COMPARISON OF CATALOGS

A PREPRINT

September 23, 2024

Table of contents

1 The data

In this script we will compare 2 catalogs Kovlakas et al. (2021) and Karachentsev and Kaisina (2013)

- The data have been joined based on their position in the sky (Ra, Dec).
  - We assume that every galaxy within 2 arc seconds of the initial coordinates is the same galaxy.
- We use TOPCAT to create two joins, an inner and an outer join
- We will use the inner join for 1-1 comparisons
- If we see that the data are similar we can use the outer join
- For the comparison we keep the parameters names exactly they are given in the catalogs

The dataset we are going to use for the comparison (inner join) consists of 288 galaxies and 168 columns.

2 Catalog Completeness

Checking for completeness in galaxy catalogs is essential to ensure that the data accurately represents the true population of galaxies. Incomplete catalogs can lead to biased results in statistical studies, such as the distribution of galaxy luminosity, mass, or star formation rates. Additionally, missing galaxies, especially those at faint magnitudes or large distances, can distort cosmological measurements and hinder our understanding of galaxy formation and evolution.

Completeness checks are crucial for addressing selection biases, ensuring accurate redshift distributions, and validating galaxy simulations. They help identify gaps in the data and guide follow-up observations, ensuring that the catalog provides a reliable sample for scientific analysis. Without these checks, conclusions drawn from the data may be inaccurate or incomplete.

Checking for completeness in galaxy catalogs is essential to ensure that the data accurately represents the true population of galaxies. Incomplete catalogs can lead to biased results in statistical studies, such as the distribution of galaxy luminosity, mass, or star formation rates. Additionally, missing galaxies, especially those at faint magnitudes or large distances, can distort cosmological measurements and hinder our understanding of galaxy formation and evolution. Completeness checks are crucial for addressing selection biases, ensuring accurate redshift distributions, and validating galaxy simulations.

Distance-based corrections are applied to mitigate these biases by adjusting for the underrepresentation of galaxies at greater distances. As galaxies move farther away, they become fainter and harder to detect, leading to a drop in the number of detected galaxies. Methods like **volume corrections** (e.g., V/Vmax) and **luminosity function-based corrections** help account for these effects by estimating the true galaxy population based on the observed sample. These corrections ensure that statistical analyses, even in incomplete catalogs, more accurately reflect the full galaxy population.

Table	Number of galaxies
Inner join	288
Outer join	2901
LCV	1316
HECATE	2901
Unique galaxies in LCV	1028
Unique Galaxies in Hecate	2613

2.1 Completeness of the Inner join

$$\text{Completeness (X)} = \frac{\text{(Galaxies in Inner Join)}}{\text{(Galaxies in X)}} \times 100\%$$

Completeness (HECATE)= 10 %

Completeness (LCV)= 22 %

2.2 Completeness in Outer join

$$\text{Completeness (X)} = \frac{\text{(Galaxies in Outer Join form X)}}{\text{(Galaxies in X)}} \times 100\%$$

Completeness (HECATE)= 90 %

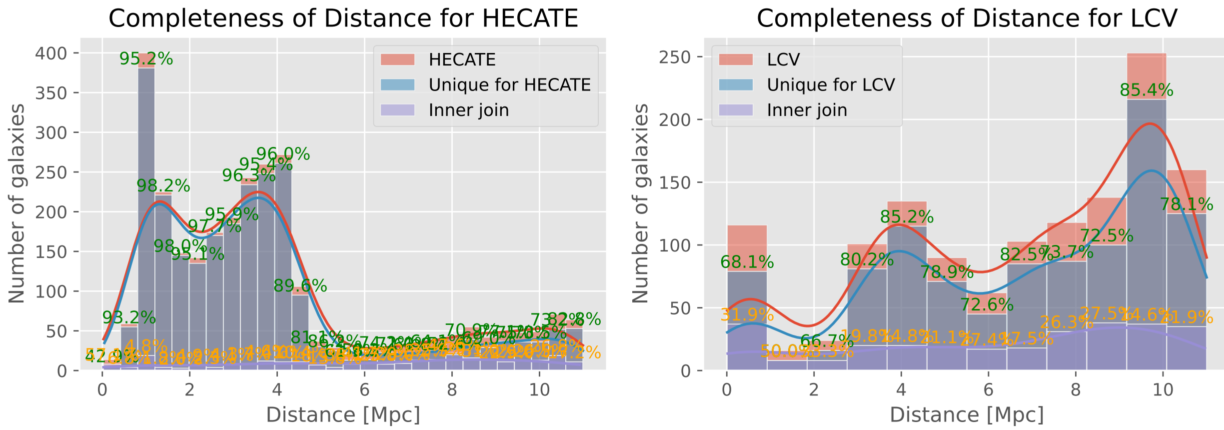
Completeness (LCV)= 78 %

Combined Completeness $= \frac{\text{Total galaxies in Outer}}{\text{Unique galaxies in HECATE + LCV}} = 80$ %

2.3 Completeness of the Catalogs, based on the Distance and the Morphological Type

As we can see from the histograms Figure 1 and Figure 2 the sample of nique galaxies of each catalog, gets smaller by an almost constant proportion (Inner join).

This means there is no bias in the selection of the galaxies.



(a) HECATE (b) LCV

Figure 1: Histograms showing the Distance Completeness of the Catalogs

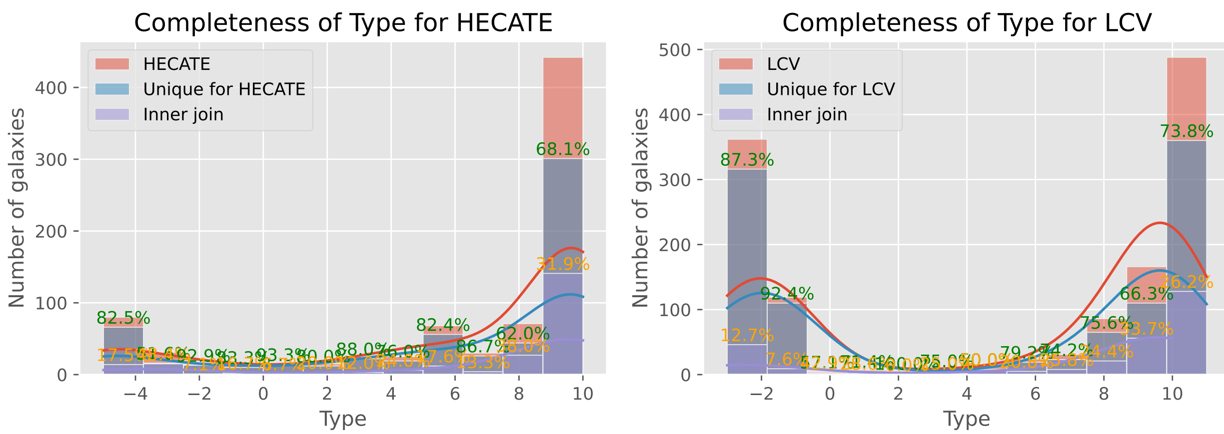
Table 2: Table showing the Completeness of the Catalogs, per bin

(a) HECATE

Bin Start	Bin End	Unique %	Inner %
0	0	43%	57%
0	1	93%	7%
1	1	95%	5%
1	2	98%	2%
2	2	98%	2%
2	2	95%	5%
2	3	98%	2%
3	3	96%	4%
3	4	96%	4%
4	4	95%	5%
4	4	96%	4%
4	5	90%	10%
5	5	81%	19%
5	6	86%	14%
6	6	62%	38%
6	6	58%	42%
6	7	74%	26%
7	7	71%	29%
7	7	62%	38%
7	8	64%	36%
8	8	52%	48%
8	9	71%	29%
9	9	69%	31%
9	9	77%	23%
9	10	75%	25%
10	10	73%	27%
10	11	73%	27%
11	11	83%	17%

(b) LCV

Bin Start	Bin End	Unique %	Inner %
0	1	68%	32%
1	2	50%	50%
2	3	67%	33%
3	4	80%	20%
4	5	85%	15%
5	6	79%	21%
6	6	73%	27%
6	7	83%	17%
7	8	74%	26%
8	9	72%	28%
9	10	85%	15%
10	11	78%	22%



(a) HECATE

(b) LCV

Figure 2: Histograms showing the Type Completeness of the Catalogs

3 How are we going to compare the data?

3.1 Scatter plots and R^2 calculation

1. R^2: Measures the proportion of variance explained by the linear model.
2. Slope of the Fitted Line: Should be close to 1 for a 1-1 correlation.[1]

3. Pearson Correlation ρ: Measures the strength and direction of the linear relationship between two variables, ranging from -1 to 1. [2]

4. Plots: Plots are essential for visually assessing the relationship between two datasets, identifying correlations, trends, and outliers, and evaluating the fit of linear models.

