Anomaly detection of SDSS photometric redshifts using Data Analysis Techniques

A Machine Learning Analysis with data taken from SDSS

Gautham Gururajan[1] and Shantanu Desai[2,⋆]

[1] Undergraduate, Indian Institute of Technology(IIT), Hyderabad
e-mail: ep17btech11008@iith.ac.in
[2] Professor, Indian Institute of Technology(IIT), Hyderabad
e-mail: shantanud@iith.ac.in

ABSTRACT

Aims. To detect anomalies in training data consisting of spectroscopic redshifts using Elliptical Envelope technique and later test with some machine learning frameworks to learn efficiently information acquired from the SDSS Sky Server - Casjobs, from the DR15.
Methods. We use MySQL to query data with 21 easily measurable features.We then apply an array of data analysis techniques to logically deduce the reliability of datapoints, and hence construct different datasets. Using our base anomaly detector as the Elliptical Envelope technique, we contaminate our training dataset with different concentrations of unreliable redshifts and then evaluate some parameters($|\mu|$, $\Delta_{z'}$>0.15, 1σ, 2σ) that could give us an estimate of how well the anomalies were filtered out. In the end, we use some machine learning algorithms including the novel Self Organising Map (SOM) through a SOTA python library SuSi as well as Adaboost regression of Decision Trees (ADR) through sklearn to see how well the redshift predictions are with and without anomalies.
Results. Our experiment shows an improvement of evaluation metrics after using the Elliptical Envelope method and testing on machine learning redshifts, about a 2% improvement of outlier fraction through the SOM architecture and a 5% improvement of outlier fraction through the ADR architecture. We would also like to point out that removing 2-5% of the galaxies not only improves performance, but also decreases runtime in the case of a huge dataset (which is almost a given in machine learning redshift estimation.)
.

Key words. Methods – Machine Learning, Statistics Galaxies – Distances and Redshifts Cosmology – Observations

.

1. Introduction

For decades, scientists have been trying to understand our universe, and the key to doing so lies in it's expansion theory. Similar to a chocolate chipped cookie being baked, if you imagine yourself on a chocolate chip, no matter how you view the cookie, the farther away another chip is, the faster it recedes. Directly drawing parallels, all the other galaxies are moving away from ours as the universe expands. And because the universe is uniformly expanding, the farther a galaxy is from Earth, the faster it is receding from us.
Through the Doppler effect, the light emitted from these galaxies is shifted towards the red end of the spectrum and the distance of these galaxies from us is measured by a quantity known as *redshifts*.

These quantities have often been measured through spectroscopic techniques, as explained below. The most straightforward way to measure this distance would be through it's electromagnetic spectral energy distribution (SED), which is made of continuum and emission/absorption lines. Emission and absorption lines are sharp features which can be easily identified in the SED. Also, it is common knowledge that due to the expansion of the universe, the SED is shifted towards longer wavelengths by a factor of 1 + z where z is the redshift.
The main difficulties in distance estimation is of finding a pair of characteristic features in the SED and measuring the amount by which they have been stretched.
By using a necessary reference point, the measured redshift is then mapped to a corresponding distance value. Two well known features shape the SED con-

⋆ Project Guide for EP4275

tinuum, The *Balmer break* (Below 4000Å), which is explained by absorption of photons more energetic than the Balmer limit at 3646Å and the combination of numerous absorption lines by ionised metals in stellar atmospheres, and The *Lyman Break* (Below 1216Å), which is explained by absorption of light below the Lymann limit at 912Å and the absorption by the intergalactic medium along the line of sight.

When SEDs of sufficient wavelengths resolution are available, the emission/absorption lines can be identified, and the redshift precision can be measured to be better than 10^{-3}.

Despite having sophisticated multi-object spectrographs, only a few meaningful spectra for a few percent of the sources detected through deep imaging surveys are obtained. At least two well identified spectral features are required to obtain a robust redshift measurement. It is a fact that the success rate of measuring spectroscopic redshifts can be lower than 50-70% in deep spectroscopic surveys.

An alternative to this costly and time intensive method

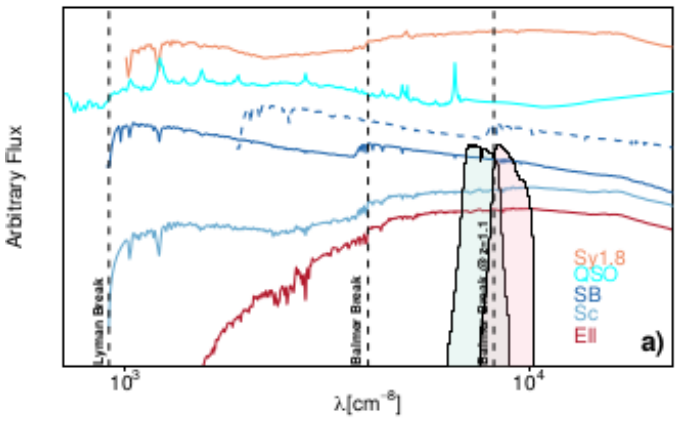

Fig. 1. Examples of SEDs for various type of galaxies(elliptical:Ell; Starburst:SB; spiral with small bulge: Sc) and AGN (luminous quasar: QSO; low luminosity, obscuredAGN: Sy1.8 (Seyfert 1.8)) The Lyman and Balmer breaks, among the key features in determining the redshifts, at rest-frame are indicated by vertical dashed lines. One template and the Balmer break are also plotted at redshift 1.1. For clarity, the transmission curves of i and z filters, covering the wavelength rage between 700 and 1100 nm are also indicated. Taken from [19]

