ISSN 2347 - 3983
Volume 8. No. 9, September 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter58892020.pdf
https://doi.org/10.30534/ijeter/2020/58892020

Search for Sub-classes with in Type-Ia Supernovae using Hierarchical Agglomerative Clustering (HAC)

Neha Malik[1], Shashikant Gupta [2], Vivek Jaglan[3], Meenu Vijarania[4]
[1]Amity University, Gurgaon, Haryana, India, nehag78@gmail.com
[2]G D Goenka University Gurgaon, Haryana, India, shashikantgupta.astro@gmail.com
[3]Graphic Era Hill University, Dehradun, Uttarakhand, India, jaglanvivek@gmail.com
[4]Amity University, Gurgaon, Haryana, India, mvijarania@ggn.amity.edu

ABSTRACT
Search of sub-classes within Type Ia Supernovae is a topic of debate in cosmology. Researchers have considered diverse perspectives while searching for sub-classes in various studies. In the present study, SNe Ia have been investigated based on morphology of host galaxy and location of the progenitor within the host galaxy. Hierarchical Agglomerative Clustering technique has been applied on a sample of 43 type Ia Supernovae for the purpose. Our results show that spiral galaxies (Sb, Sc, Sd) favour brighter SNe Ia as compared to S0 and ellipticals. Also, the frequency of occurrence of brighter SNe is less towards the outer regions of the host galaxies. The reason could be the difference in metalicity among the host galaxies as well as the variation in metalicity at different locations within the host galaxy.

Key words: Clustering, luminosities, supernovae, morphology.

1 INTRODUCTION

Supernovae (SNe) explosions are among the most violent astronomical events and are linked with the death of stars. A particular sub-class referred as Supernova Type Ia plays a major role in Cosmology. SN Ia results from the thermonuclear explosion of carbon-oxygen white dwarf that accretes mass from its neighbouring donor or from merger of stars in a binary system. Explosion takes place when the mass of progenitor white dwarf exceeds over the Chandrasekhar limit of 1.4M⊙ [11]. The light-curve of Supernova Ia shows a sharp maximum and decays afterward. Observations indicate that the shape of light curve of all SNe Ia are similar and can be calibrated to measure the luminosity at the peak [4]. Thus, Supernova 1a can be considered as standard candles.

However, recent studies have shown deviations from the above standard candle approach [2, 10, 12, 13]. The decline rate of SNe Ia light curve is calculated in terms of a parameter termed as Δm_{15}, and it is strongly correlated with the peak luminosity, (M_B), of the explosion. Both parameters have played a vital role in establishing SNe Ia as standard candle. In this work, we are investigating whether the peak luminosity or decline rate depend on the nature of host galaxy or on the position of the supernova in the galaxy? An efficient classification technique is required, to find sub-types by using properties of SNe Ia. We propose to use the hierarchical agglomerative clustering technique to search for possible subgroups in the data. This is an unsupervised clustering technique as it does not require the training examples. We introduce this technique in section 3. The SNe Ia data used for our analysis is presented in section 2. The results and conclusions have been presented in section 4 and section 5 respectively.

2 DATA SET

A dataset of 43 Supernovae has been collected from various well-established research work and catalogue in the field. The dataset has been presented in Table 1. Column 1 contains tag of SN Ia [1–3]. Column 2 and 3 contains SN offset from the galaxy nucleus, in the E/W and N/S direction respectively [5]. Supernova offset is determined using equation:

$$offset = \sqrt{(x^2 + y^2)}. \quad (1)$$

where 'x' is supernova offset from the galaxy nucleus in the E/W direction and 'y' is supernova offset from the galaxy nucleus, in the N/S direction, respectively.

Column 4 provides information about morphology type code (T) of the host galaxy [6] (RC3). Column 5 and 6 contain decline rate (Δm_{15}) i.e. the amount in magnitudes of B-band light curve decays in the first 15 days after maximum light (defined by [4]) [1–3] and its measurement errors. Column 7 provides information about absolute B-band peak magnitudes (M_B) of the SN [1–3]. Column 8, 9 and 10 specify SN parent galaxy name, morphology type of the host galaxy and supernova type respectively [5, 6].

