COMPARISON OF CATALOGS

A PREPRINT

October 6, 2024

Table of contents

1 The data

In this script we will compare 2 catalogs, HECATE (Kovlakas et al. (2021)) and LCV (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, for $z \approx 0$. 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, 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.

2.1 Completeness of the LCV Catalog

The local volume selection has been made by taking into account galaxies with:

1. Radial velocities of
$$V_{LG} < 600 \text{ km} \cdot \text{m}^{-1} \tag{1}$$
2. Distances of
$$D < 11 \text{ Mpc} \tag{2}$$

A simultaneous fulfillment of both conditions (1) and (2) is not required.

2.1.1 Difficulties in calculating the Completeness

As explained in (Karachentsev, Makarov, and Kaisina (2013)), it is difficult to calculate the completeness of the catalog.

1. Completeness within a 10 Mpc radius is difficult to assess due to: - Variability in galaxy properties (luminosity, size, surface brightness, gas content) - Errors in distance measurements (Tully–Fisher method errors of ~20–25%), especially at the 10 Mpc boundary.
    - Accurate distances are mainly known within ~5 Mpc.
    - Non-Hubble motions (~300 km/s)[1] may make up half of the adopted velocity constraint (Equation 1)
    - Solution for our usage: only keep the galaxies inside a radius of $D = 11Mpc$
    - HI sources in surveys with low angular resolution: “The presence around our Galaxy of hundreds of high-velocity clouds with low line-of-sight velocities and small W50 widths also provokes the inclusion of false “nearby” dwarf galaxies in the LV”(Karachentsev, Makarov, and Kaisina (2013))
2. Astro-Spam and Survey Errors: - Automatic surveys produce false detections (e.g., stars misclassified as galaxies, high-velocity clouds mistaken for dwarfs).
3. Conditional Completeness Estimate: - Galaxies brighter than $M_B^c = -11^m$ or with linear diameters larger than $A_{26} = 1.0$ kpc show (40–60)% completeness. - “among the members of the Local Group (D < 1 Mpc), only half of the galaxies have absolute magnitudes brighter than 11m. Consequently, more than half of the ultra-faint dwarf companions around normal galaxies, like the Sombrero galaxy (D = 9.3 Mpc), still remain outside our field of view.”
4. Undetected Ultra-Faint Dwarfs: - Many faint dwarf galaxies remain undetected beyond the Local Group. Estimated population of undetected dwarfs could be as large as 10^3–10^4 within the LV.
5. Surface Brightness Distribution: - Surface brightness remains consistent across distances, except for ultra-faint dwarfs ($SB < 31\text{mag} \cdot \text{arcsec}^{-2}$). - Faintest dwarf galaxies are detectable only nearby due to their resolution into individual stars.
6. Luminosity-Size Relationship: - Observations align with cosmological models predicting $L \sim A^3$ (Navarro, Frenk, and White (1996)), though deviations occur for extremely low surface brightness galaxies. - “The deviation from it at the extremely low surface brightness end is due to a systematic overestimation of dwarf galaxy sizes, the brightness profiles of which lie entirely below the Holmberg isophote.”

2.2 Completeness of HECATE

The completeness of HECATE is difficult to assess due to: - Unknown selection function of HyperLEDA and selection effects from other cross-correlated catalogs. - Estimation based on comparing B-band luminosity distribution with the galaxy luminosity function (LF).

[1] *Non-Hubble motion refers to the component of a galaxy's velocity that deviates from the uniform expansion of the universe, described by the equation* $V = H_0 d + v_p$,where V is the total velocity, $H_0 d$ represents the Hubble flow, and v_p is the peculiar velocity. “Essay” (n.d.)

- HECATE is:
  - Complete down to $L_B \sim 10^{9.5} L_{B,\odot}$ for distances less than 33 Mpc.
  - Complete down to $L_B \sim 10^{10} L_{B,\odot}$ for distances between 67 Mpc and 100 Mpc.
  - Incomplete at distances greater than 167 Mpc, even for the brightest galaxies.
- Completeness estimates based on B-band luminosity density:
  - >75% complete for distances less than 100 Mpc. $\sim 50\%$ complete at distances of $\sim$ 170 Mpc.
  - Completeness exceeds 100% within 30 Mpc due to the overdensity of galaxies around the Milky Way.
- Completeness in terms of stellar mass (M*):
  - Similar to B-band completeness when measured with K_s-band luminosity as a tracer for stellar mass.
  - Overdensity at small distances and cut-off at large distances are observed.
- Completeness in terms of star formation rate (SFR):
  - ~50% complete between 30 and 150 Mpc.
  - Lower SFR completeness due to limitations in WISE-based SFR estimates, which lack full sky coverage.
  - Despite IRAS's limited depth, it provides >50% coverage for star-forming galaxies in the local neighborhood.
  - HECATE's nonuniform SFR and stellar mass coverage, affecting the reliability of stellar population parameters.