is by measuring the flux of a source in broader filters. We can obtain a sample sufficient enough to give us details such as shape of the continuum, extra-galactic nature of sources and an estimate of redshift based on broader features such like the Lyman and Balmer breaks, or strong emission and absorption lines. This low resolution, less accurate distance estimate is called a *photometric redshift*. Although these measurements are relatively much cheaper with respect to computation cost, they pay the price by being much less precise(in the order of a 100 times factor). The figure from [19] shows us how we can consider broader features like the breaks and the arbitrary flux information that is relevant to this make meaningful estimations of redshifts

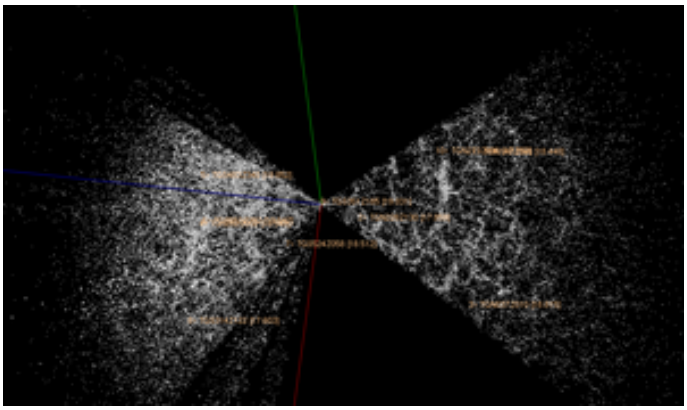

Fig. 2. A map of our surrounding universe taken from the 2df Galaxy Redshift Survey

Knowing redshifts allow us to make a proper map of the universe (Fig.1) and also can help us understand past events, and hence, having well measured redshifts is paramount to such analyses. Subsequent sections will shed more light on this. The data used in this analysis is from the publicly available archive, `CasJobs`, which we will talk about later. The data was surveyed through the `SDSS`.

This machine learning approach works under the assumption that the mapping is constructed from an unbiased and uncontaminated training dataset. This leads us to the question - In such cases, are there any chances of removing such 'biases' and 'contaminants' through some data analysis techniques, and can it really produce significant changes in results? And the answer to this is, yes. and will be discussed in detail in the upcoming sections.

Research shows that the contamination of a training sample can adversely affect the recovered machine learning redshifts. Ref. [3] uses simulated spectra to show how the cosmological constraints for a weak lensing survey are degraded in the presence of even 1% of spectroscopic outliers in the training sample.

Machine Learning

Clearly, accurate measurements of spectroscopic redshifts are time intensive and costly, and so such measurements are made only to some galaxy subsamples and then 'mappings' are 'learnt' between some photometrically observable features and of galaxies and spectro-

scopic redshifts of these samples so that the same mappings can be used to estimate redshifts of other galaxies with ease, but comes at the cost of low precision.
Of course, there are other methods such as Spectral Energy Distribution (SED) - template fitting techniques or hybrid techniques that are also used depending on what type of features are available. More details on how SED techniques are used are mentioned here.

Band	λ_{eff}
U	365 nm
G	475 nm
R	658 nm
I	806 nm
Z	900 nm

Table 1. Approximate Effective midpoints for the U-G-R-I-Z bands.

SDSS

The *Sloan Digital Sky Survey* [21] began the operation phase at around May, 2000 and is currently operating in its fifth phase SDSS-V. The main intention for such a project was simple, to carry out a contiguous survey of all measurable galactic/extra-galactic objects.
This is being done through through imaging and spectroscopic surveys, with a dedicated 2.5m wide angle optical telescope equipped with a large format CCD camera to image the sky in five optical bands (Table 1).
Also it was equipped with two digital spectrographs to obtain the data of more than a million galaxies and more than a hundred thousand quasars selected from imaging data.
Spectra of more than three million astronomical objects have been obtained, along with documented deep multicolor images of more than one third of the sky.
It periodically updates the publicly accessible archives in phases known as *Data Releases*. Our paper makes use of the Data Release-15 [1]

DR-15

As seen in Ref. [1], the data release was for the fourth phase of SDSS (SDSS-IV) spanning its first three years of operation (July 2014-July 2017). This is the third data release for SDSS-IV, and the fifteenth from SDSS (Data Release Fifteen; DR15). New data has come from MaNGA [2] (Mapping Nearby Galaxies at Apache Point Observatory) - they released 4824 datacubes, as well as the first stellar spectra in the MaNGA Stellar Library (MaStar), the first set of survey-supported analysis products (e.g. stellar and gas kinematics, emission line, and other maps) from the MaNGA Data Analysis Pipeline (DAP), and a new data visualisation and access tool they call Marvin.