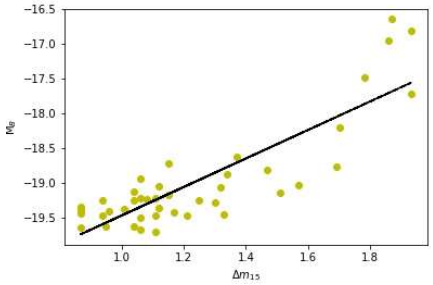

Figure 1: A graph between Δm_{15} and M_B with best fit line

2.1 Data pre-processing

Pre-processing of data is a critical process in data science and machine learning. It transforms the data into some relevant format that can be easily rendered to algorithm. Clustering process calculates distance between feature values, so it is imperative to normalize these values at some scale. This process is also referred as scaling or standardizing the dataset. Scaling is done in our analysis using following equation 2.

$$x_s = \frac{x_i - \overline{x}}{\sigma_x} \qquad (2)$$

Where i ranges from 1 to n, 'n' is the number of rows in dataset and 'x_s' is the new scaled value of parameter. In this case resultant feature values are not restricted with in some range. This method enables feature values centred around the mean with a unit standard deviation

2.2 Dealing with measurement errors

It is an obvious fact that data is prone to sampling errors. Data collection, especially in the case of astronomical data where controlled experiments are not possible, is itself a complicated process and bound to have measurement errors. To get effective clustering results, dealing these measurement errors is essential. We weight the data values according to the error, so that the data points with large error gets smaller weight. This is achieved by dividing the original data value by the square of its error as shown in the following equation 3.

$$x_i^{new} = \frac{x_i}{e_i^2} \sum_{i=1}^{n} (e_i)^2 \quad , \qquad (3)$$

where 'x_i' is the measured value and 'e_i' is measurement error. Here 'x^{new}_i' is new recalculated feature value. The decline rate (Δm_{15}) is the most vital parameter in clustering process, so its value is recalculated using equation 3. In this study, algorithm is implemented on recalculated value of this parameter. Graphs plotted in fig 9, 10 and 11 are also using the new recalculated value of Δm_{15}.

3 METHODOLOGY

Hierarchical Agglomerative Clustering (HAC) is a bottom up approach that starts from one cluster and iteratively merges clusters with other clusters until all the data items belong to one cluster [8, 9]. It is an unsupervised learning technique that uses dendrograms-a tree like structure to represent the process of clustering. In HAC, the inter-cluster distance is calculated during each step. For three-dimensional cartesian space, distance can be calculated using Euclidean distance formula given in equation 4.

$dist = \sqrt{(a^2 + b^2 + c^2)}$. (4)

Other measures like Manhattan distance can also be used for distance calculation. Based on distance, HAC uses following techniques to calculate linkage among clusters:

1. Single nearest distance or single linkage – It is distance between closest members of two clusters. If C1 and C2 are two clusters then single linkage is calculated as:

$$\min\{dist(x,y) : x \in C1,\ y \in C2\}. \qquad (5)$$

2. Complete farthest distance or complete linkage – It is distance between the farthest members of the two given clusters. If C1 and C2 are two clusters then it is calculated as:

$$\max\{dist(x,y) : x \in C1,\ y \in C2\}. \qquad (6)$$

3. Average distance or average linkage – It is the average of distance between all pairs of one cluster to every pair in other cluster. If C1 and C2 are two clusters then it is calculated as:

$$\frac{1}{|C1| \cdot |C2|} \sum_{x \in C1} \sum_{y \in C2} dist(x,y) \qquad (7)$$

2. Ward linkage - It is the distance between clusters which is calculated as summation of square of differences within all clusters. If C1 and C2 are two clusters then it is calculated as:

$$\frac{1}{|C1| \cdot |C2|} \sum_{x \in C1} \sum_{y \in C2} dist(x,y)^2 \qquad (8)$$

Steps in HAC are:

1. Find two data points with minimum/maximum distance (based on above-mentioned techniques).
2. Join these two data points to form one cluster.
3. Consider this new cluster as one data point.
4. Repeat steps 1-2 until all data points get clustered into a single cluster.