In this section we will check the completeness

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

2.3 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 %

2.4 Completeness in Outer join

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

Completeness (HECATE)= 90 %

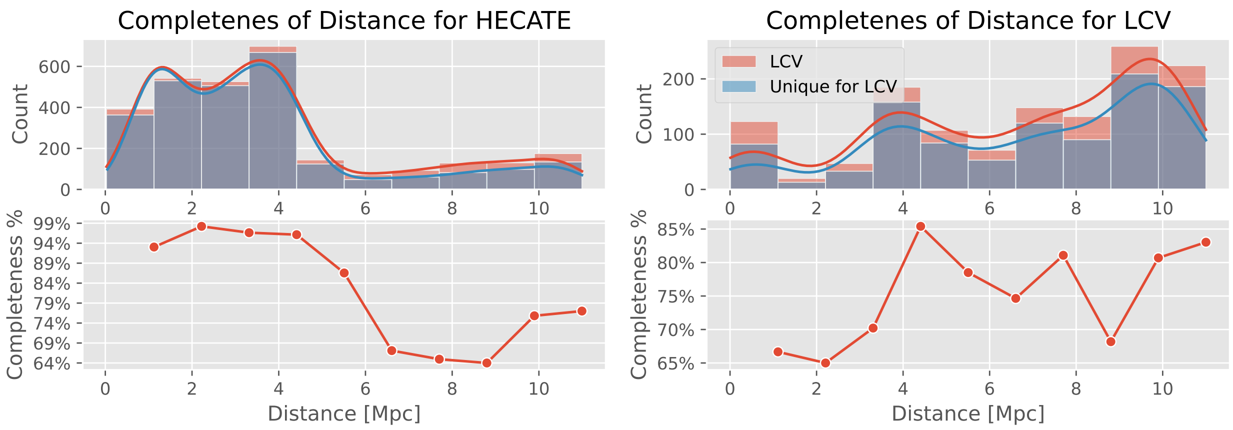
Completeness (LCV)= 78 %

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

2.5 Completeness of the Data

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

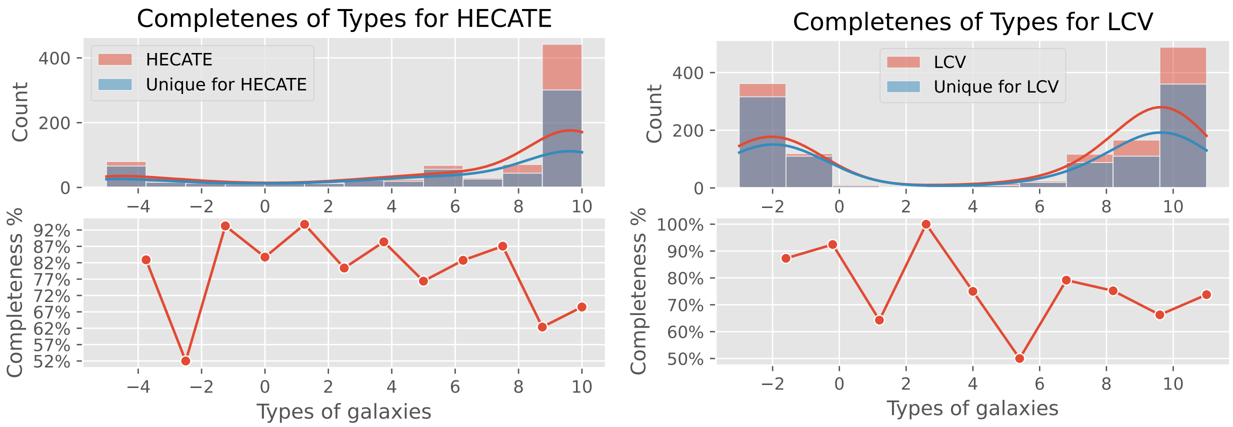
This means there is no bias in the selection of the galaxies.



(a) HECATE

(b) LCV

Figure 1: Histograms showing the Distance Completeness of the Catalogs



(a) HECATE

(b) LCV

Figure 2: Histograms showing the Type Completeness of the Catalogs

3 How are we going to compare the data?

For the comparison of our data we will mainly use two methods: linear regression and comparison of the distributions.

For the linear regression out main tools are scatter plots and R^2

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.
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. Beside scatter plots and histograms, another useful plot is the Correlation Heatmap.

> 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.

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

For the distrabution comparisons we will use histograms and Kernel Density Estimate (KDE) plots. The KDE plots visually represent the distribution of data, providing insights into its shape, central tendency, and spread.

Finally, we will examine the percentage change for each galaxy.

- 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

LCV	HECATE	Description	Pearson Correlation [-1,1]
Dis	D	Distance	0.881

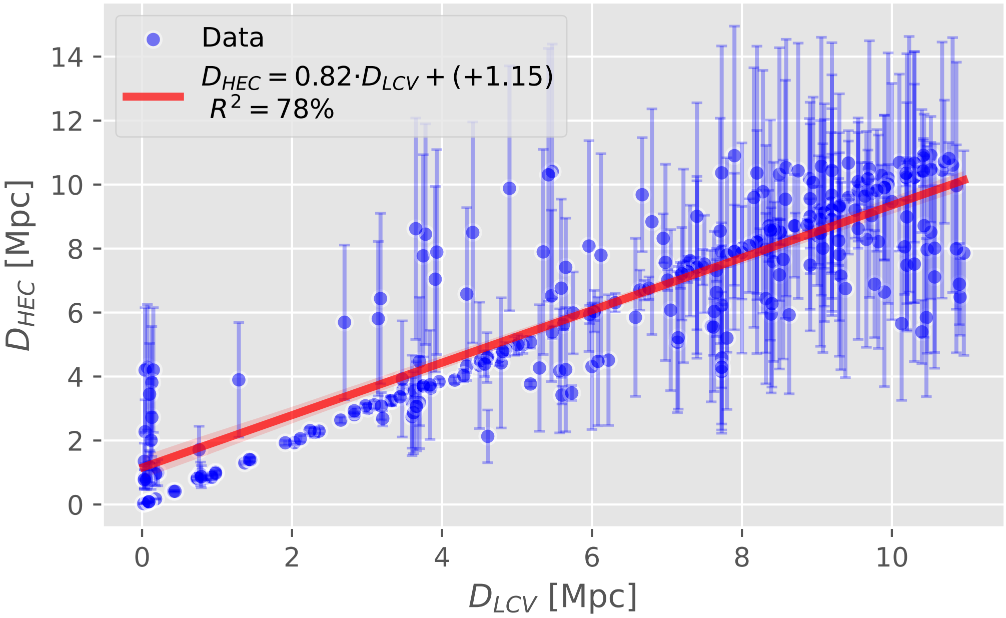


Figure 3: Comparison of the Distances

- The average error of the distance in the HECATE catalog is $\overline{\delta D} = \pm 1.6$ Mpc, so the intercept is included in the error.
- So we can assume that the Distances are the same

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	

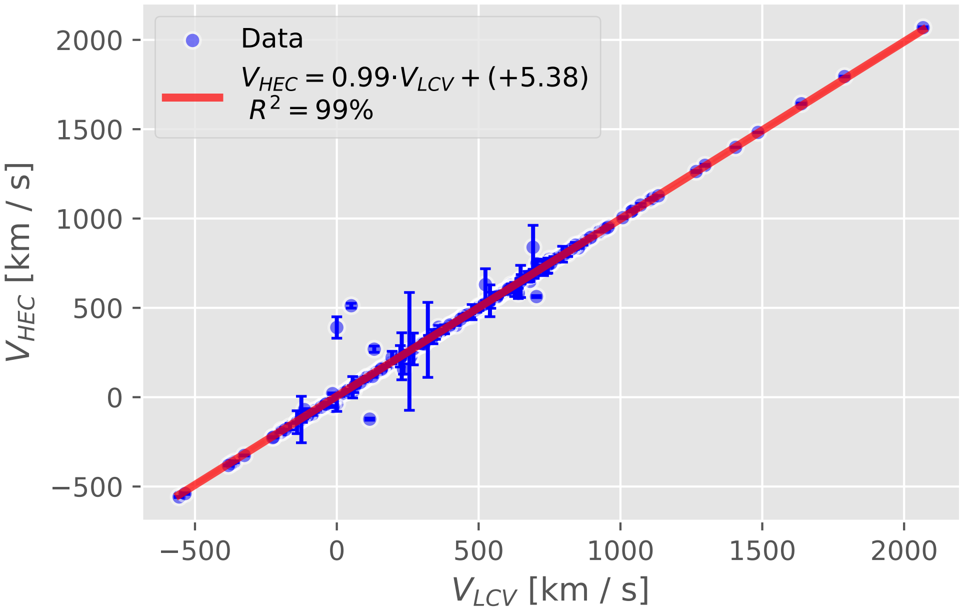


Figure 4: Comparison of the Radial Velocities

- The average error of the radial velocity in the HECATE catalog is $\overline{\delta V} = \pm 12 \text{km} \cdot s^{-1}$, so the intercept is included in the error.
- So we can assume that the radial velocities are the same

[?] 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.

