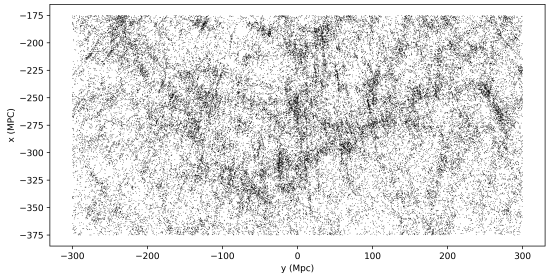


Figure 11. Projection map of the SDSS great wall.Sample galaxies with $z < 0.13$ are plotted.

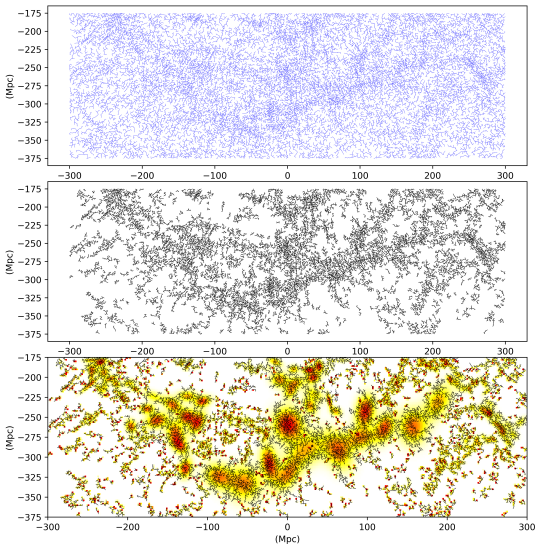


Figure 12. Projected map of the SDSS great wall after HCA. The up panel shows the full approximate Euclidean MST spanning the data.The middle presents the final (truncated) graph showing clusters.The bottom with various color depths indicates galaxy density.

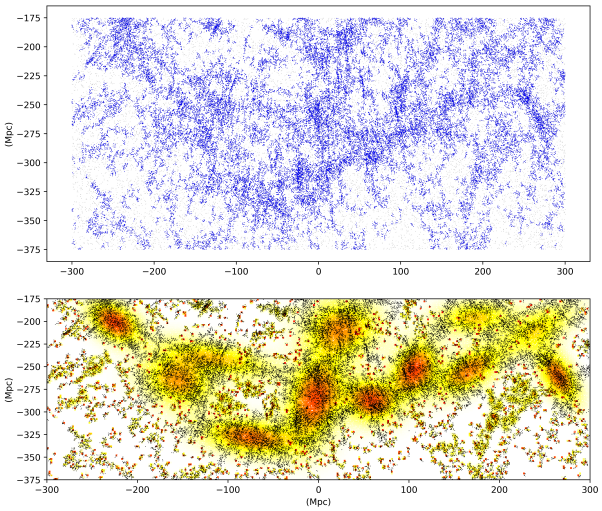


Figure 14. Projected map of the SDSS great wall after OPTICS. The up panel shows the samples in clusters identified out of background. The bottom panel presents the density distribution.

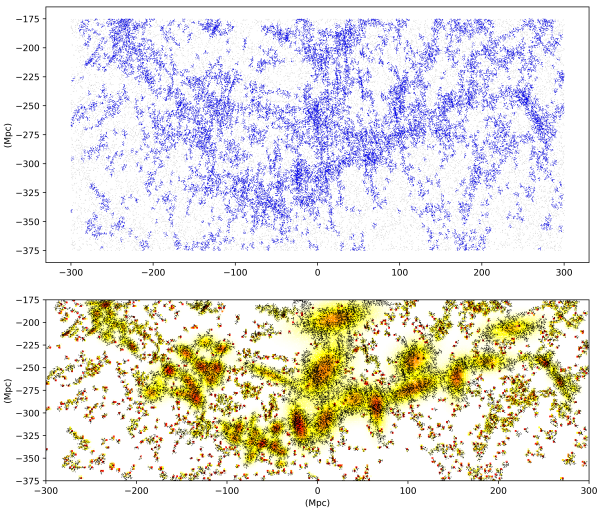


Figure 13. Projected map of the SDSS great wall after DBSCAN. The up panel shows the clusters identified out of background. The bottom panel presents the density distribution.

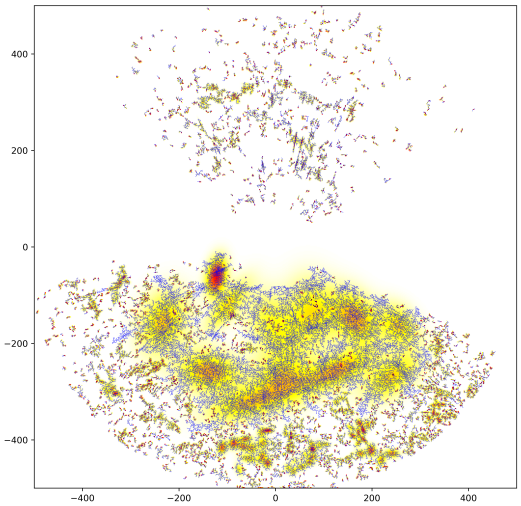


Figure 15. Projection map of the SDSS cosmic web after HCA. The blue line indicates the final (truncated) graph showing clusters.Various red color depths indicates galaxy density.

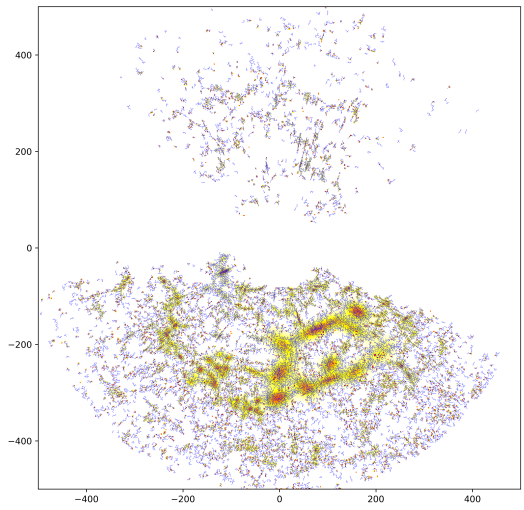


Figure 16. Projection map of the SDSS cosmic web after DBSCAN. The blue line the samples in clusters identified out of background.Various red color depths indicates galaxy density.