- Histograms: Because not all of our data have the same number of counts, the comparison with histograms between data that are not the same, doesn't help us right now.[3] This is why we will only use histograms for comparing the distribution of same-data columns normalized by their maximum value

- Correlation Heatmaps: A correlation heatmap is a graphical tool that displays the correlation between multiple variables as a color-coded matrix. It's like a color chart that shows us how closely related different variables are. In a correlation heatmap, each variable is represented by a row and a column, and the cells show the correlation between them. The color of each cell represents the strength and direction of the correlation, with darker colors indicating stronger correlations.
- Kernel Density Estimate (KDE) plot: The KDE plot visually represents the distribution of data, providing insights into its shape, central tendency, and spread.

5. Percentage change: We can calculate the percentage change of the data for each galaxy and then we can see if the data are similar, based on minimum, the maximum and the mean value of the difference.

$$\text{Percentage change} = \frac{V_{Hecate} - V_{LCV}}{V_{Hecate}} \cdot 100\%$$

4 Comparable data

4.1 Coordinates

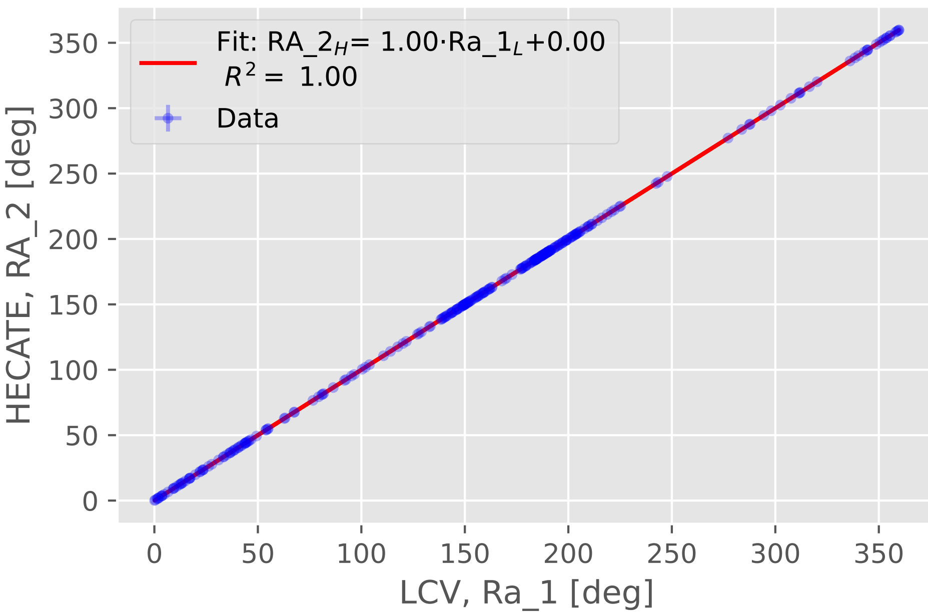
[1] Some data seem to have a very good linear correlation but they have many outliers. This is why we will clip the outliers with $\sigma > 3$

[2] In simple linear regression, R^2 is the square of the Pearson correlation coefficient ρ.

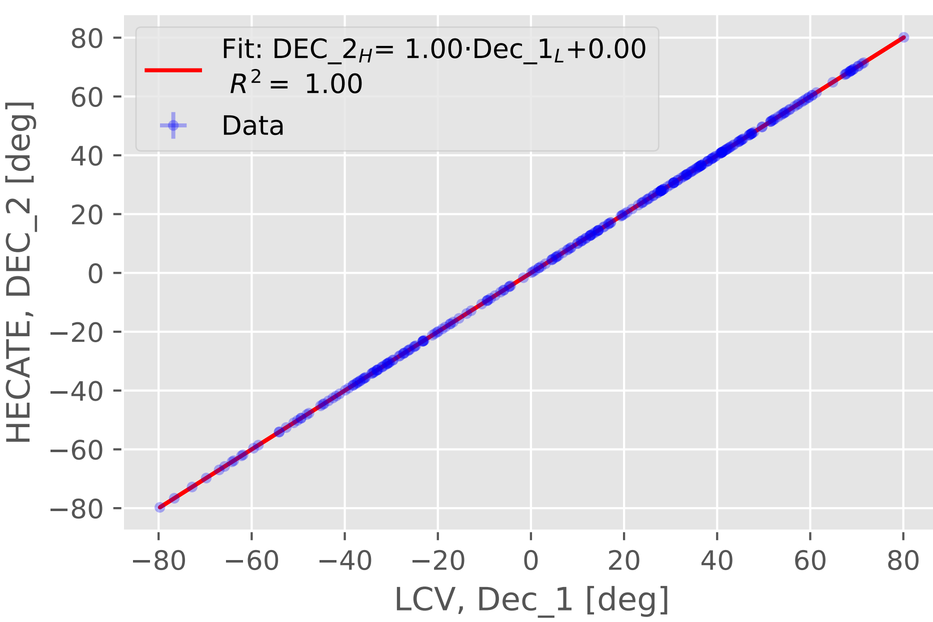
[3] When we will use the outer join table we could use histograms due to the large number of counts.

LCV	HECATE	Description	Pearson Correlation [-1,1]
Ra_1	RA_2	Right Ascension	1.0
Dec_1	DEC_2	Declination	1.0
Dis	D	Distance	0.881