The strong correlation between the velocities reported by the HECATE and LCV catalogs can be attributed to the fact that both measure the same intrinsic galaxy velocities but reference them to different frames. For instance, velocities can be corrected for the movement of the Earth (heliocentric frame) or the for the Virgo-infall. These corrections shift the measured velocity slightly, but the fundamental measurement remains consistent. This explains why the velocities are highly correlated, as they represent the same physical phenomenon but adjusted for different reference points.

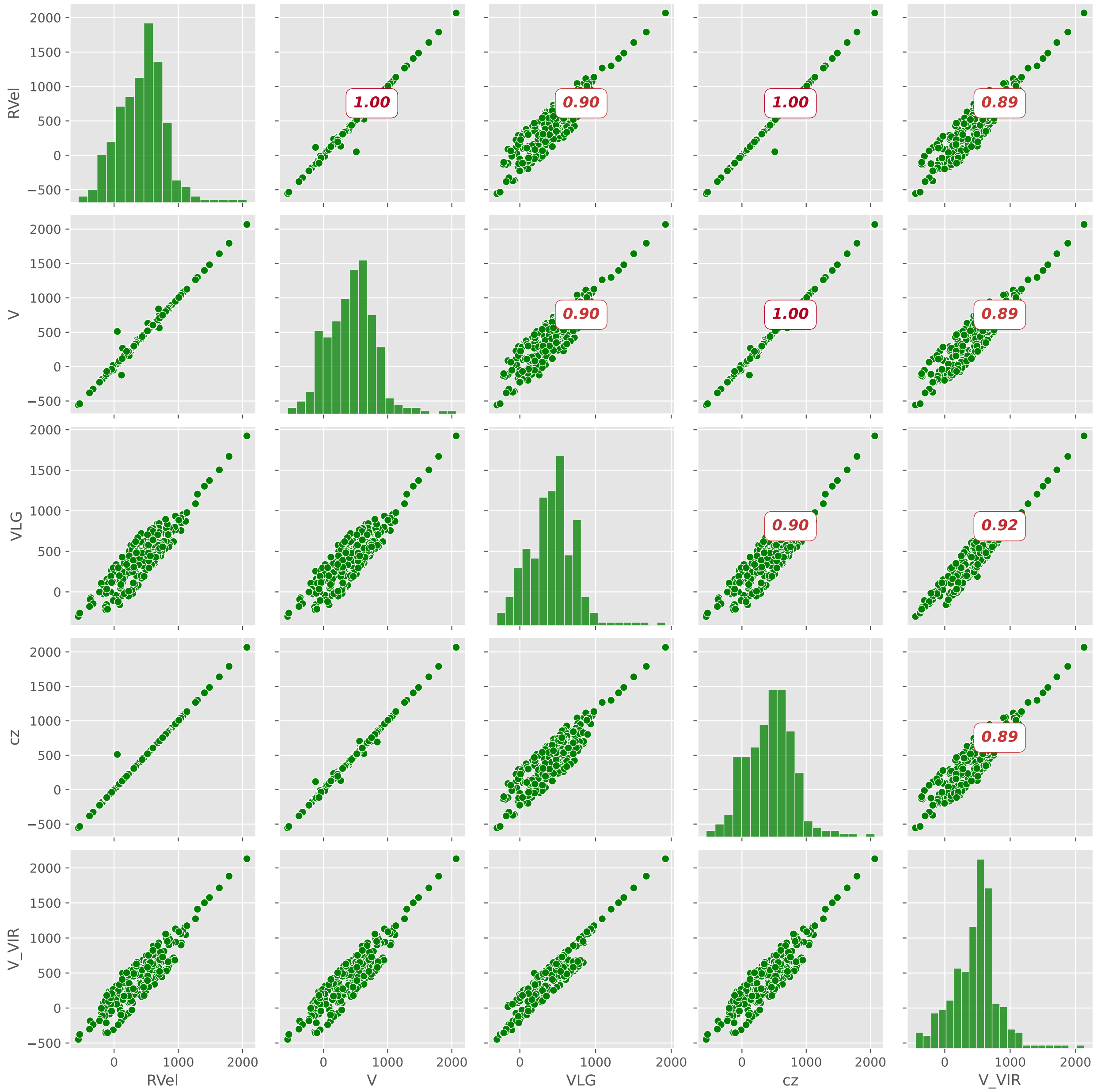


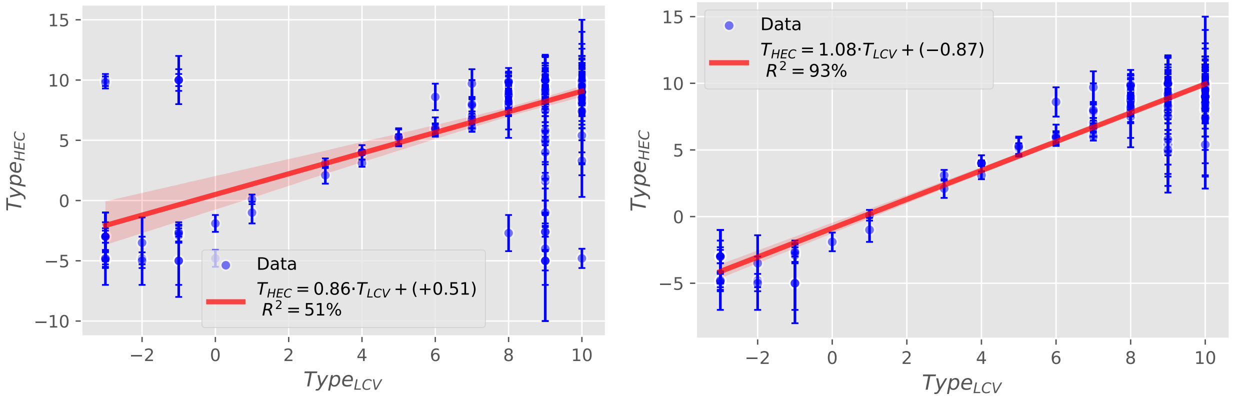
Figure 5: The correlation Matrix of the Velocities. The lower left triangle is is composed of the scatter plots of the various velocities, the diagonal shows their distrubution and the upper right triangle shows the correlations of the Velocities