Getting the data

For every machine learning problem, the main concern is of getting a database large enough to provide results that are not very skewed towards a particular behaviour. However, in our case it is relatively easy to obtain data, as associations like the SDSS make sure that all observed data is made available to the public through their sky server, named CasJobs [14].
Catalog Archive Server Jobs System allows you to access different types of schema. (A schema is a unique object that contains tables which contain information, A Catalog is a collection of schema : *To summarize, Cluster* > *Catalog* > *Schema* > *Table* > *Columns*/*Rows*), and on selecting one, you may issue your Query.
Since they would ideally need to have an arbitrarily large number of databases, MySQL[15] is provided to make queries instead of SQL.
We select the DR15-schema [1] and query out some features, most of which we think would be helpful to predict photometric redshifts (The code is attached in the appendix section). We subsequently issue our query and get 2,697,380 galaxies.

Terminology and Features

In this analysis, we choose eight features for the outlier estimation and this is by all means just a method to check for outliers, the final performance of this anomaly detection is measured by checking certain parameters such as Outlier Fraction (Δ>0.15), 1σ deviation, 2σ deviation and Median Absolute Deviation($|\mu|$) of predicted redshift.
As seen in the paper by Hoyle et.al [11], different photometric features have different 'importances', and so, to make a powerful prediction, we would use those features. A similar analysis has been done by the author, using a variety of feature importance techniques, to summarize that, Gini Coefficients were used through Decision Trees to give the optimal runtime with the best features as compared to many other feature importance techniques. More can be read about this here. The most important features can be seen in Figure 2.
But as mentioned, the motive of this paper is to identify outliers and so, we use features that could be helpful for this.
In this work we have concentrated on the following eight

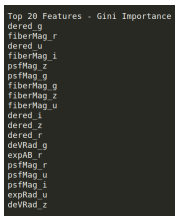

Fig. 3. Some features that are arranged in decreasing order of importance for the selection of features for a regression analysis. Although, the order of these features change based on hyperparameter tuning, the top features remain the same in almost all cases

features for outlier estimation; the spectroscopic redshift and error, r band magnitude, the following colors (color a_b is dered_b (dereddened magnitude in band 'b') - dered_a (dereddened magnitude in band 'a')) : *g_i, g_r, r_i, z_r*, and the Petrosian radius measured in the r band (Table 2). Of course will only use the photometric quantities in the ML prediction of redshift.
We also see that using the SDSS data for such an experiment is only possible due to a criterion called 'zWarning' which we will explain along with some other important evaluation criteria.

1.0.1. zWarning

A given area in the sky may be passed through by multiple runs of the telescope. Due to unavoidable repositioning and re-calibrating errors, a single galaxy may report multiple values for the same features and such a thing happening is quite common.
Merging different survey phases is also common and so each galaxy/celestial object is allowed some error. Having such a quality is unique to the SDSS as far as the author knows. The readings from these objects are then passed through the processing pipeline and if a huge deviation exists, the 'zWarning' parameter is set to a non-zero value, stating that this set of readings is unreliable. In fact, some unreliable readings are not marked with zWarning!=0 and are hence followed up on at later dates, and this is the exact problem we plan to address. Using the galaxies with non zero zWarning, amongst other filters (Which will be discussed soon), we prepare an undoubtedly saturated sample of unreliable redshifts, which we later augment to an uncontaminated data set to allow a proper analysis of outlier detection.

1.0.2. Redshift scaled residual vector

For a proper way to evaluate our results after performing regression, we require a framework that could help relate the quantities of redshift before and after anomaly (outlier) removal. The Redshift scaled residual is defined as:

$$\Delta'_z = \frac{z - specz}{1 + specz} \tag{1}$$

Where z denotes redshift predicted through regression, and *specz* is the true spectroscopic redshift. Each Δ'_z acts as a single transformed unit that can aid us in giving meaningful results through checking quantities such as Standard deviation, Median Absolute Deviation etc., and what is expected is that these quantities be decreased after applying the base anomaly detector.
We discuss those quantities now.

Description	Feature
Magnitude	dered_u dered_g dered_r dered_i dered_z
Radii	petroRad_r
Spectroscopy	specz specz_er

Table 2. The list of all features used in this work. A brief description of each of these features can be found on the SDSS Sky Server web page.

1.0.3. Evaluation Metrics

Here 68% and 95% spread (1σ and 2σ) are calculated for Δ'_z. The MAD $|\mu|$ - Median Absolute Deviation) is also calculated corresponding to the median value of Δ'_z.
We also look at a quantity called Outlier Fraction that checks the number of galaxies that tells us the fraction of galaxies that satisfy $\Delta'_z > 0.15$.
A point to note here is that the mentioned 'Outlier' in 'Outlier Fraction' is different, although related to the anomalies/outliers that we talk about in the rest of the paper.

Constructing our datasets

In order to realise the idea of augmenting a completely impure dataset to a pure dataset, we need to construct the aforementioned. So we will begin with cascading some filters.