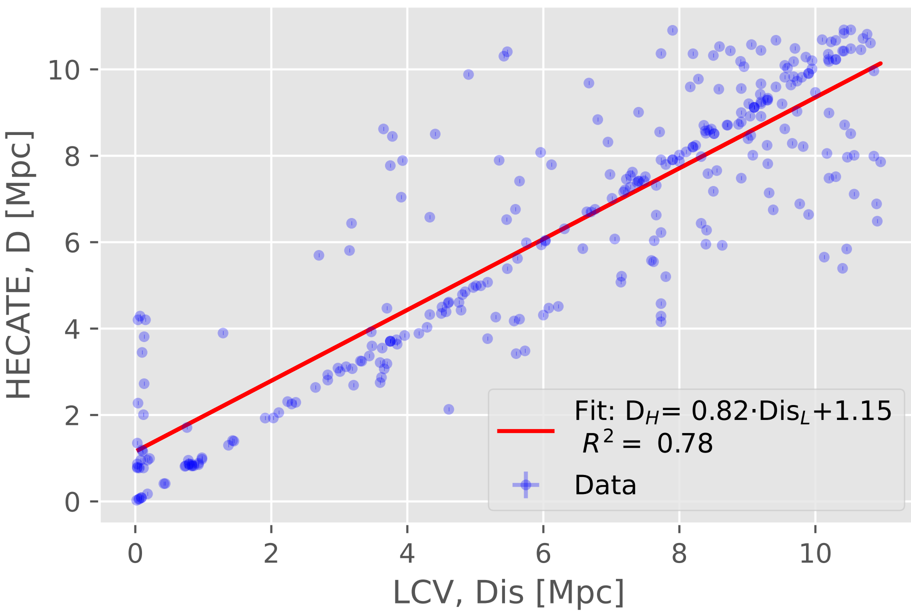
4.1.1 Right Ascension



4.1.2 Declination

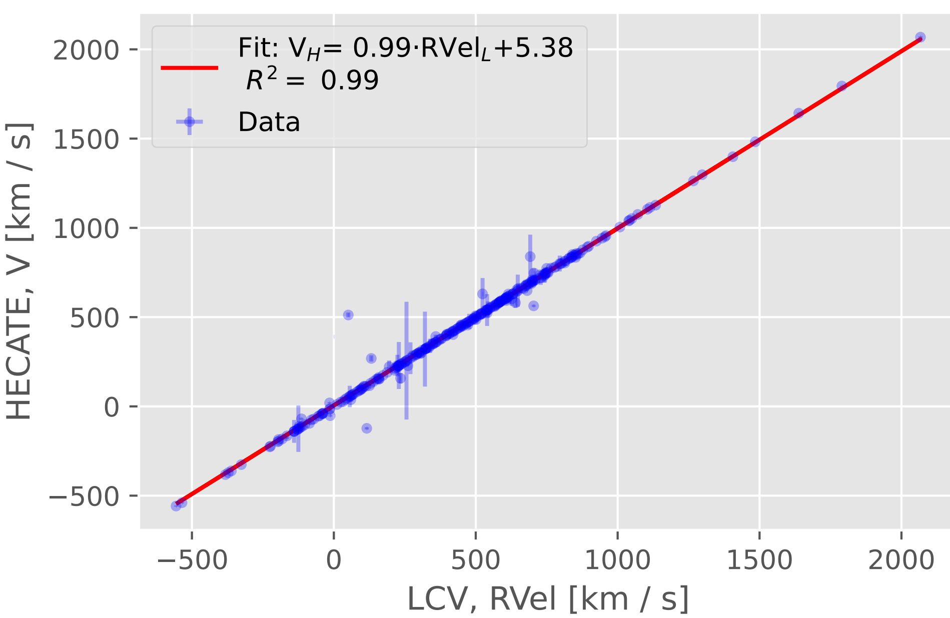


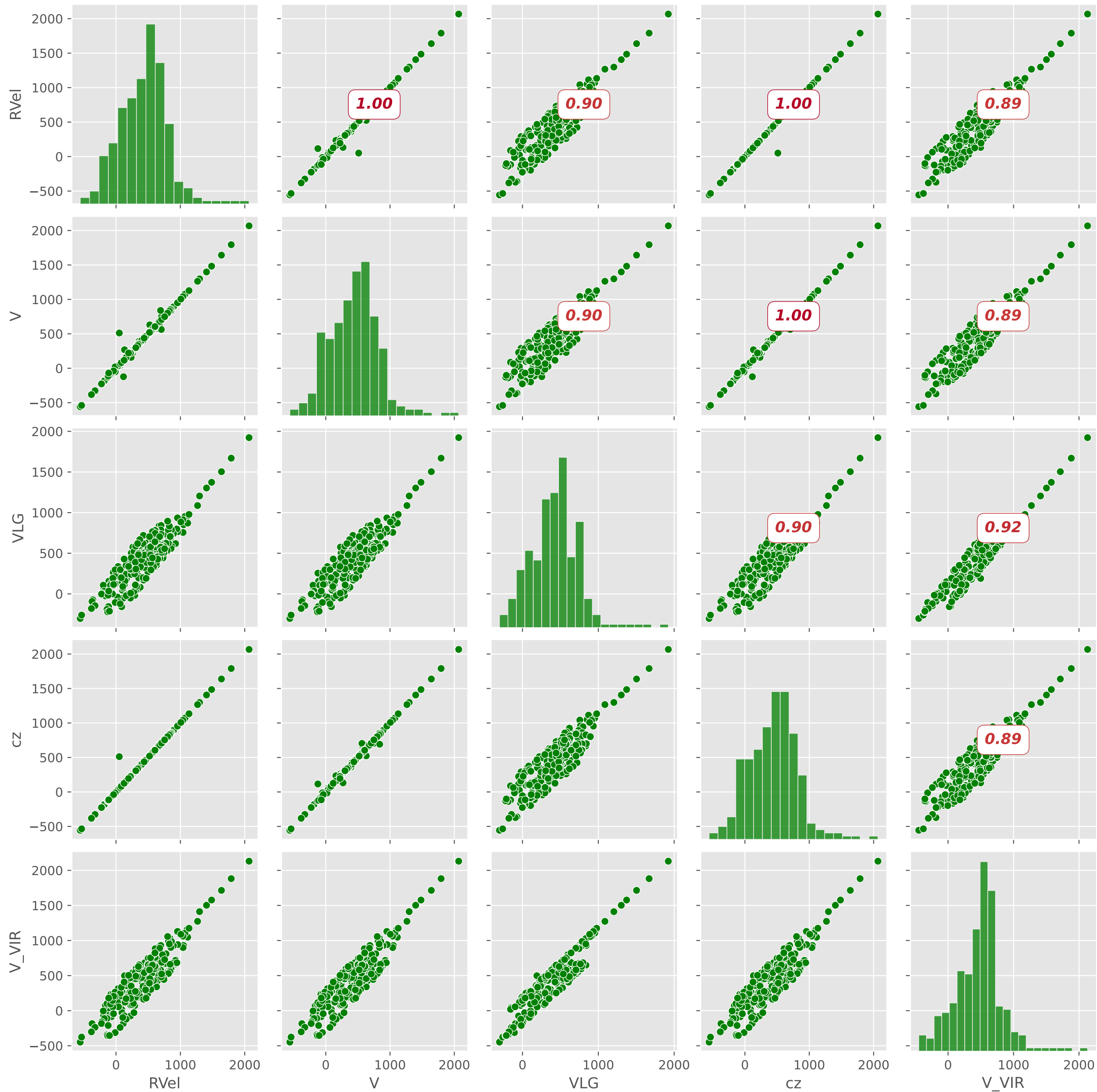
4.1.3 Distance



4.2 Velocities

LCV	HECATE	Description	Linear Correlation
RVel (km/s)	V (km/s)	Heliocentric radial velocity	0.994
VLG (km/s)		Radial velocity	
cz (km/s)		Heliocentric velocity	
	V_VIR (km/s)	Virgo-infall corrected radial velocity	



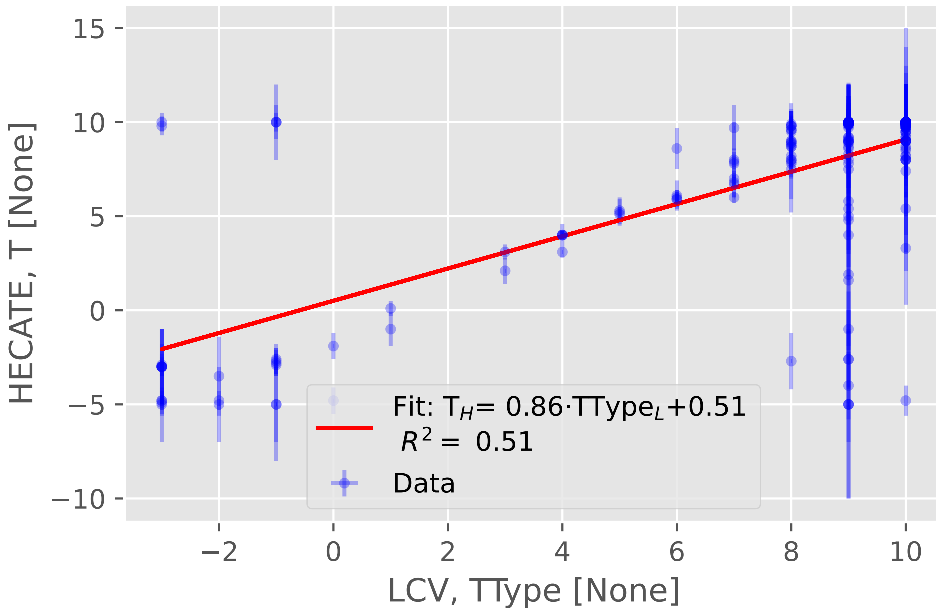


[?] The close correlation between all of the velocities, could be due to the fact that all of them measure the velocity of each galaxy, but from a different frame of reference.