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

"Morphological type of galaxy in the numerical code according to the classification by de Vaucouleurs et al. (1991). It should be noted that about three quarters of objects in the LV are dwarf galaxies, which require a more detailed morphological classification. For example, dwarf spheroidal galaxies and normal ellipticals are usually denoted by the same numerical code T < 0, although their physical properties drastically differ. The classification problem arises as well for the "transient" type dwarf galaxies, T r, which combine the features of spheroidal (Sph) and irregular (Ir)systems. Due to small classification errors, such objects may "jump" from one end of the T scale to the other." (Karachentsev, Makarov, and Kaisina (2013))



(a) All the galaxies, without the correction

(b) Comparison with the correction $|T_{HECATE} - T_{LCV}| < 7$

Figure 6: Comparison of the Types of galaxies

- we can assume that the galaxies from the upper left and lower right regions of the plot are classified differently because of this
- this might explain the large errors E_T of HECATE
- why we see a difference in the distribution of (Figure 2)

We can remove the "problematic" galaxies by only keeping the ones with:

$$|T_{HECATE} - T_{LCV}| < \sigma_{T_{LCV}} = 4.8$$

- The average uncertainty of the morphological type in the HECATE catalog is $\overline{\delta T} = \pm 1.4$, so the intercept is included in the error.
- So we can assume that the Distances are the same

4.3.2 Inclination

	inc	INCL	Percentage Change [%]
count	287	209	202
mean	60	59	-0
std	19	23	0
min	9	0	-3

	inc	INCL	Percentage Change [%]
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 (Figure 7). 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 (Figure 8a). We need to see the linear correlation of the decimal logarithms (Figure 8b).

$\overline{R_1} = 3.9$ [arcmin], $\overline{a_{26}} = 9.3$ [arcmin], so the intercept is negligable.

$$R_1 = 0.48 \cdot a_{26} - 0.34 \sim \frac{1}{2} a_{26}$$

$$\log(R1) = 0.89 \log(a_{26}) - 10^{0.38} = \log(10^{-0.38} a_{26}^{0.89}) = \log(0.41 \cdot a_{26}^{0.89}) \Rightarrow R1 \simeq \frac{a_{26}}{2}$$

4.4 Luminosities

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

$$\log(L_{K,HEC}) = 1.16 \log(L_{K,LCV}) - 1.78 = \log\left(\frac{{L_{K,LCV}}^{1.16}}{10^{1.78}}\right) \Leftrightarrow L_{K,HEC} = 0.02 \cdot {L_{K,LCV}}^{1.16}$$

To assess whether the two galaxy catalogs have a 1-1 correlation, I applied two types of linear fits to the data: a free fit and a forced fit. The free fit allows both the slope and intercept to vary freely, meaning the line can take on any slope and can cross the y-axis at any point. In contrast, for the forced fit, I set the slope to exactly 1 and fixed the line to pass through the origin, meaning the equation becomes y=xy=x. This forced fit represents a perfect 1-1 relationship, where the values from one catalog should match the other exactly, without any shift or scaling.

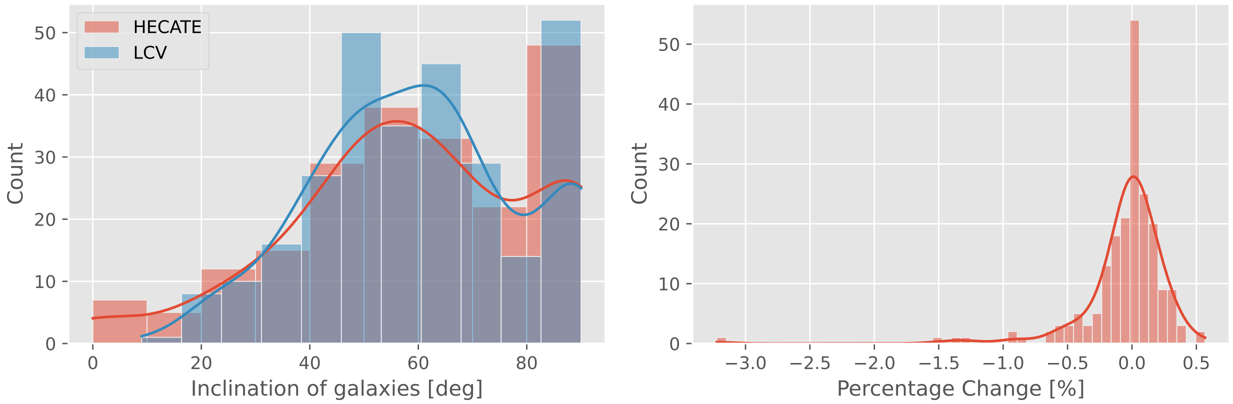
After fitting the data using both methods, I calculated the R^2 value for each fit. In this case, both the free fit and the forced fit produced very similar R^2 values, around 87.13%[3], indicating that both fits explain about 87% of the variance in the data (Figure 9).

The fact that both the free and forced fits have nearly identical R^2 values suggests that the data closely follows a 1-1 relationship. In the free fit, the best-fitting line had a very small intercept, which shows that allowing the line to shift vertically did not significantly improve how well the line fits the data. This strongly indicates that the data from the two catalogs are well-aligned and follow the relationship y=x quite closely, which is what we expect if the catalogs are in good agreement.

While these results are promising, it's important to note that achieving an R^2 of 87%—while high—doesn't necessarily prove a perfect 1-1 correlation. There may still be small discrepancies or uncertainties in the data that prevent a perfect match. Nonetheless, the fact that fixing the slope to 1 and forcing the line through the origin gave such a good result suggests that the two catalogs are very consistent with each other.