Table 1: Table of 43 Supernovae

SN Name	x	y	T	Δm_{15}	Δm_{15}-errors	M_B	Galaxy	Morphology-Type	SN-Type
SN1992A	3	62	-1.9	1.47	0.05	-18.81	NGC1380	S0	Ia
SN1990N	63.2	1.8	3.8	1.08	0.05	-19.23	NGC4639	SBbc	Ia
SN1994D	9	7	-2	1.32	0.05	-19.06	NGC4526	S0	Ia
SN1998bu	4.3	55.3	2	1.04	0.05	-19.12	NGC3368	Sab	Ia
SN1981B	41	41	4.5	1.11	0.07	-19.21	NGC4536	SBbc	Ia
SN1991bg	2	57	-4.7	1.93	0.1	-16.81	NGC4374	E	Ia pec
SN1993H	1	12.3	1.9	1.7	0.1	-18.2	E445-G66	SBab	Ia
SN1989B	15	50	3	1.34	0.07	-18.87	NGC3627	SBb	Ia
SN2003kf	9.2	14.3	3	1.01	0.05	-19.37	M-02-16-02	Sb?	Ia
SN1996X	52	31	-5	1.25	0.05	-19.24	NGC5061	E0	Ia
SN1999ee	13	3.5	4	0.94	0.04	-19.46	IC5179	Sbc	Ia
SN2003du	8.8	13.5	8	1.06	0.06	-18.93	UGC 9391	SBdm	Ia
SN2001el	22	19	5.9	1.15	0.04	-18.71	NGC1448	Sc	Ia
SN1997br	20.6	51.6	7	1.04	0.15	-19.62	E576-G40	SBd:pec	Ia pec
SN1999cw	21.1	1.5	1.5	0.94	0.05	-19.24	M-01-02-01	SBab pec:	Ia
SN1991T	26	45	3.8	0.95	0.05	-19.62	NGC4527	SBbc	Ia pec
SN1983G	17	14	-2.2	1.37	0.1	-18.62	NGC4753	S0	Ia
SN2002bo	11.6	14.2	1	1.17	0.05	-19.42	NGC3190	Sa pec	Ia
SN2002er	12.3	4.7	1	1.33	0.04	-19.45	UGC10743	Sa?	Ia
SN1984A	15	30	1	1.21	0.1	-19.46	NGC4419	SBa	Ia
SN1989A	21	18	4.1	1.06	0.1	-19.21	NGC3687	SBbc	Ia
SN2002dj	8.9	2.8	-5	1.12	0.05	-19.05	NGC5018	E3:	Ia
SN1999by	100	91	3	1.87	0.1	-16.64	NGC2841	Sb	Ia pec
SN1997cn	6.8	11.7	-5	1.86	0.1	-16.95	NGC5490	E	Ia pec
SN1986G	120	60	-2.2	1.78	0.07	-17.48	NGC5128	S0	Ia pec
SN1990O	21.8	3.9	1	0.96	0.1	-19.4	M+03-44-03	SBa	Ia
SN1990T	24.8	1.9	-2	1.15	0.1	-19.17	PGC0063925	S0	Ia
SN1991S	4.4	17.3	1.8	1.04	0.1	-19.24	UGC 5691	Sab	Ia
SN1991U	2.2	5.8	3.8	1.06	0.1	-19.49	IC4232	Sbc	Ia
SN1991ag	4.4	22.1	7.9	0.87	0.1	-19.4	IC4919	SBd	Ia
SN1992K	1.9	15.4	1.9	1.93	0.1	-17.72	E269-G57	SBab	Ia pec
SN1992P	4.3	9.8	4	0.87	0.1	-19.34	IC3690	Sbc	Ia
SN1992al	19	12	5.1	1.11	0.05	-19.47	E234-G69	SBc:	Ia
SN1992bc	16	5	5	0.87	0.05	-19.64	E300-G09	Sc	Ia
SN1992bk	12	21	-5	1.57	0.1	-19.03	E156-G08	E	Ia
SN1992bl	15	22	1	1.51	0.1	-19.13	E291-G11	SBa	Ia
SN1992bo	47.3	54.7	-1.5	1.69	0.05	-18.76	E352-G57	S0/a	Ia
SN1993ah	1	8	-2	1.3	0.1	-19.28	E471-G27	S0	Ia
SN1937C	30	40	8.9	0.87	0.1	-19.39	IC4182	Sm	Ia
SN1960F	38	24	8.2	1.06	0.12	-19.67	NGC4496A	SBd	Ia
SN1972E	38	100	8	0.87	0.1	-19.44	NGC5253	Sd	Ia
SN1998aq	18	7	3	1.12	0.03	-19.35	NGC3982	Sb:	Ia
SN1974G	32	14	5	1.11	0.06	-19.7	NGC4414	Sc	Ia

Hierarchical Agglomerative Clustering (HAC) creates clusters at different levels with granularity while disclosing the hidden structure in dataset in detail. Some experts' tools like R, MATLAB and Python are available with user friendly libraries to implement Hierarchical Agglomerative Clustering (HAC) [19, 20] and other clustering algorithms [18]. The process followed from data collection to result interpretation is displayed in figure 2.