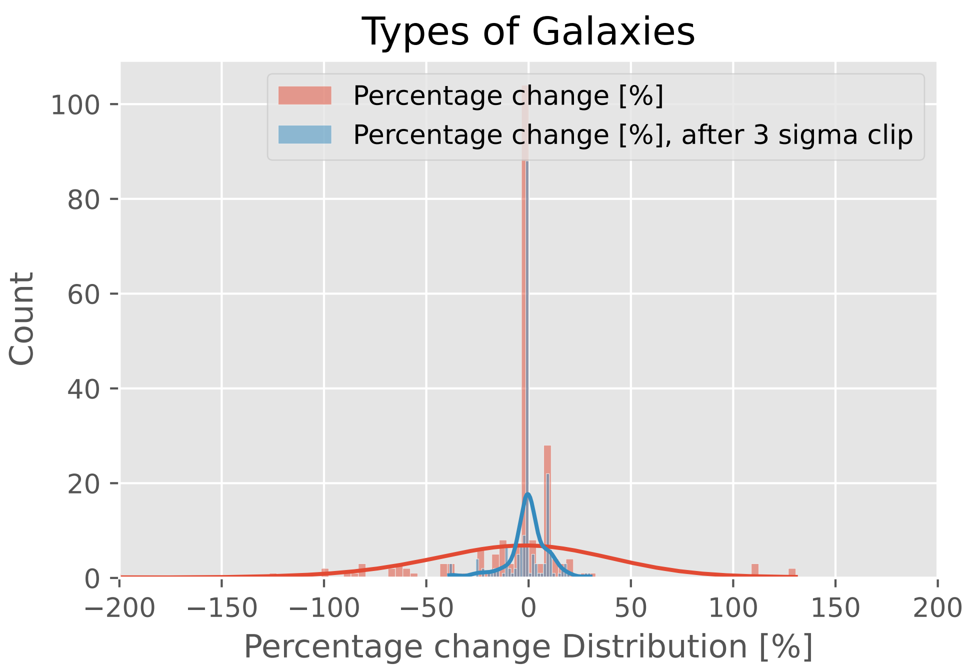
4.3 Morphology and Geometry

LCV	HECATE	Description	Pearson Correlation [-1,1]
TType	T (with errors)	Numerical Hubble type following the de Vaucouleurs system	0.7107
inc	INCL	Inclination (deg)	0
a26_1 (Major)	R1 (Semi-major axis)	angular diameter (arcmin)	0

4.3.1 Galaxy Types



Percentage change:



	diff_T	diff_T_clip
count	229	191
mean	-30	-1
std	119	10
min	-1000	-39
25%	-11	-2
50%	0	0
75%	0	2

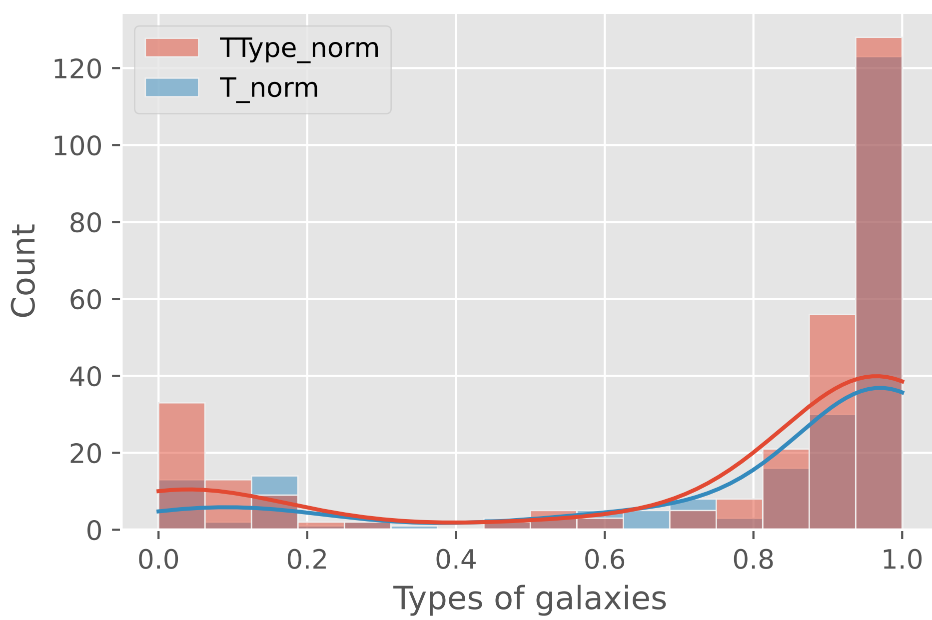
	diff_T	diff_T_clip
max	131	30

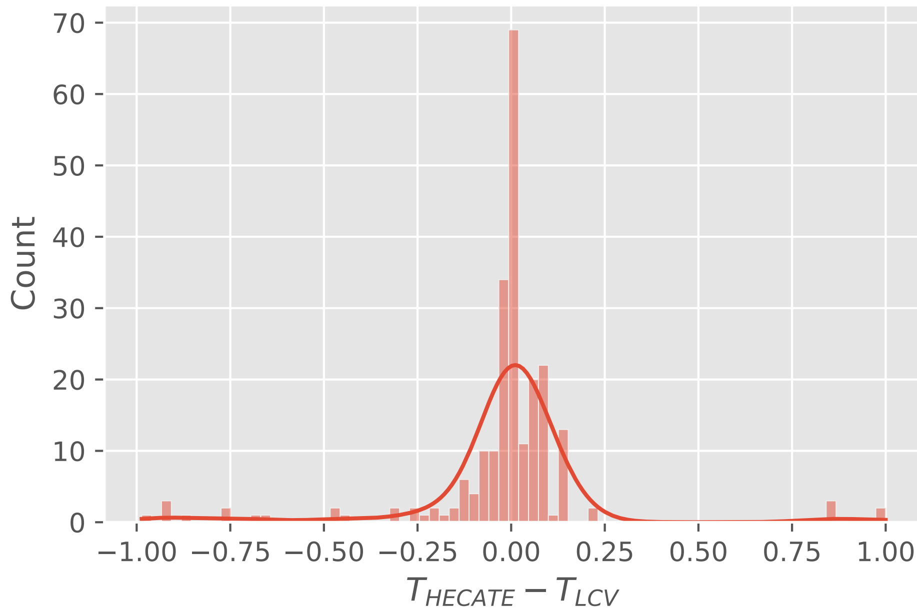
[?] After the sigma clip we only lose 39 galaxies (14%) and we can see that both the median and the mean of the percentage change are close to 0%.This is why we can assume that the Types of the galaxies are the same for the two catalogs

4.3.1.1 Normalize the scale of galaxy types It is very possible that the two catalogs use different scaling methods, as indicated by the use of decimal numbers in HECATE.