4.5 Magnitudes

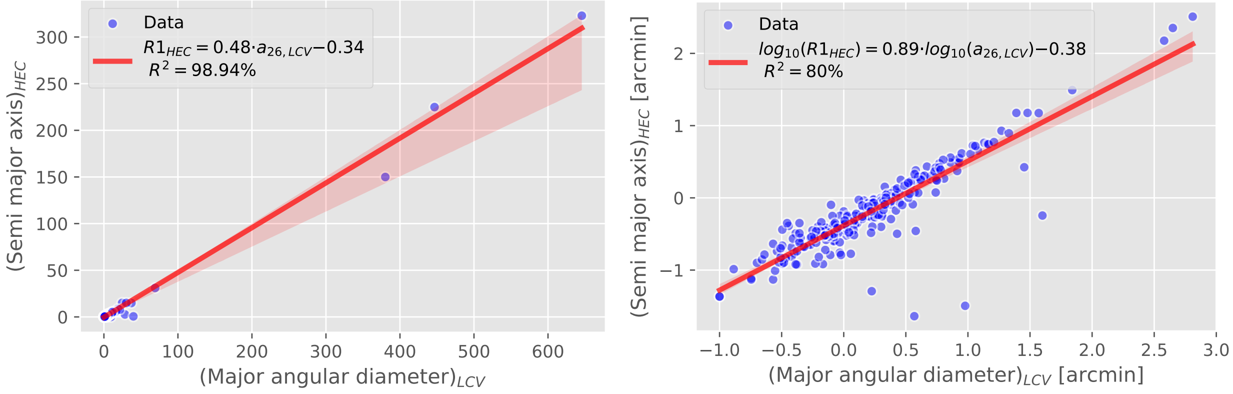
[3] The difference of the linear regression with the free parameters and the forced regression is $\Delta R^2 = 1.14 \cdot 10^{-13}\%$



(a) Distribution of the Inclination of the galaxies

(b) Distribution of the Percentage Change

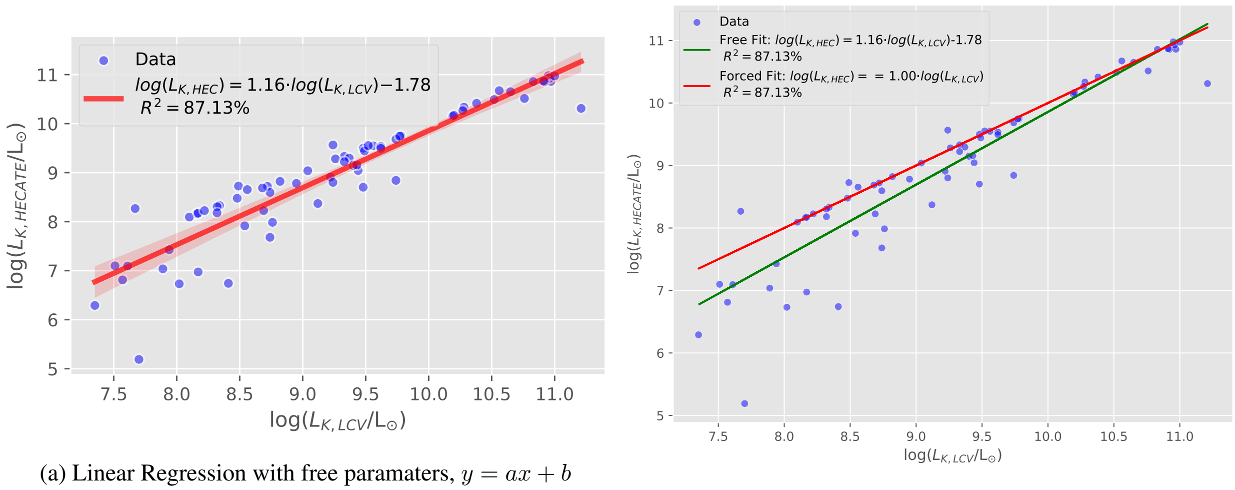
Figure 7: Comparison of the Inclination of the galaxies



(a) Linear scale

(b) log_{10} scale

Figure 8: Comparison of the Major Axises of the galaxies

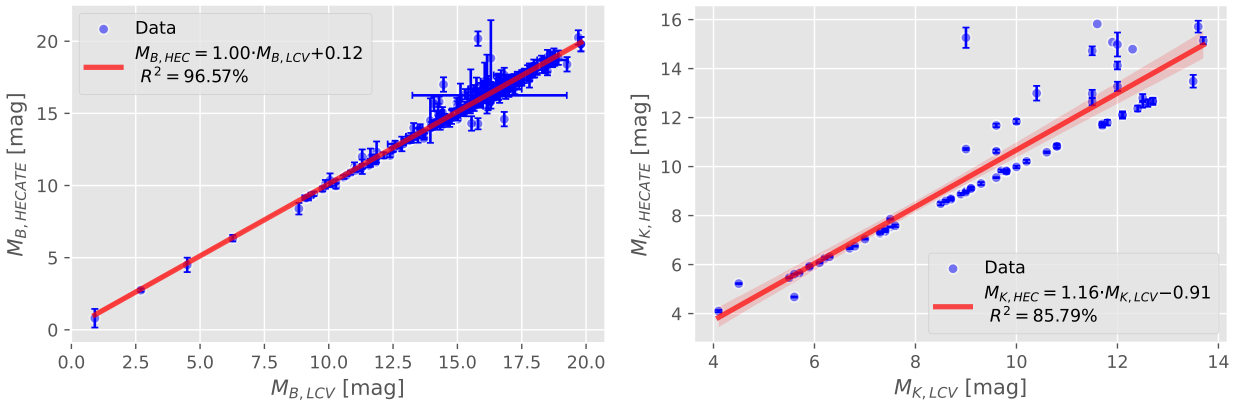


(a) Linear Regression with free paramaters, $y = ax + b$

(b) Linear Regrassion $y = x$

Figure 9: Comparison of the L_K of the galaxies.