3.1 Validation of clustering results.

Validation of clusters obtained after applying algorithm is done to justify the clustering results. In Hierarchical Agglomerative Clustering (HAC), dendrograms display the process of clustering and can also act as validity tools to optimise the process of clustering [15].

In HAC, as groups become larger, they become more dissimilar. The height of dendrograms represents dissimilarity/distance and width represents the sample index [16]. The height of every parent node in dendrograms is

proportional to the value of dissimilarity between itself and its children. Generally, in dendrograms, longest vertical lines that are not intersecting any horizontal line (horizontal line represents clusters formation) are searched. Then a new horizontal line (the cut-off) is drawn through these vertical lines at both extremities [15]. The number of vertical lines passed by this new horizontal line is optimal number of clusters.

Another technique commonly used for validation of clusters is silhouette score technique [7,18]. Silhouette Score Coefficient finds how much the data points are integrated in their respective cluster as compared to other clusters. Silhouette coefficient value lies between +1 and -1. The best score is +1 and the worst score is -1. Value 0 shows that clusters are overlapping. The Silhouette Coefficient is calculated as:

$$S = \frac{b - a}{max(a, b)} \quad (9)$$

where 'a' is the mean value of intra-cluster distance and 'b' is the mean value of nearest-cluster distance. Figure 3, 4 and figure 5, 6 shows dendrogram and silhouette score plot for morphology code vs. decline rate (T -Δm_{15}) and location within galaxy vs. decline rate (offset-Δm_{15}) respectively.

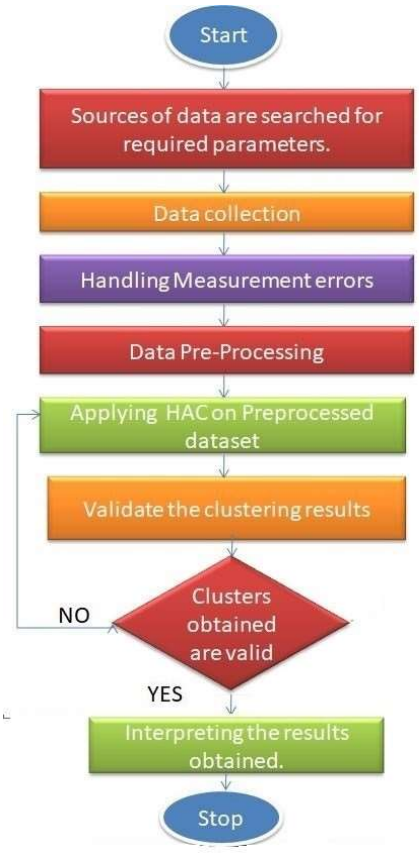

Figure 2.: Flowchart of Research Process

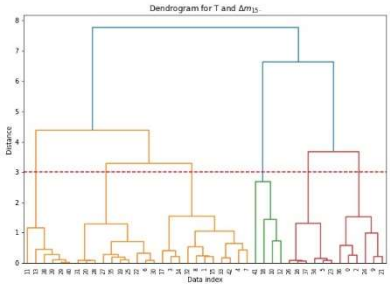

Figure 3: Cluster validation with Dendrogram for T and Δm_{15}

Similarly, figure 7, 8 shows dendrogram and silhouette score plot when HAC is implemented on combination of three parameters i.e. morphology code, decline rate and location within galaxy (T, offset and Δm_{15}).