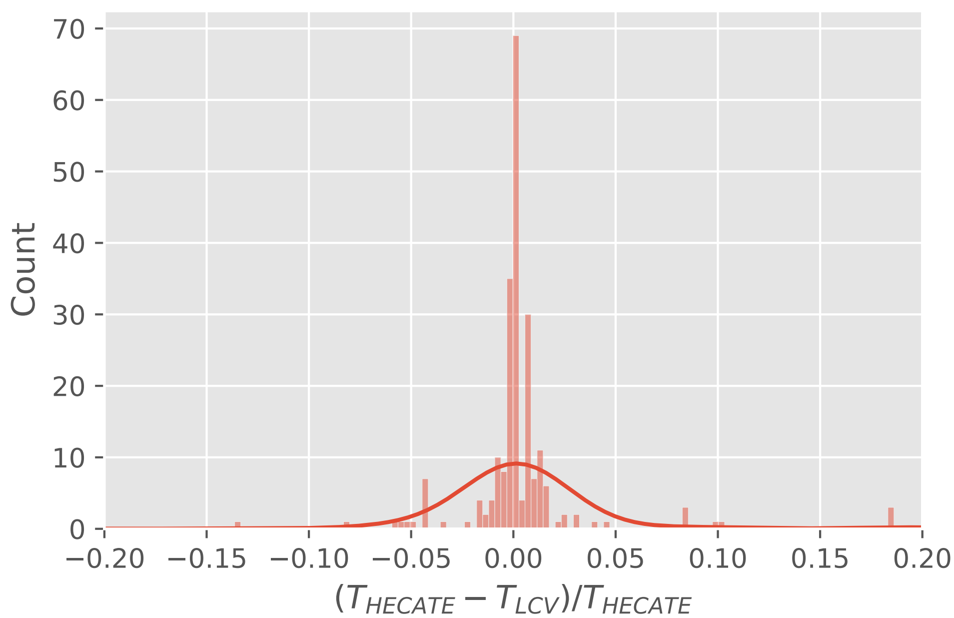
	T	TType	T_norm	TType_norm
count	229	287	229	287
mean	7	7	1	1
std	5	5	0	0
min	-5	-3	0	0
25%	7	6	1	1
50%	10	9	1	1
75%	10	10	1	1
max	10	10	1	1

Also, as we can see the minimum values are lower by 2 in HECATE, which complies with the linear fit.



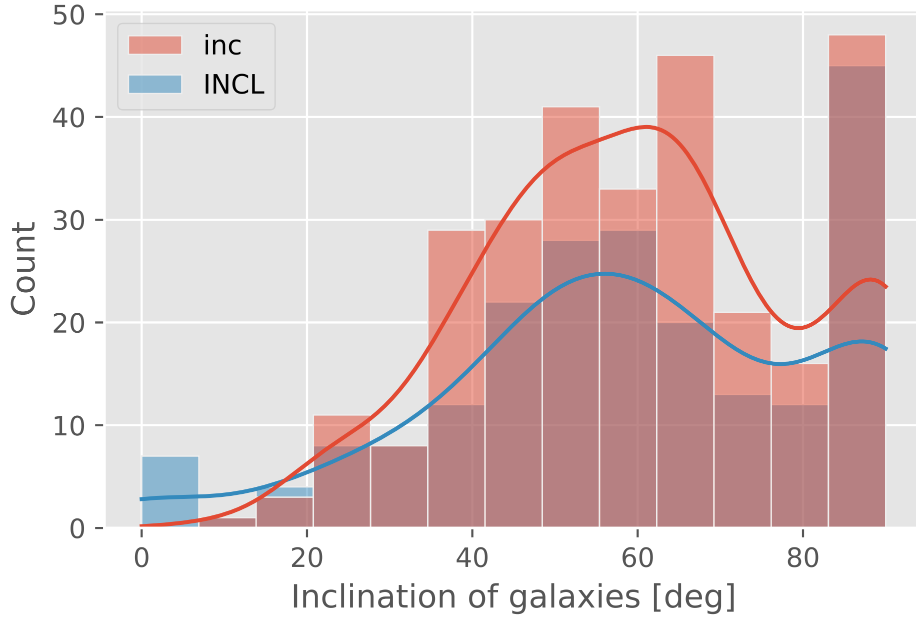


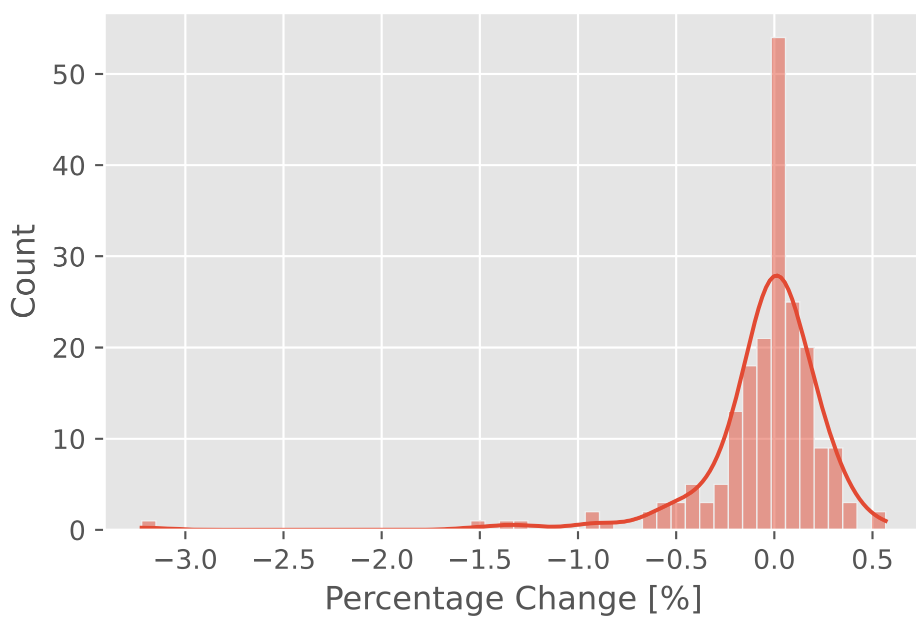
Count
70
60
50
40
30
20
10
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
$T_{HECATE} - T_{LCV}$



Count
70
60
50
40
30
20
10
0
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
0.20
$(T_{HECATE} - T_{LCV})/T_{HECATE}$