Filter 1 - zWarning

We start with 2.5M galaxies. As mentioned in the previous section, the primary filter that we apply is to check whether or not a galaxy is flagged with a zWarning!=0, and if so, we know that it is questionably reliable, and

this becomes our base unreliable sample dataset, leaving us with 112221 unique objects' samples.

Filter 2

This is a much more intuitive logic based filter. For those galaxies that have questionable measurements, we check for galaxies with copies (To check if there are multiple datapoints that correspond to the same Galaxy/Object ID). Galaxies, without copies, we cannot really say anything as the redshift may or may not be accurate and hence we remove them from our set. We check for galaxies with both poorly measured redshift and a well measured redshift with specz-err< 0.001. We are left with 9393 galaxies of which 4221 are unique.

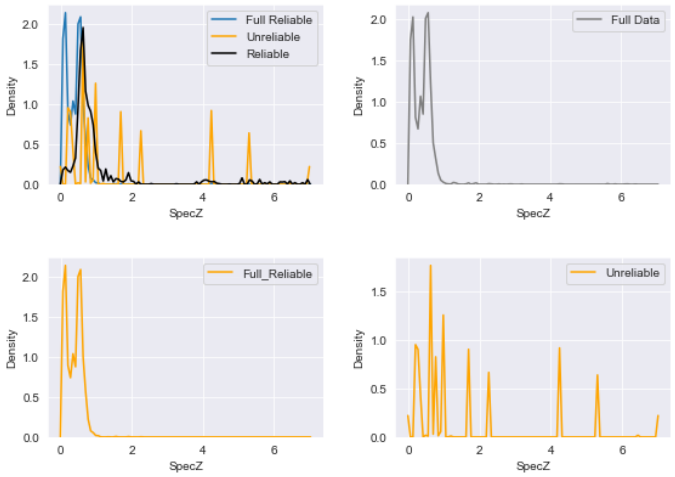

Fig. 4. The distributions of the sets disjoint with respect to reliability as well as a comparison with the distribution of the full dataset. Looking at the bottom right panel, we see the abnormal density distrubution, which is expected.

Filter 3

The last filter would be made on the fair assumption that the error on redshift learning estimate would be less than 0.01 and so, we filter out galaxies with difference in well measured and poorly measured redshift > 0.01 as the most impure subset of redshifts. At this step, we are left with 2426 unique galaxies. We look at the distributions of the above and can clearly see that the unreliable redshifts peak abnormally towards higher redshifts which suggests anomalous behaviour.

2. Anomaly Detection

We finally get to our anomaly detection and will be using a method called Elliptical Envelope (EE) technique as our base detector. We make use of the scikit-learn package [16], and also make use of the fact that high level gaussians can be modelled into the data as seen in Fig.2. The EE models the data with all possible covariences between features, and then constructs a ellipse with all point inside a boundary signifying inliers and outside signifying outliers.

The EE routine makes use of the FAST-Minimum Covariance Determinate (MCD).As [18] shows, The minimum covariance determinant (MCD) method is a highly robust estimator of multivariate location and scatter. Its objective is to find h out of n (Total Number of observations) observations whose covariance matrix has the lowest determinant to estimate the size and shape of the ellipse. FAST-MCD is able to detect an exact fit—that is, a hyperplane containing h or more observations.

The FAST-MCD technique, in detail selects non overlapping samples of data and computes mean (μ), covariance matrix (C), in each feature dimension of each subsample. The Mahalanobis distance d_{mh} , is computed for each multidimensional data vector $\boldsymbol{x}$, in each subsample and the data are ordered ascendingly. Mahalanobis distance is defined by:

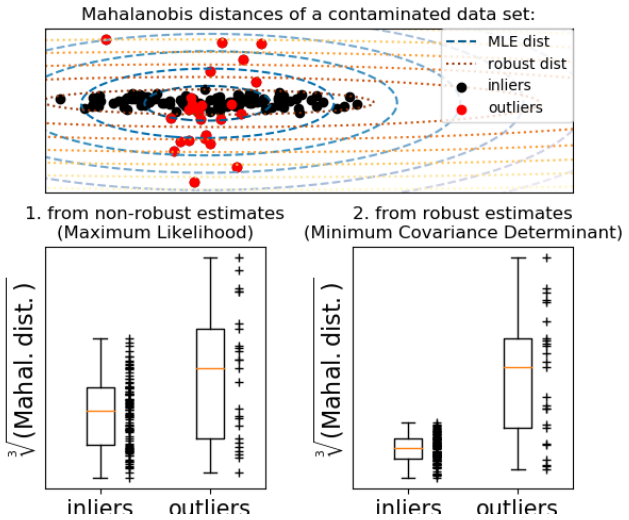

Fig. 5. With the assumption that the data comes from a known distribution (here, Gaussian), we try to define the 'shape' of the data and define outliers as the readings that stand out from this shape. In a robust way, inlier location and covariances are estimated and the d_{mh} is used to derive a measure of anomaly. The figure illustrates the strategy.

$$d_{mh} = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})C^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \quad (2)$$

In simpler words, d_{mh} measures how many σ a datapoint is from the mean of it's distribution.