Figure 4: Cluster validation with Silhouette score for T and Δm_{15}

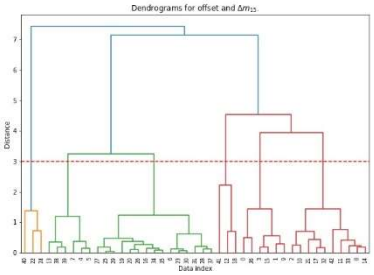

Figure 5: Cluster validation with Dendrogram for offset and Δm_{15}

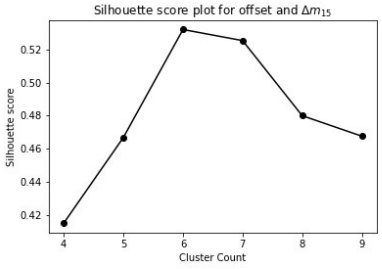

Figure 6: Cluster validation with Silhouette score for offset and Δm_{15}

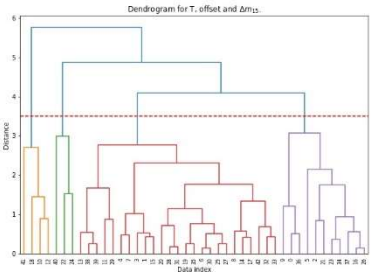

Figure 7: Cluster validation with Dendrogram for T, offset and Δm_{15}

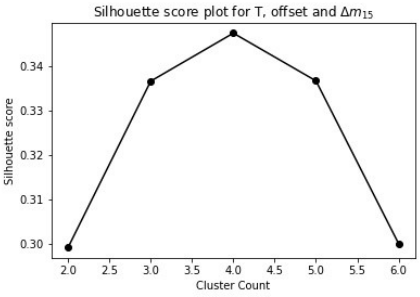

Figure 8: Cluster validation with silhouette score for T, offset and Δm_{15}

Table 2: Correlation among various parameters of 43 SNe

SN.Parameters	Correlation
$\Delta m_{15} - M_B$	0.8506
T - Δm_{15}	-0.5971
T - M_B	-0.4333
offset - Δm_{15}	0.2790

4 RESULTS

First, correlation coefficient is calculated among various columns of the data presented in table 1. The numerical values of these coefficients are presented in table 2. A strong correlation between Δm_{15} and M_B as expected can be seen in the first row of table 2 and in figure 1 as well. However, there exist a significant correlation between Δm_{15} and T that indicates the dependence of SNe decline rate on galaxy morphology. Firstly, Hierarchical Agglomerative Clustering (HAC) is implemented on the pair of Δm_{15} and T. Figure 3 shows the dendrogram of this implementation which exhibits 6 clusters within the set of 43 SNe. Figure 4 validates this result through Silhouette score. The properties of these clusters in terms of correlations among various parameters are shown in table 3. Table 4 shows the gross properties of these clusters. It shows that on average the brighter SNe (smaller M_b) favour high value of T. This means that brighter SNe prefer the spiral galaxies (Sb, Sc and Sd). On the other hand, the fainter SNe with larger M_b favour host galaxies with smaller T which belong to either elliptical or lenticular (S0) galaxies.

Next, the HAC is applied on the pair of offset and Δm_{15}. The dendrogram along with the Silhouette score graph is shown in figure 5 and 6 respectively. Here also 6 clusters are observed within the set of 43 SNe of table 1. The correlations among various parameters for these 6 clusters are shown in table 5 and the gross properties of these clusters are presented in table 6. Again, it is observed that the brighter SNe belong to smaller offset values, i.e. close to the centre of the host galaxy. Then, HAC is implemented on three parameters i.e. T, offset and Δm_{15}. Table 7 contains correlation that is quite good, within clusters when HAC algorithm is implemented on T, offset and Δm_{15}. Table 8 shows the mean value of parameters within each cluster and it also shows that higher luminosities exists at lower value of offset and higher value of morphology T. Figure 9, 10 and 11 displays the graph showing clusters formation for each case.

5 CONCLUSIONS

HAC technique is implemented on a set of 43 Type Ia SNe to test the dependence of SNe brightness on host galaxy morphology and the location of SNe in host galaxy. Results indicate that most luminous SNe (smaller absolute magnitude) occur in spiral (Sb, Sc, Sd) galaxies, i.e., late type galaxies. On the other hand, dimmer SNe Ia occur in both spiral (S0) as well as elliptical galaxies.

Table 3: Correlations within clusters when HAC is implemented on T and Δm_{15}.