4.3.2 Inclination

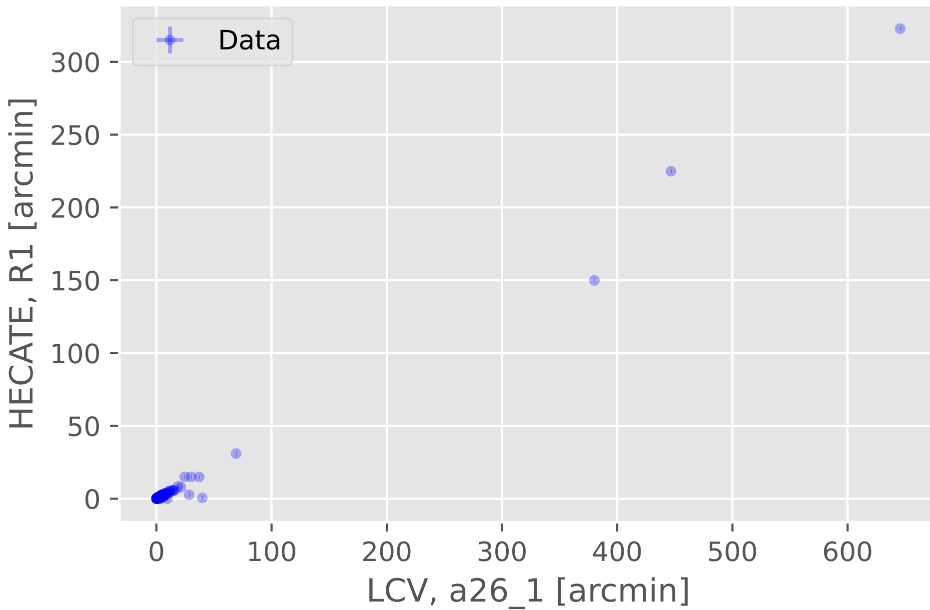




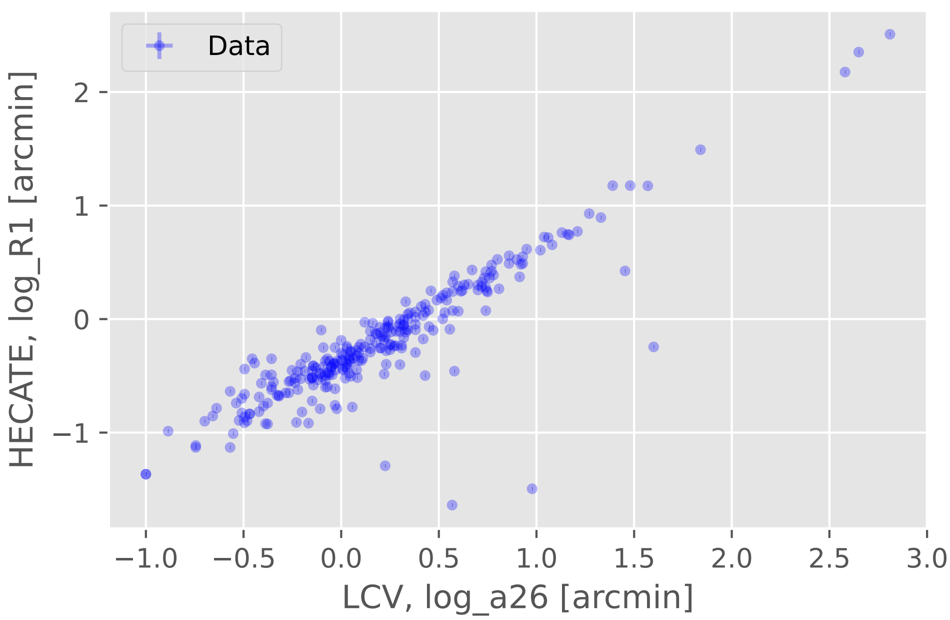
	inc	INCL	Percentage Change [%]
count	287	209	202
mean	60	59	-0
std	19	23	0
min	9	0	-3
25%	47	47	-0
50%	60	59	0
75%	72	78	0
max	90	90	1

We can see that for values in the range $[\sim 30^{\circ}, \sim 80^{\circ}]$, the values of the LCV inclination are higher. However, since their means, median, min and maxes are similar and the percentage change is practically 0% (mean, median, σ = 0 with a range $[-3\%, 1\%]$), we can ignore the differences and assume they are the same values.

4.3.3 Major Axis

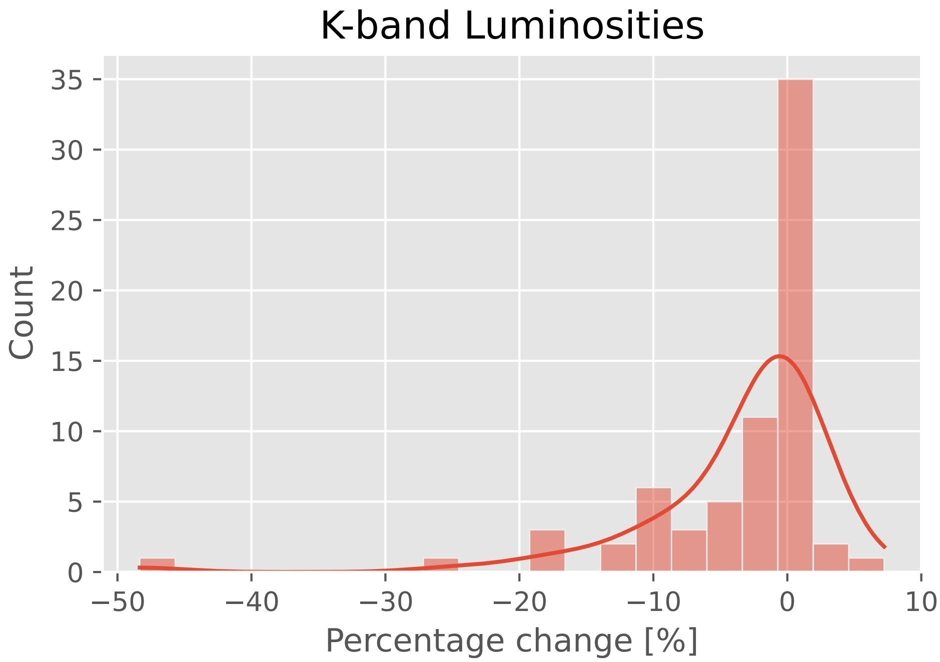


it is not very clear if we truly have a correlation or not. We need to see the linear correlation of the decimal logarithms.



4.4 Luminosities

LCV	HECATE	Description	Pearson Correlation [-1,1]
logKLum	logL_K		0

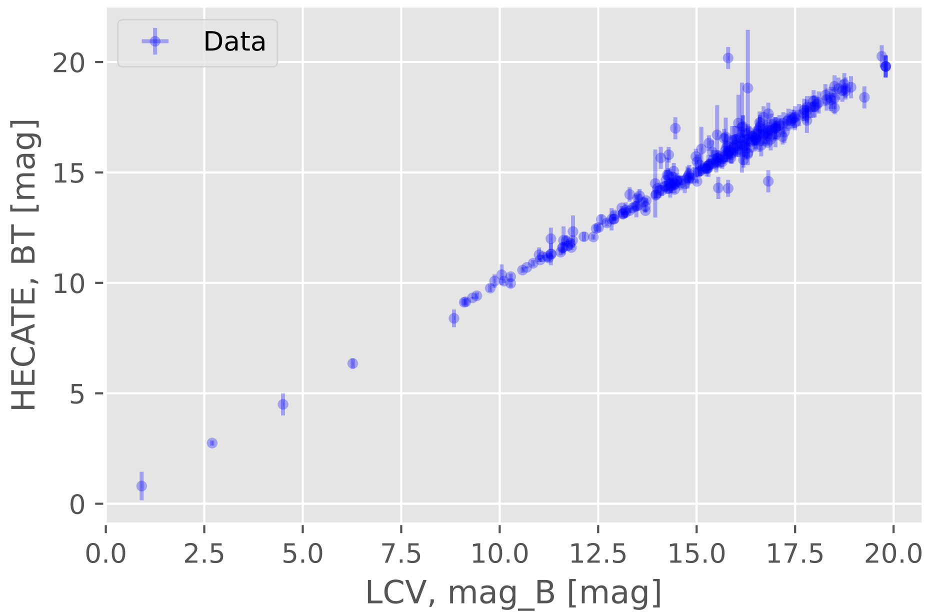


	log(L_K)_{LCV}$	log(L_K)_{HEC}	Percentage Change [%]
count	287	70	70
mean	8	9	-4
std	1	1	8
min	3	5	-48
50%	8	9	-0
max	11	11	7