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

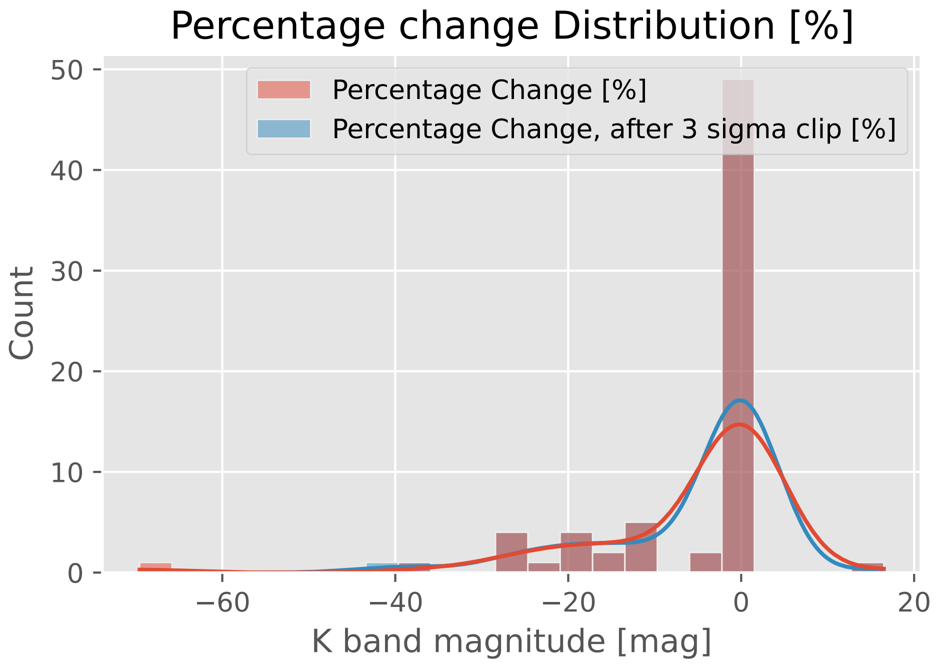


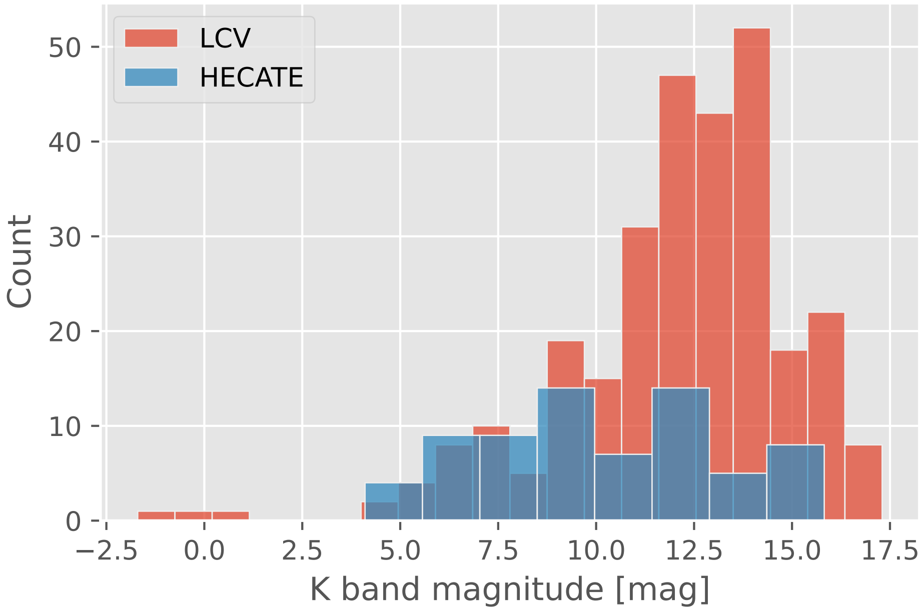
(a) M_B

(b) M_K

Figure 10: Comparison of the Magnitudes of the galaxies

- M_B: it is a 1-1 correlation, since the average error $M_{B,HECATE} = 0.4\ mag$, so the itercept is smaller than the error (Figure 10)
- M_K: we need to examine it more, since the intercept is bigger than the error $M_{K,HECATE} = 0.09\ mag$ (Figure 10)





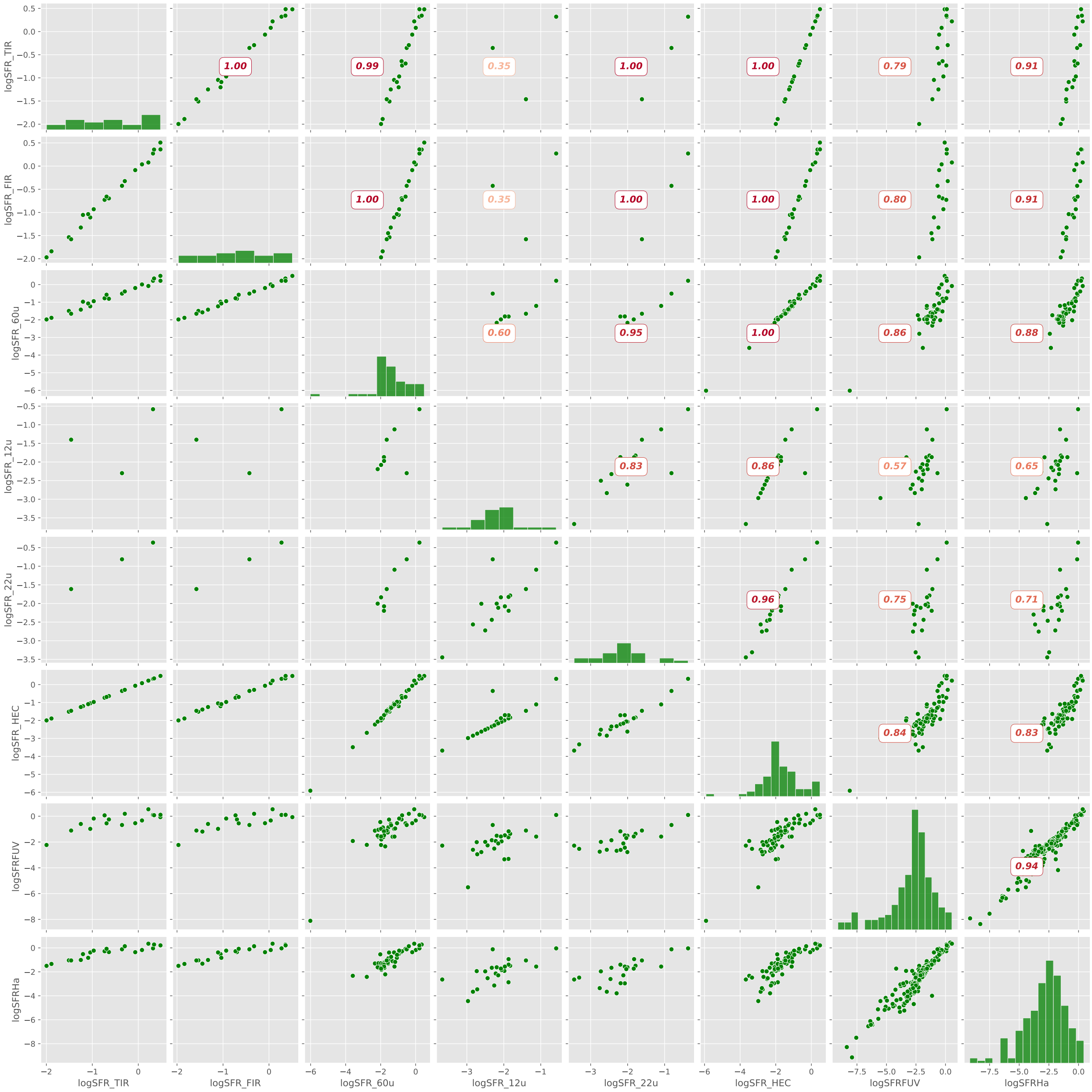
	Percentage Change [%]	Percentage Change, after 3 sigma clip [%]
count	70	70
mean	-5	-5
std	12	10
min	-70	-42
50%	-0	-0
max	16	16

[?] this 5% mean in the difference probably exists because of the low number of galaxies we compare, so we can ignore it.