Subsamples with small d_{mh} are continuously computed along with values of μ and C and is iterated till the determinant of the C converges. From all the subsamples, the covariant matrix with the smallest determinant is chosen and forms an eclipse with data inside it being classified as inliers and outside it as outliers. Clearly this

method is thorough and robust and displays better performance as compared to outlier detection compared to other conventional techniques as seen in the figure taken from here. A hyperparameter of EE is ContPar, the contamination parameter, which is apriori assumed as the fractional contamination of the dataset, and as we see it is not in need of being hyper-tuned and can give accurate results with a rough estimate.

In our experiment, we build our training+testing sets with a random 100<N_{ur}<2420 unreliable redshifts and 3*N_{ur}<N_r<100000 completely reliable redshifts and check for a single test run due to limited computation power, but this gives us a general idea of the working as well as how well a model can perform despite not being tuned properly.

Fig 4. contains our findings and we take ContPar=0.140 for the rest of the analysis as it is visible that this value has a very good tradeoff. Note that changing ContPar to values in the neighbourhood of 0.140 will yield similar results as there is less need to hypertune, as mentioned before.

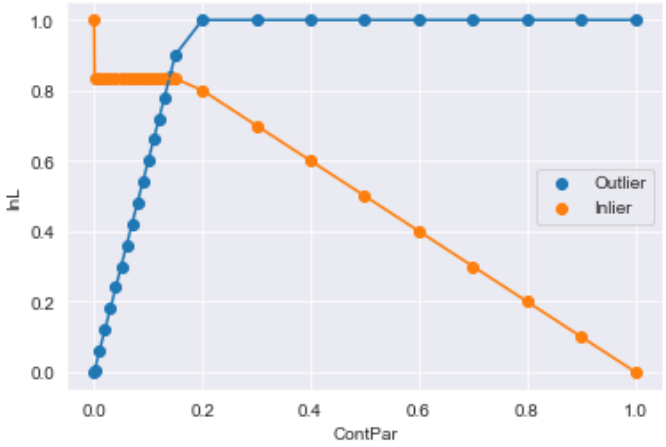

Fig. 6. A plot of correct fraction of outliers/inliers detected with change in the contamination parameter, we see that above a certain contamination parameter, the fraction of outliers detected becomes 1 while the fraction of inliers decreases, this is similar to the case of overfitting.

Machine Learning Redshifts

We focus on two models, the Self Organizing Maps (SOM) and the Adaboost Regression (ADR) Techniques, and compare them with each other as well as their performance in evaluating the above mentioned metrics before and after applying the EE method, and now discuss it.

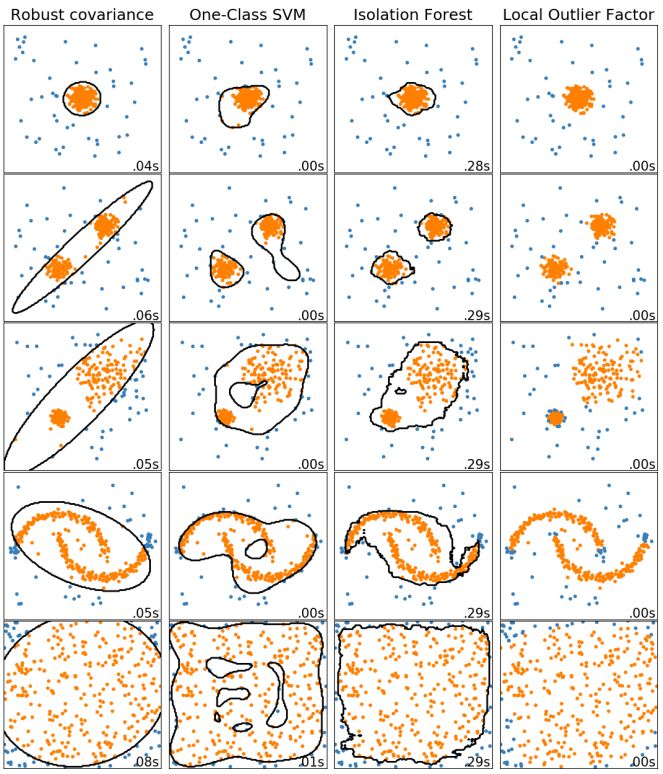

Fig. 7. A comparison of many anomaly detectors over a dataset with clearly marked outliers, we see the ellipses in the robust covariance (EE) method providing efficient detection, unlike models like One-Class SVM which is known to be sensitive to outliers and thus does not perform very well for outlier detection. The isolation forest performs well, but as seen the run time is almost 5-6 times more in this case.

SOM

We use the SOTA `SuSi` python library [17] in this supervised regression problem. As far as a machine learning analysis goes, the SOM [12] (Also called a Kohonen map) is weakly represented. Since we work with datasets of sizes that may or may not be small (2000 < len(data) < 102000), we require an algorithm that is well suited for this type of uncertainty. In brief, the SOM, being a type of ANN has a unique characteristic : to find out the neighborhood relation of output neurons, this relation helps to improve generalization on small datasets.