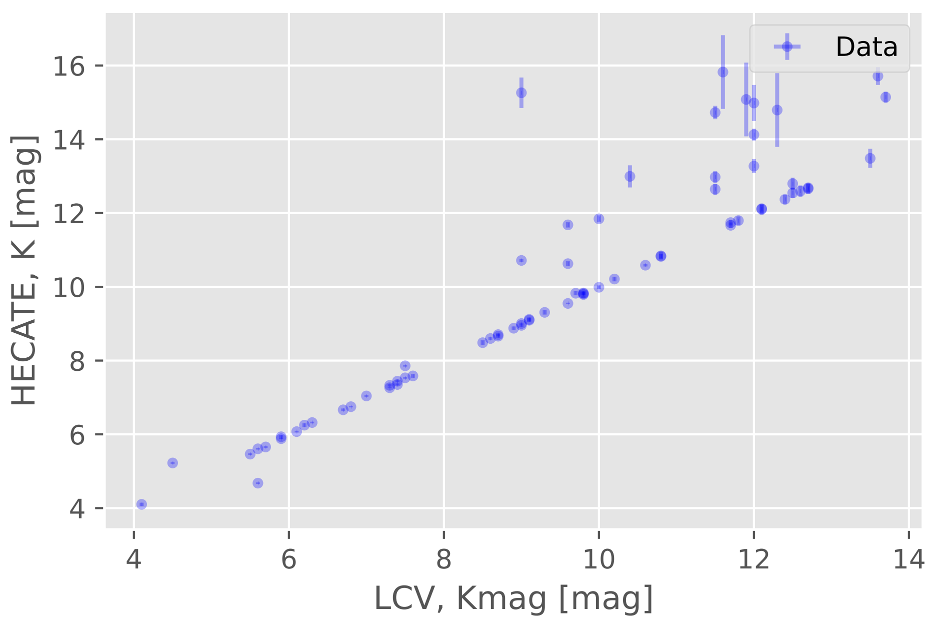
4.5 Magnitudes

LCV	HECATE	Description	Pearson Correlation [-1,1]
mag_B (with errors)	BT (with errors)		0
Kmag	K	2MASS band magnitude (both)	0

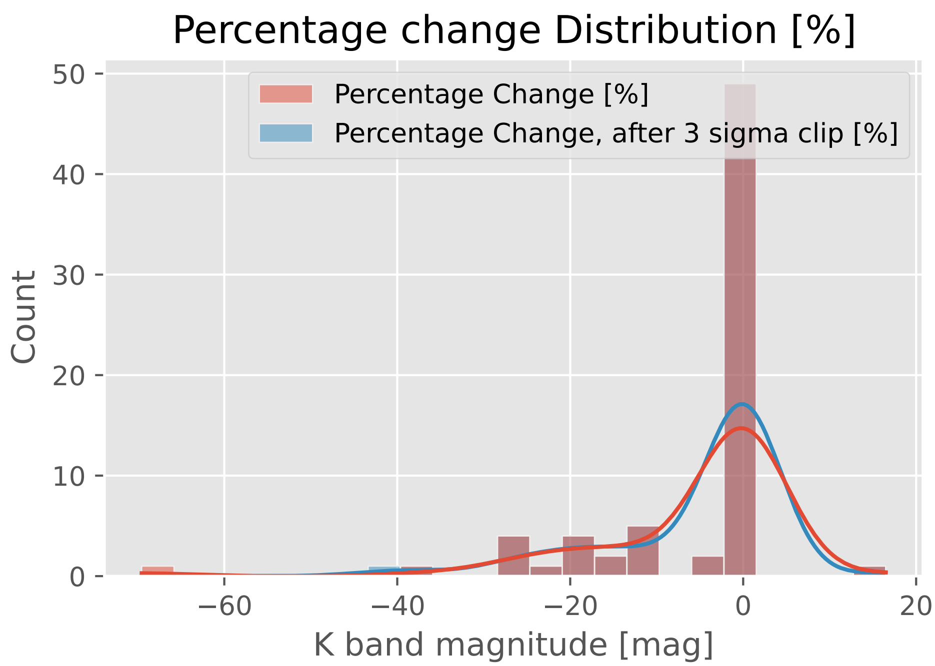
4.5.1 B mag



4.5.2 K mag



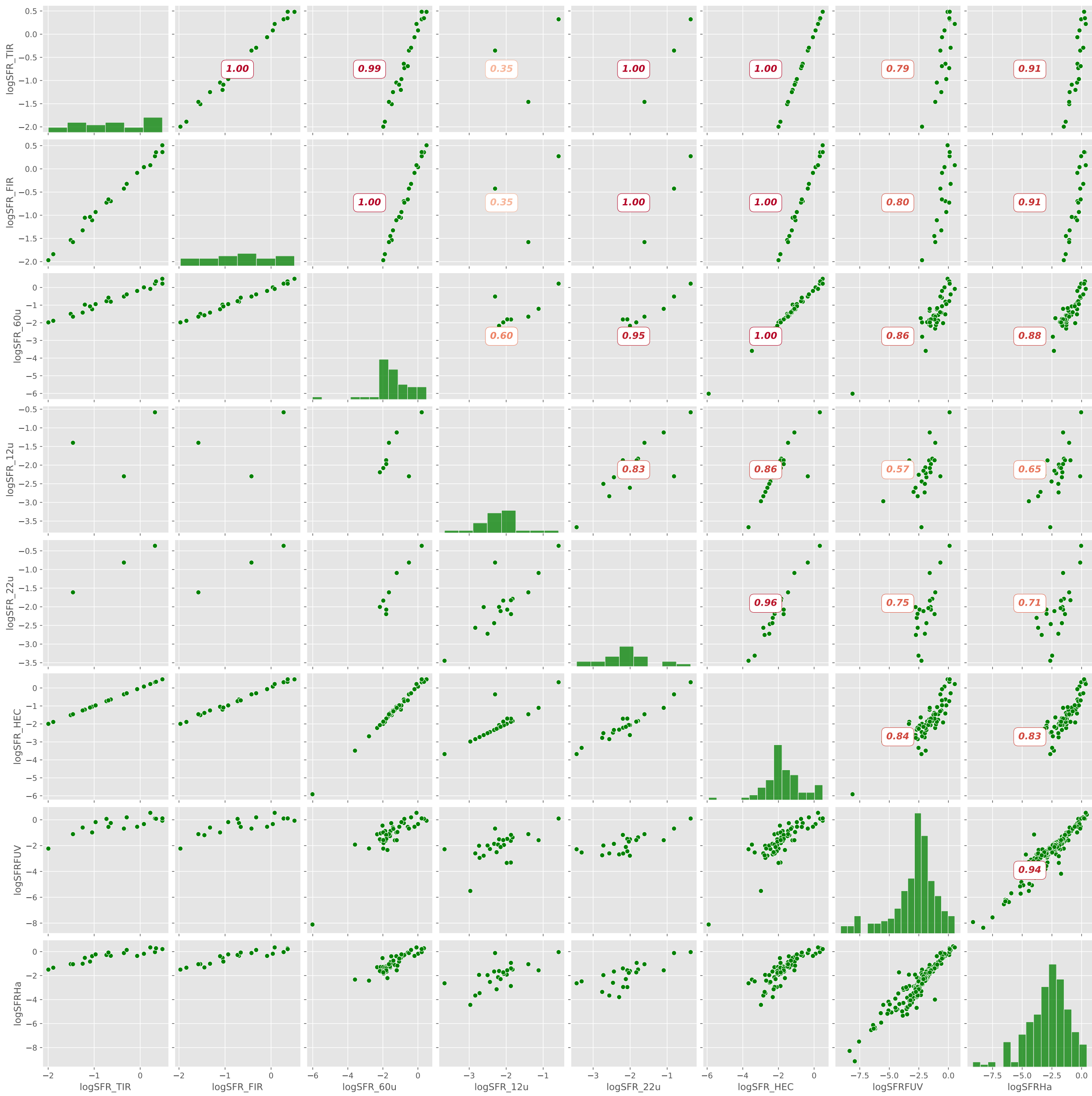
	Percentage Change [%]	Percentage Change, after 3 sigma clip [%]
count	70	70
mean	-5	-5
std	12	10
min	-70	-42
50%	-0	-0
max	16	16



[?]