4.6 SFR

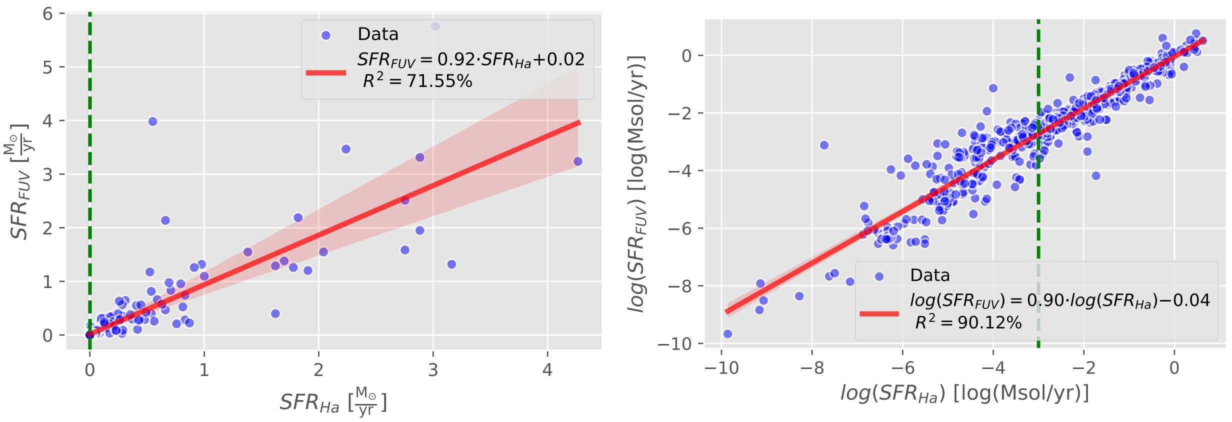
LCV	HECATE	Description	Count
	logSFR_TIR	Decimal logarithm of the total-infrared SFR estimate [Msol/yr]	21
	logSFR_FIR	Decimal logarithm of the far-infrared SFR estimate [Msol/yr]	22
	logSFR_60u	Decimal logarithm of the 60um SFR estimate [Msol/yr]	48
	logSFR_12u	Decimal logarithm of the 12um SFR estimate [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

LCV	HECATE	Description	Count
	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



The SFR according to (Kroupa et al. (2020)), can be calculated from the mean of SFR from the Ha and FUV, for $SFR > 10^{-3}\ M_{\odot}\, yr^{-1}$. As we can see from the plots (Figure 11) it is a good aproximation for all SFR's

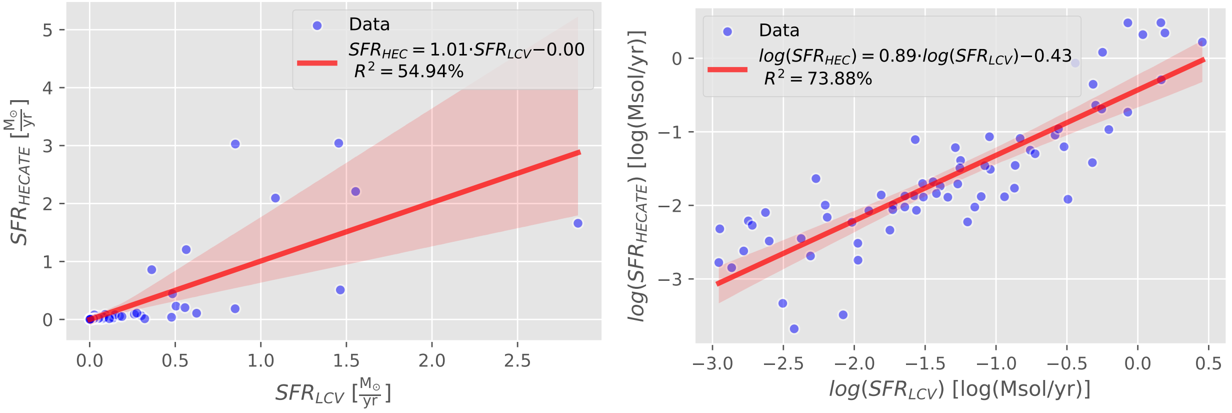
$$SFR = \frac{SFR_{FUV} + SFR_{H\alpha}}{2}, \text{if both exist, else: } SFR = SFR_i,\ i = FUV,\ H\alpha$$



(a) Linear scale

(b) Decimal logarithmic scale

Figure 11: Comparison of the $SFR_{FUV} - SFR_{Ha}$ of the galaxies



(a) Linear scale

(b) Decimal logarithmic scale

Figure 12: Comparison of the SFR's of the galaxies

[?]EXPLANATION this low correlation maybe comes from the inaccuracy/approximation of (Kroupa et al. (2020))

The reason is that this aproximation is accurate if we consider a stable SFR over time. However, we tend to observe higher Ha for more recent SF activity. (we discussed it in our last meeting, I need a refrence).