The composition of a SOM is *nxn* units organized on a 2D-grid lattice, with the unit prototypes being updated with a set of rules, with decreasing learning rate and a gaussian neighbourhood. The map activation is equal for all points, with the output function being a linear regression function.

The map activation $\boldsymbol{a(t)} = a_{ij}(t)_{(i,j)}$ is equal for each unit located at position (i, j) to the euclidean distance between its prototype $\boldsymbol{w}_{ij}^{S}$ and the current input $\boldsymbol{x(t)}$:

$$a_{ij}(t) = \|\boldsymbol{w}_{ij}^{S}(\boldsymbol{t}) - \boldsymbol{x(t)}\|$$

The SOM output activity $\boldsymbol{o(t)} = o_{ij}(t)_{(i,j)}$ is the activity cascaded to the other modules of the integrated architecture, in our case, being a regression problem , would be a linear regression function. It is defined by applying an output function to the map activation, ie, at pixel (i,j):

$$o_{ij}(t) = \phi(\boldsymbol{a(t)}, \theta_{ij})$$

where θ_{ij} represents the optimal parameters for the function ϕ. Depending on conditions, many possible functions for ϕ exist, some of them being the Gaussian Similarity function, the BMU (Best Matching Unit), Softmax etc.
Finally we have the linear regression itself, which gives the final predicted value $\boldsymbol{p(t)} = p_k(t)_k$ which is:

$$p_k(t) = \sum_{(i,j)} w^L_{kij}(t) o_{ij}(t)$$

This concludes the construction of the architecture, with our problem using the softmax activation and the linear regressor to predict values.
The study in [8] shows us that varying the activation function and hypertuning the parameters can help increase the performance greatly, in their case by around 7 times, but this is not the focus of our analysis.
We apply the SOM method to our data for regression after and before the EE technique and check the relative improvement of our metrics, as expected, four our optimal value of ContPar = 0.140, we have a positive improvement.

Fig. 8. This plot shows a convincing relative improvement for the parameters and can hence be seen as a successive removal of anomalies for the SOM/EE architecture.

Another point to note is that the performance decreases for higher ContPar, which is understandable because if our ContPar exceeds the true, contamination fraction, inliers would be classified as anomolous as well, and so, quantities like σ would only decrease.

AdaBoost Regressor

AdaBoost or Adaptive Boosting, is an ensemble boosting technique through which the output of other week learning algorithms taken into a weighted sum is the output of the boosted regressor.

$$Ada_n = \sum_{i=0}^{n} w_i a_i$$

Where w_i and a_i represent weights and 'weak learners' respectively. Here, our 'weak' learners will be Decision Trees, and weights are chosen based on errors by each learner.
We follow the `sklearn` implementation of AdaBoost [7]. Given below is a brief overview of the working of selecting the weights, described in [6], explained in [11].
Each galaxy is assigned a weight w_i, and a Decision Tree Regressor is trained on a bootstrapped dataset of size N. Each element has a probability of getting selected given by:

$$p_i = \frac{w_i}{\sum_{i=1}^{N} w_i}$$

A new model is thus produced which is added to the ensemble of models. The training set loss L_i , for each element i is calculated as

$$L_i = \frac{| Pz(F_i) - Pz_i |}{sup_j | Pz(F_j) - Pz_j |}$$

Where $Pz(F_i)$ is the function represented by the corresponding tree. L_i is normalized in such a way, that $L_i \in [0, 1]$. Average loss $\bar{L}$ is given by:

$$\bar{L} = \sum_{i=1}^{N} L_i pi$$

Where the sum runs on elements over the whole set. Confidence β is defined as:

$$\beta = \frac{\bar{L}}{1 - \bar{L}}$$

and the weights for each model are iteratively updated by multiplying the weights for each element in the training set by β^{1-L_i}
This weight update procedure gives less weight to elements with a low prediction error L_i and therefore these objects are less likely to be included in the training set drawn in the next boosting iteration. This focuses subsequent learners on elements with a high prediction error. We train a number of Decision Tree Regressors in this fashion and update the weights for the training set. The

number of trees M included into the ensemble is decided by the user of the algorithm.If we query a new object with input features F_i , we obtain a prediction $Pz_j(F_i)$ for each tree in the ensemble $j \in 1, ..., M$. The final machine learning redshift prediction $Pz(F_i)$ is then given as the weighted median of the redshift predictions of the models in the ensemble with respect to $\log(\frac{1}{\beta_j})$ (as described in [6]).

As we see, since this model requires to compute many smaller prediction models, it has a large time complexity.

In our problem, we find use the linear loss function only because according to [5] the exponential loss function has been seen to be sensitive to label noise in classification. Also, we see that hypertuning the parameters can lead to a great improvement, but this analysis does not require so.