4.6 SFR

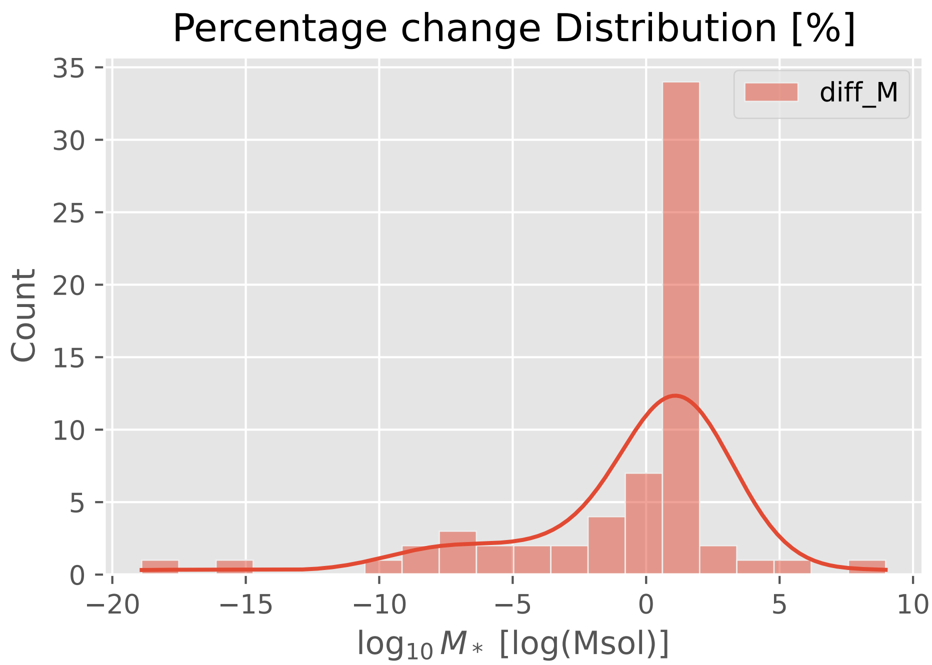
LCV	HECATE	Description	Count
	logSFR_TIR	Decimal logarithm of the total-infrared SFR estimate [Msol/yr]	21
	logSFR_FIR	Decimal logarithm of the far-infrared SFR estimate [Msol/yr]	22
	logSFR_60u	Decimal logarithm of the 60um SFR estimate [Msol/yr]	48
	logSFR_12u	Decimal logarithm of the 12um SFR estimate [Msol/yr]	26
	logSFR_22u	Decimal logarithm of the 22um SFR estimate [Msol/yr]	23
	logSFR_HEC	Decimal logarithm of the homogenised SFR estimate [Msol/yr]	73
	logSFR_GSW	Decimal logarithm of the SFR in GSWLC-2 [Msol/yr]	0
SFRFUV		FUV derived integral star formation rate	220
SFRHa		H{alpha} derived integral star formation rate	223



4.7 Masses

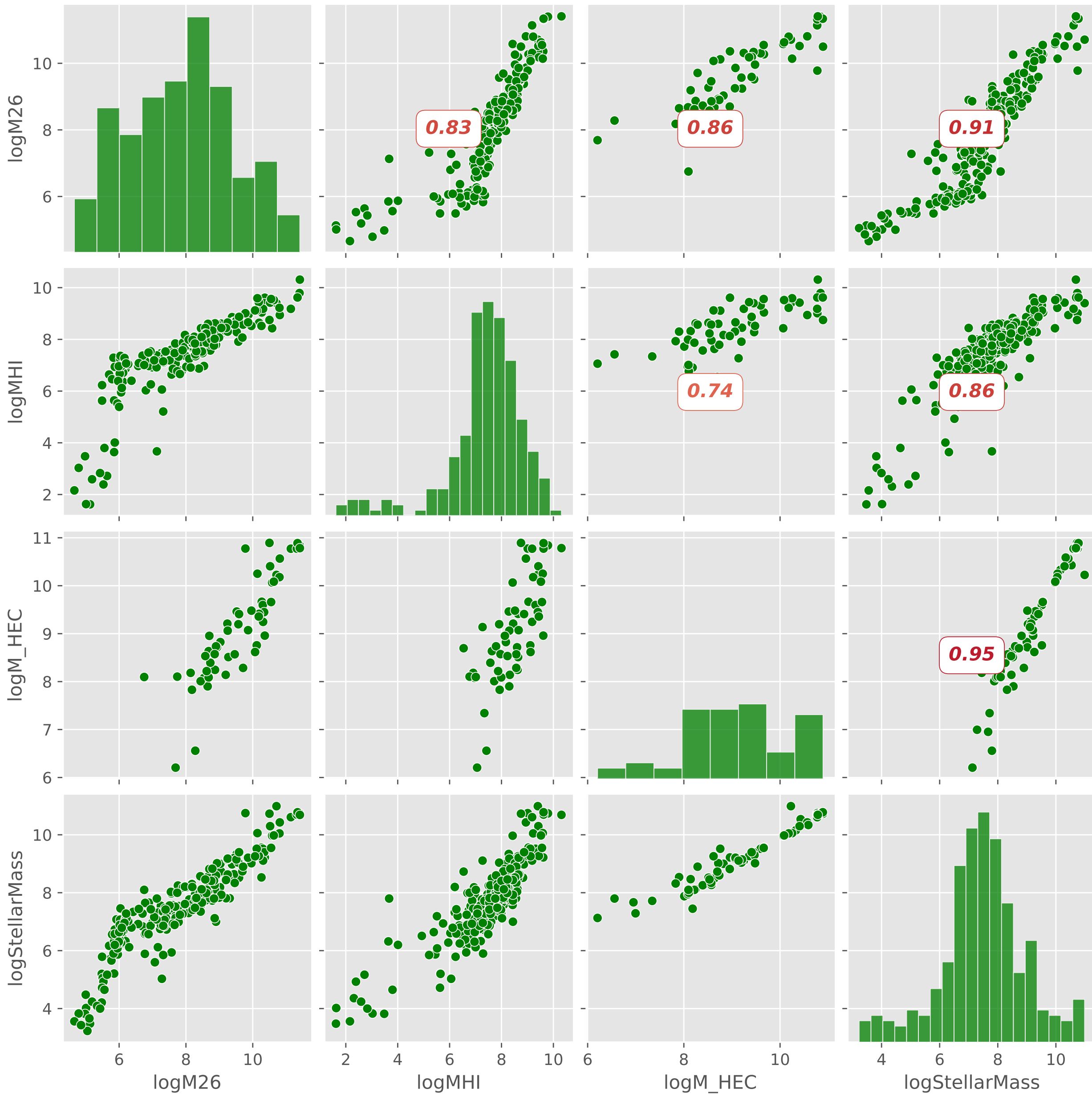
LCV	HECATE	Description	Count
logM26		Log mass within Holmberg radius	233
logMHI		Log mass within Holmberg radius	233
	logM_HEC	Decimal logarithm of the stellar mass [Msol]	64
	logM_GSW	Decimal logarithm of the stellar mass in GSWLC-2 [Msol]	0
logStellarMass		Stellar Mass from $M_*/L = 0.6$	287

4.7.1 Stellar Masses Comparison



	Percantage Change [%]
count	64
mean	-1
std	5
min	-19
50%	1
max	9

4.7.2 Heatmap



Karachentsev, Igor D., and Elena I. Kaisina. 2013. "STAR FORMATION PROPERTIES IN THE LOCAL VOLUME GALAXIES VIA H AND FAR-ULTRAVIOLET FLUXES." *AJ* 146 (3): 46. https://doi.org/10.1088/0004-6256/146/3/46.

Karachentsev, Igor D., Dmitry I. Makarov, and Elena I. Kaisina. 2013. "UPDATED NEARBY GALAXY CATALOG." *AJ* 145 (4): 101. https://doi.org/10.1088/0004-6256/145/4/101.

Kovlakas, K., A. Zezas, J. J. Andrews, A. Basu-Zych, T. Fragos, A. Hornschemeier, K. Kouroumpatzakis, B. Lehmer, and A. Ptak. 2021. "The Heraklion Extragalactic Catalogue (HECATE): A Value-Added Galaxy Catalogue for Multimessenger Astrophysics." *Monthly Notices of the Royal Astronomical Society* 506 (September): 1896–1915. https://doi.org/10.1093/mnras/stab1799.