Cluster no.	Data points	Δm_{15}, M_B	T, Δm_{15}
1	11	0.5643	-0.1193
2	6	0.7946	0.7588
3	4	0.0673	-0.5536
4	6	0.0398	-0.4299
5	6	0.9117	-0.8822
6	10	0.8663	-0.2499

Table 4: Mean values within clusters when HAC is implemented on T and Δm_{15}.

Cluster no.	Clusters	$\overline{M_B}$	$\overline{T}$
1	11	-19.3536	3.4273
2	6	-18.7333	-2.9333
3	4	-19.2425	3.475
4	6	-19.4083	8
5	6	-18.31	-3.4833
6	10	-18.783	2.35

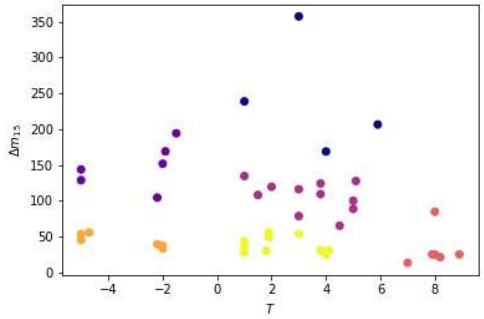

Figure 9: A graph between T and Δm_{15} with six clusters

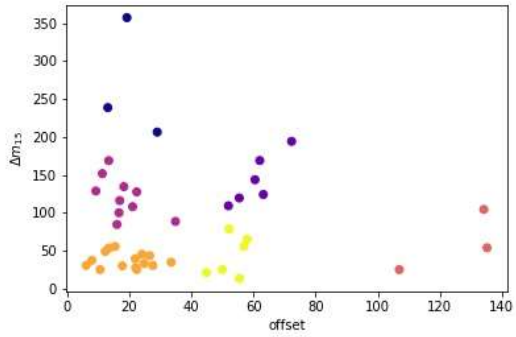

Figure 10: A graph between offset and Δm_{15} with six clusters.

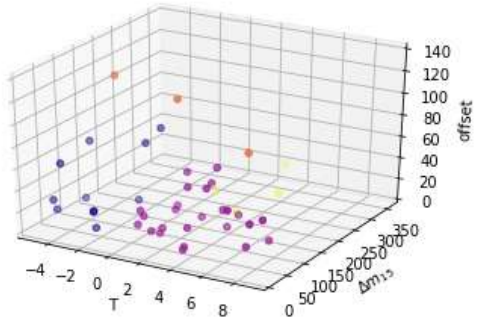

Figure 11: A graph between T, offset and Δm_{15} with four clusters.

Table 5: Correlations within clusters when HAC is implemented on offset and Δm_{15}.

Cluster no.	Data points	Δm_{15}, M_B	offset,M_B	offset,Δm_{15}
1	3	-0.4934	0.9642	-0.7064
2	6	0.8767	0.7848	0.8714
3	10	0.3852	-0.6365	-0.0857
4	3	0.9769	0.9653	0.9988
5	15	0.8417	-0.2569	-0.1199
6	6	0.9577	0.4761	0.4399

Table 6: Mean values within clusters when HAC is implemented on offset and Δm_{15}.

Cluster no.	Clusters	$\overline{M_B}$	$\overline{\text{offset}}$
1	3	-19.17	20.5165
2	6	-19.13	60.9317
3	10	-19.334	18.0966
4	3	-17.8533	125.449
5	15	-18.9093	19.1865
6	6	-18.9283	52.954

Table 7: Correlations within clusters when HAC is implemented on T, offset and Δm_{15}.

Cluster no.	Data points	Δm_{15},M_B	T, Δm_{15}	offset,M_B	offset,Δm_{15}
1	11	0.8464	-0.2426	0.1252	0.3659
2	25	0.8277	-0.4261	-0.2236	-0.150
3	3	0.9769	-0.8159	0.9653	0.9988
4	4	0.0672	-0.5536	0.9682	0.0248

Table 8: Mean values within clusters when HAC is implemented on T, offset and Δm_{15}.