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

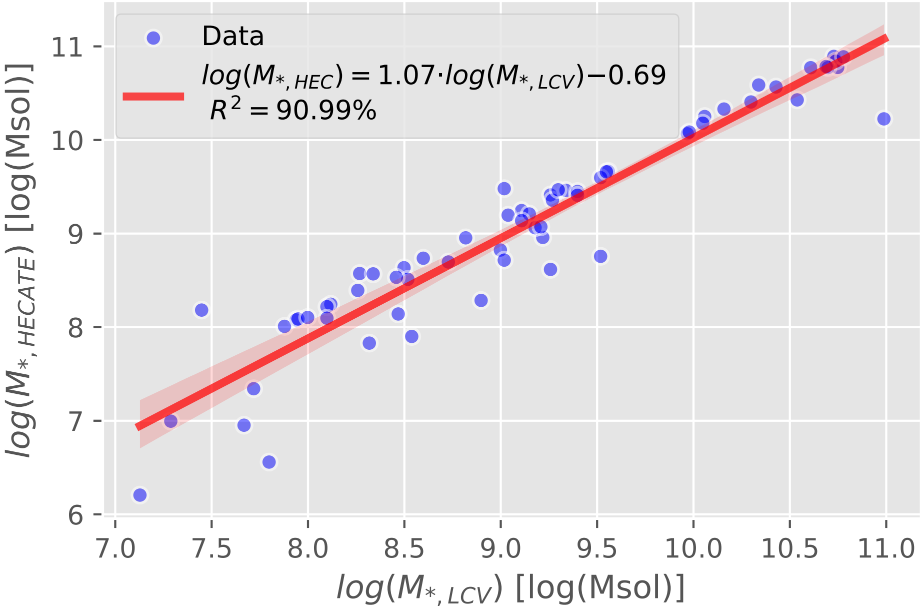
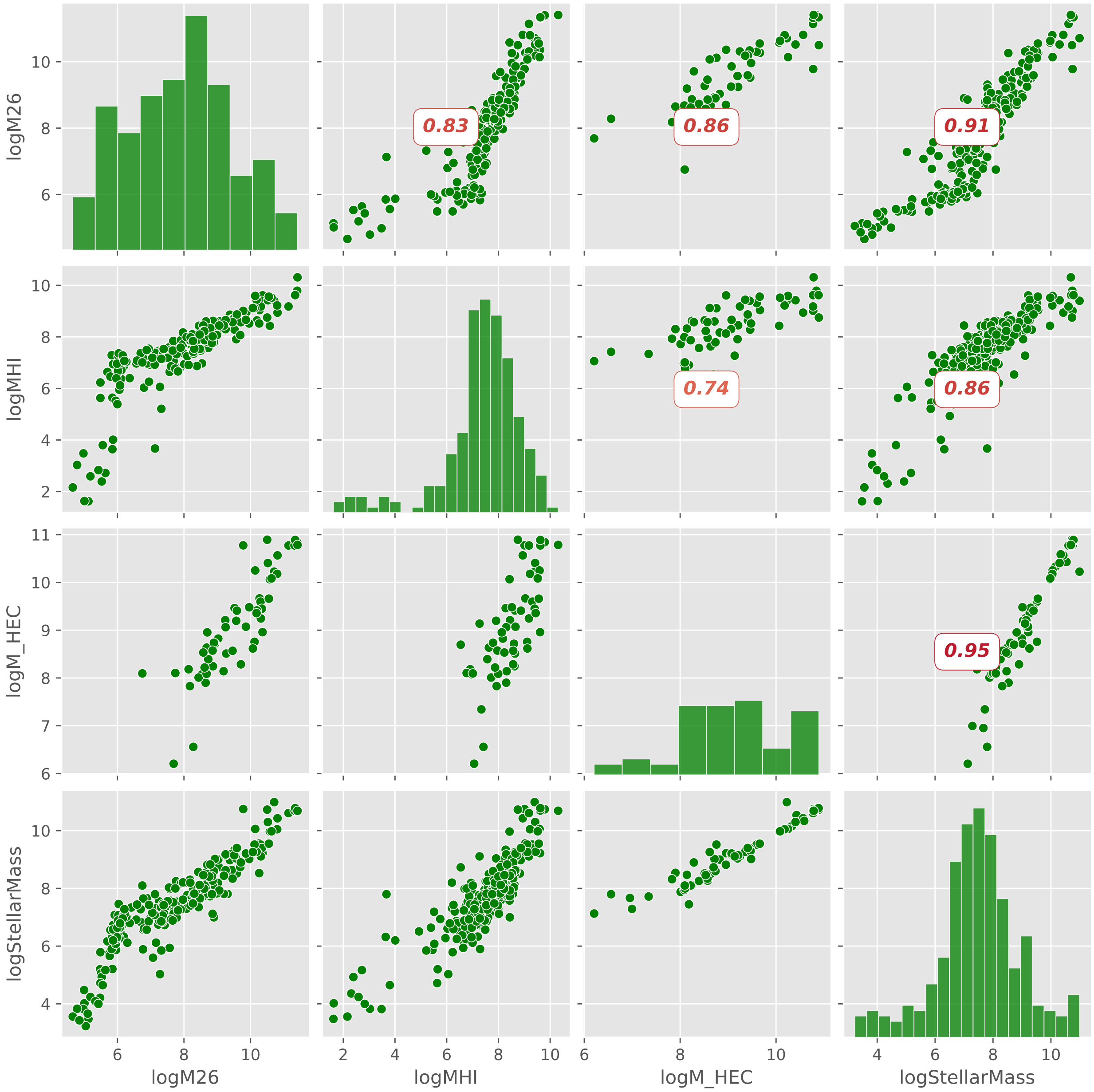


Figure 13: Comparison of the Stellar Masses of the galaxies

$$\log M_{*,HECATE} = 1.07 \cdot \log M_{*,LCV} - 0.69, \text{ but } \log M_{*} > 6 \Rightarrow M_{*,HEC} \sim M_{*,LCV}$$

As we can see the aproximation of Mass/Light=const.=0.6 is a pretty good approximation, for the calculation of $\log(M_{*}/M_{\odot})$, especially for high-mass galaxies

4.7.2 Heatmap



The heatmap is used to visually assess the correlations between different mass estimates of the galaxies, including the mass within the Holmberg radius (M_{26}), the HI gas mass (M_{HI}), the stellar mass from the HECATE catalog (M_{HEC}), and the stellar mass derived assuming a mass-to-light ratio of 0.6 (StellarMass). The purpose of this visualization is to quickly identify whether these different mass measurements show consistent correlations, which would suggest that they capture similar aspects of the galaxies' mass distributions.

This quick check using the heatmap helps confirm that, while different catalogs and methods might have their own specific ways of calculating galaxy masses, the overall mass estimates align well, offering a solid basis for further analysis of galaxy properties.

"Essay." n.d. Accessed October 5, 2024. https://ned.ipac.caltech.edu/level5/Glossary/Essay_pecmotion.html.

Karachentsev, Igor D., and Elena I. Kaisina. 2013. “STAR FORMATION PROPERTIES IN THE LOCAL VOLUME GALAXIES VIA H AND FAR-ULTRAVIOLET FLUXES.” *The Astronomical Journal* 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.” *The Astronomical Journal* 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.

Kroupa, P, M Haslbauer, I Banik, S T Nagesh, and J Pflamm-Altenburg. 2020. “Constraints on the Star Formation Histories of Galaxies in the Local Cosmological Volume.” *Monthly Notices of the Royal Astronomical Society* 497 (1): 37–43. https://doi.org/10.1093/mnras/staa1851.

Navarro, Julio F., Carlos S. Frenk, and Simon D. M. White. 1996. “The Structure of Cold Dark Matter Halos.” *The Astrophysical Journal* 462 (May): 563. https://doi.org/10.1086/177173.