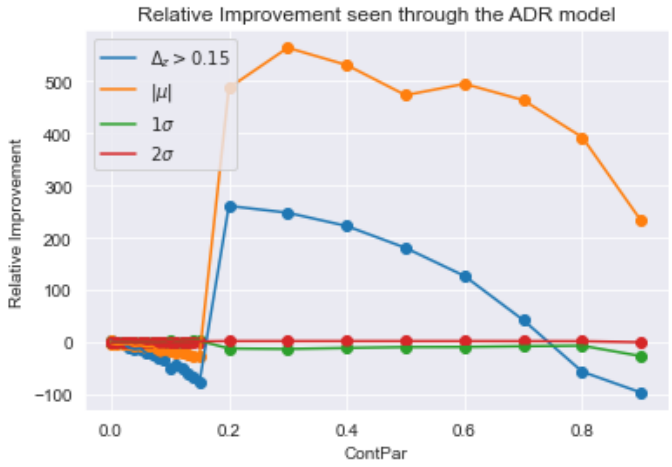

Fig. 9. The relative improvement of quantities exceeds or is comparable to the SOM analysis for our ContPar=0.140, but a point to note is that this would also be the case for different contamintaions

Results

Looking at our results, we see that ADR outperforms the SOM, which is in agreement with the results stated in [10], we also see that the Elliptical Envelope technique gives us a clean best case 5% improvement in the Outlier Fraction (Which is said to be the most reliable metric for outlier analysis [9]). The other metrics also show a positive improvement in case of clean datasets with appended outliers. To conclude the project, we summarize it briefly. With the help of `CasJobs` and `MySQL`, we were able to query our required samples and store them. We used 2.5M galaxies initially and applied a series of cascading filters that helped us to classify the degree of reliability of the data.

A point to note would be that this is possible only due to the feature that is apparently only exclusive to the SDSS, the `zWarning`. This quantity flags possible readings that could have been miscalculated, and due to this, we are able to build a sample of galaxies that has a high concentration of unreliability.

Algorithm	Sample	$\|\mu\|$	$\Delta > 0.15$	1σ	2σ
SOM	Inlier and Outlier	0.0418	0.049	0.026	0.001
	Inlier	0.0306	0.021	0.016	0.006
ADR	Inlier and Outlier	0.0232	0.065	0.409	0.205
	Inlier	0.0170	0.014	0.015	0.005
SOM	(NPC)Inlier and Outlier	0.0275	0.0064	0.044	0.9952
	(NPC)Inlier	0.0462	0.0282	0.150	0.029
ADR	(NPC)Inlier and Outlier	0.0136	0.0072	0.041	0.010
	(NPC)Inlier	0.0527	0.0359	0.100	0.003

Fig. 10. Results of our analysis, NPC relates to No Pre-Contamination. The top 2 rows of each algorithm presents our results before and after applying the EE method. The cleaner(Non Augmented) galaxies (NPC) also have their results denoted as seen.

Then, with the help of these datasets, we choose a random number of reliable and unreliable galaxies which are then passed through as an input to the EE method. We then see that for this particular configuration, the best possible `ContPar` exists at `ContPar`=0.140, and around it (Hyper tuning would not give much difference in results.)

With the help of this constructed ellipse, we seperate all data points into Outlier/Inlier (If it reports a value of -1, it lies outside and +1 means it lies inside), for different values of `ContPar`, and then use the machine learning frameworks.

We then apply our ADR and SOM machine learning frameworks and evaluate our metrics before and after removing outliers for different `ContPar`. In our results table, we report the values pertaining only to `ContPar`=0.140, but we also note that the trends would be similar for different values of around this neighbourhood. This would conclude our analysis.

Conclusions

The aim of our project was to build a suitable set of filters so that the data would not be skewed by unreliable readings. We performed these operations to get optimally get rid of Outliers introduced to the dataset by some technical biases, which could have a great impact(of around 5%) on some outlier measuring metrics, which is indeed significant, as stated in [3] (where they show that for a

weak lensing survey, even 1% of outilers degrades the analysis).

In recent statistical applications to astrophysics, due to the large number of precise methods of classification and regression, working under the assumption that the input data is unbiased and without outliers is dangerous, and this all the more emphasises the importance of anomaly detection.

The need for accurate redshifts has been highlighted by applications in weak lensing tomography which has been one of the main probes in ongoing and future surveys and cosmological experiments such as DES [4], LSST [20], Euclid [13] etc.

A standard LSST 'gold sample' (S/N > 20)takes around 100 hours on a 10m telescope to determine a redshift of a single object per fibre.

Even with a next generation 5000 fibre telescope, it would take around 50,000 telescope years to measure LSST gold samples (Around 4 billion galaxies).

Given that the world's biggest telescope would take 50,000 years to complete the necessary observations, it would be highly impractical to move on solely with this method (spectroscopy), and so we rely on other methods.

The requirements of these surveys say that any error can go to some limit which is photometrically achievable through the correct methods, and this is why the photometric method is extremely relevant.