Cluster no.	Clusters	$\overline{M_B}$	$\overline{\text{offset}}$	$\overline{\text{T}}$
1	11	-18.62	33.216	-3.3
2	25	-19.24	30.929	3.928
3	3	-17.85	127.63	2.933
4	4	-19.24	18.753	3.475

Variation in the composition of different types of galaxies could be a leading cause of this distribution. In terms of location within galaxies, it conclude that brighter Supernovae Ia occur near the nucleus of galaxies due to abundance of heavy elements On moving towards outer regions of galaxies the abundance of heavy elements decreases, and hence, luminous supernovae Ia are rare in the outer regions. On comparing the results with previous literature [14, 17] it is found that our results coincide with the results of previous existing studies. Thus, it reckons that there is a good proportion of possibility on existence of sub-classes with in SNe Ia based on variation in metallicity.

REFERENCES

1. Mario Hamuy et al, "**The absolute luminosities of the calan/tololo Type Ia Supernovae**", The Astrophysical Journal Vol. 623, no. 2, pp.1011–1016, 2005.
2. S.Benetti et al, "**The diversity of Type Ia Supernovae: evidence for systematics?"** The Astrophysical Journal Vol. 112, no. 6, pp.2391–2396, 1996.

3. A.Saha et al,, "**Cepheid Calibration of the Peak Brightness of SNe Ia. XI. SN 1998aq in NGC 3982**", The Astrophysical Journal Vol. 562, no. 1, pp.314, 2001.
4. M.M. Phillips et al, "**The Absolute Magnitudes of Type Ia Supernovae.**" The Astrophysical Journal Vol.413, no.2, p.p. L105–L108, 1993.
5. Barbon et al, "**Asiago Supernova Catalogue**." Astronomy and Astrophysics Supplement Vol.81, no. 3/DEC, p.p. 421, 1989.
6. Harold G. Corwin, Jr. et al, "**Corrections and Additions to the Third Reference Catalogue of Bright Galaxies**", The Astrophysical Journal Vol. 108, Number 6, p.p. 2128-2144, 1994.
7. Peter J.Rousseeuw, "**Silouettes: a graphical aid to the interpretation and validation of cluster analysis**", Journal of Computational and Applied Mathematics, Vol. 20,p.p. 53–65, 1987.
8. Daniel Mullner, "**fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python**", Journal of Statistical Software, Vol. 53, no. 9, May 2013.
9. Daniel Mullner, "**Modern hierarchical, agglomerative clustering algorithms**".stat.ml, 2011.
10. Chiaki Kobayashi1, "**Subclasses of Type IA Supernovae as the origin of [*α*/FE] ratios in dwarf spheroidal galaxies**".The Astrophysical Journal Letters, Vol. 804,p.p. L24, 2015.
11. P.A.Mazzali et al, "**A Common Explosion Mechanism for Type Ia Supernovae"**, Science Vol. 315 (5813): 825–828, 2007.
12. X. Wang et al, "**Improved Distances to Type Ia Supernovae with two Spectroscopic Subclasses**", The Astrophysical Journal, Vol. 699, p.p.L139–L143, 2009a.
13. David Branch et.al, "**Comparative Direct Analysis of Type Ia Supernova Spectra. II. Maximum Light**", Publications of the Astronomical Society of the Pacific, Vol. 118, Number 842, 560-571, 2006.
14. Hideyuki Umeda et.al. "**The Origin of Diversity of Type IA Supernovae and Environmental Effects**", The Astrophysical Journal Letters, Vol. 522, Number 1,1999.
15. Antoine E. Zambellia, "**A data-driven approach to estimating the number of clusters in hierarchical clustering**", F1000 Research, Vol. 5, 2016.
16. NuraKawa, "**Agglomerative Hierarchical Clustering**", Online tutorials, 2017.
17. M. Sullivan et al, "**The Dependence of Type Ia Supernova Luminosities on their Host Galaxies**", Monthly Notices of the Royal Astronomical Society, Vol. 406, Issue 2, p.p. 782-802, 2010.
18. Neha Malik et.al, "**Supernova Type Ia Diversity: A Study using DBSCAN Algorithm**", IJATCSE, Volume 9, No.3, p.p. 3398-3402, May-June 2020.
19. Aida Mustapha et al, "**Machine Learning Supervised Analysis for Enhancing Incident Management Process**", International Journal of Emerging Trends in Engineering Research, Vol. 8, no. 1.1, pp. 199-204, 2020.
20. AzrelAiman Azeman et al, "**Football Match Outcome Prediction by Applying Three Machine Learning Algorithms**", International Journal of Emerging Trends in Engineering Research, Vol. 8, no. 1.1, pp. 73-79, 2020.