The author would like to add that the data had not been k-corrected, as in [10], due to a number of problems faced, but this would not change the process of analysing outliers at all. A recent search for 'Outliers' on the ADS

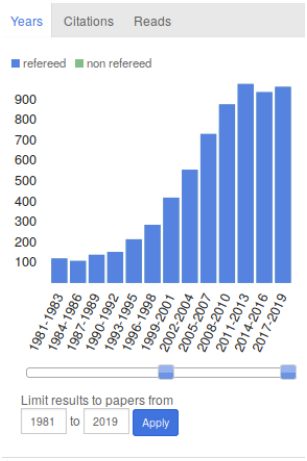

Fig. 11. A search for 'Outliers' on the ADS reveals the above.

proves that this field has produced a great inflow of techniques, with number of published papers increasing 5-fold over 15 years. We thus conclude this paper, highlighting the need for removal of unreliable data samples.

Appendix

We use the following query to CasJobs to form the basis of our dataset, from the DR15 schema.

```
select s.specObjID, q.objid,
s.z as specz, s.zerr as specz_err,
q.dered_u,q.dered_g,q.dered_r,q.dered_i,
q.dered_z,
q.modelMagErr_u,q.modelMagErr_g,
q.modelMagErr_r,
q.modelMagErr_i,q.modelMagErr_z,
q.petroRad_r,q.petroRadErr_r,
s.sourceType as specType,
q.type as photpType,s.zWarning
into mydb.specPhotoDR15v2
from SpecObjAll as s
join photoObjAll as
q on s.bestobjid=q.objid
and q.dered_g>0 and q.dered_r>0
and q.dered_z>0 and q.type=3
```

References

[1] Aguado, D. S., Ahumada, R., Almeida, A., Anderson, S. F., Andrews, B. H., Anguiano, B., Ortíz, E. A., Aragón-Salamanca, A., Argudo-Fernández, M., Aubert, M., et al. The fifteenth data release of the sloan digital sky surveys: first release of manga-derived quantities, data visualization tools, and stellar library. *The Astrophysical Journal Supplement Series 240*, 2 (2019), 23.

[2] Bundy, K., Bershady, M. A., Law, D. R., Yan, R., Drory, N., MacDonald, N., Wake, D. A., Cherinka, B., Sánchez-Gallego, J. R., Weijmans, A.-M., et al. Overview of the sdss-iv manga survey: mapping nearby galaxies at apache point observatory. *The Astrophysical Journal 798*, 1 (2014), 7.

[3] Cunha, C. E., Huterer, D., Lin, H., Busha, M. T., and Wechsler, R. H. Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. *Monthly Notices of the Royal Astronomical Society 444*, 1 (2014), 129–146.

[4] Dark Energy Survey Collaboration: Fermilab, University of Illinois at Urbana-Champaign, U. o. C. L. B. N. L. C.-T. I.-A. O., and Flaugher, B. The dark energy survey. *International Journal of Modern Physics A 20*, 14 (2005), 3121–3123.

[5] Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning 40*, 2 (2000), 139–157.

[6] Drucker, H. Improving regressors using boosting techniques. In *ICML* (1997), vol. 97, pp. 107–115.

[7] Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In *icml* (1996), vol. 96, Citeseer, pp. 148–156.

[8] Hecht, T., Lefort, M., and Gepperth, A. Using self-organizing maps for regression: the importance of the output function. In *ESANN* (2015), pp. 107–112.

[9] Hildebrandt, H. e. a. Outliers in redshift predictions.

[10] Hoyle, B., Rau, M. M., Paech, K., Bonnett, C., Seitz, S., and Weller, J. Anomaly detection for machine learning redshifts applied to sdss galaxies. *Monthly Notices of the Royal Astronomical Society 452*, 4 (2015), 4183–4194.

[11] Hoyle, B., Rau, M. M., Zitlau, R., Seitz, S., and Weller, J. Feature importance for machine learning redshifts applied to sdss galaxies. *Monthly Notices of the Royal Astronomical Society 449*, 2 (2015), 1275–1283.

[12] Kohonen, T. The self-organizing map. *Proceedings of the IEEE 78*, 9 (1990), 1464–1480.

[13] Laureijs, R., Amiaux, J., Arduini, S., Augueres, J.-L., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., et al. Euclid definition study report. *arXiv preprint arXiv:1110.3193* (2011).

[14] Li, N., and Thakar, A. R. Casjobs and mydb: A batch query workbench. *Computing in Science & Engineering 10*, 1 (2008), 18–29.

[15] MySQL, A. Mysql, 2001.

[16] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *Journal of machine learning research 12*, Oct (2011), 2825–2830.

[17] Riese, F. M., and Keller, S. Susi: Supervised self-organizing maps for regression and classification in python. *arXiv preprint arXiv:1903.11114* (2019).

[18] Rousseeuw, P. J., and Driessen, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics 41*, 3 (1999), 212–223.

[19] Salvato, M., Ilbert, O., and Hoyle, B. The many flavours of photometric redshifts. *Nature Astronomy 3*, 3 (2019), 212–222.

[20] Tyson, J. A. Large synoptic survey telescope: overview. In *Survey and Other Telescope Technologies and Discoveries* (2002), vol. 4836, International Society for Optics and Photonics, pp. 10–20.

[21] York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., Bakken, J., Barkhouser, R., Bastian, S., Berman, E., et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal 120*, 3 (2000), 1579.