A Glimpse Of Euclid's Universe Through COSMOS

Research Internship - LASTRO

CARRON FABIO

UNDER THE SUPERVISION OF :
AURÉLIEN VERDIER

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Physics

JULY - SEPTEMBER 2025

Abstract

Our universe is yet full of secrets, hidden within distant galaxies, waiting for us to unravel them. One might say, all the answers lie right before our eyes; we just have to seize them. But how can we know where to look?

To tackle this, astrophysicists rely on vast surveys that map millions of galaxies across cosmic times. At the heart of this mission is the *Euclid* satellite, launched in 2023 by the European Space Agency. However, with its first public data release not expected until October 2026, we sought a way to perform early analyses ahead of time.

In this work, we explore the possibility of simulating *Euclid*'s observations using existing data from a rich collection of space- and ground-based measurements: the COSMOS survey. We propose a method to ensure a clean, high-quality and reliable catalogue, determining which objects would realistically be detected by *Euclid*. Using this simulated dataset, we identify a specific type of galaxies known as Emission Line Galaxies (ELGs), that appear brighter in some colours than in others. To detect them, we use "colour-colour diagrams", which basically highlight these brightness patterns. Because they are easy to find and spread across the sky, ELGs are excellent tracers of the cosmic web, hence their importance.

This study is crucial because it provides a ready-to-use dataset to support preliminary analyses ahead of Euclid's first public data release. We produced a robust Euclid-like catalogue and a new method to ensure reliable results. Moreover, once *Euclid* data become available, comparisons between results will allow to verify the consistency of our method and lay the foundation for new exciting scientific opportunities. We also provide a final sample of approximately 15'000 objects assumed to be ELGs, a valuable resource ready to fuel future discoveries.

.

[0] **N.B**: see the Glossary at the end of this article for descriptions of technical terms. The first time these terms are used, they will mostly be explained, but you can always refer to the glossary for the other occurrences. It is written in the alphabetical order.

1 Introduction

Emission Line Galaxies (ELGs) are key objects in modern cosmology. Thanks to their properties, they allow us to map the large-scale structure of the Universe and place important constraints on the nature of dark energy. They act as reliable tracers of the cosmic web, making it possible to study how matter is distributed on large scales and how the Universe has expanded over time.

Over the past decade, several major spectroscopic surveys have focused on ELGs, such as the BOSS program (within SDSS-III) [4] and, more recently, the DESI project [11]. In the coming years, new surveys will greatly expand our ability to study these galaxies, and among them, the *Euclid Space Mission* stands out as a milestone.

Launched in 2023, *Euclid* is a space-based observatory that aims to study the nature of dark matter and dark energy with unprecedented precision, improving our overall understanding of the universe's structure and expansion [8]. It carries two main instruments: a visible-light camera and a near-infrared spectrophotometer. Together, they will map a large fraction of the extragalactic sky over six years. The first public cosmological data release (Data Release 1) is scheduled for October 2026 and will already cover a significant part of the final survey.

But why wait? In this work, we build a preview of what *Euclid* will see. Using the COSMOS survey, one of the most detailed maps of the sky, combining data from dozens of telescopes, we simulate what *Euclid*'s observations will look like. By adjusting COSMOS data to match *Euclid*'s expected sensitivity, we create a realistic "mock" catalogue of galaxies. Specifically, we will use data coming from the UltraVISTA telescope in the Y-, J- and H-bands - think of a band as a set of wavelengths, or colours, going from ultraviolet through visible light, all the way to infrared. These bands (Y, J and H) are in the near-infrared, and are the closest to *Euclid*'s.

With this simulated dataset, we focus on identifying Emission Line Galaxies under the same conditions that *Euclid* will face. This not only provides us with a ready-to-use catalogue for early analyses, but also allows us to test methods and prepare for the wealth of discoveries that Euclid will soon deliver.

2 Data filtering

Prior to conducting any analysis, it is essential to ensure that the dataset is both relevant to our objectives and of sufficient quality for reliable use. To this end, we apply a series of selection criteria, or masks, which serve two main purposes: (i) to restrict the sample to sources relevant for our study, and (ii) to exclude objects with insufficient observational quality. These quality and selection cuts systematically reduce the initial dataset to a refined sample suitable for subsequent analysis.

Using the COSMOS catalogue, we can already impose cleanliness conditions. This way, we first make sure that the objects we keep are galaxies, and not bright stars or active galactic nuclei. Then, we also flag sources that are contaminated by close stars or not bright enough. These

initial quality and classification cuts provide a robust baseline for the dataset. With this cleaned sample, we can now proceed to apply more specific criteria.

Next, we will ensure that the sources are observed with a sufficient precision by imposing conditions on the flux and magnitude errors (the magnitude basically giving an idea of the brightness of an object, the higher it is, the fainter the object). In astronomy, we define the Signal-to-Noise Ratio (SNR) as:

$$SNR = \frac{\text{flux}}{\text{flux error}} \tag{1}$$

We typically assume that any object with a SNR above 5 is reliable. We can use this to sort our sources based on their flux errors. Figure 1 shows the flux error with respect to the flux, in the Y-band (the results are similar in the other bands). We perform two cuts : (i) a horizontal cut, keeping objects with an error below 0.2 μJy and (ii) a cut following the SNR above 5 condition. By keeping only objects whose flux error is below the maximum of these two curves, we get a solid condition to remove any data that is not precise enough in terms of flux.

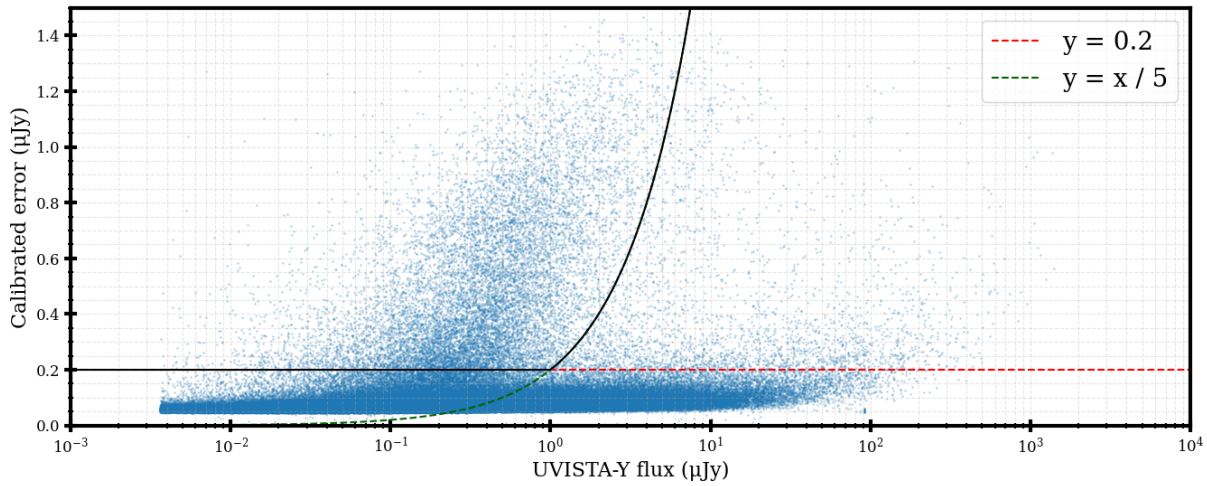


Figure 1: Cuts on the calibrated error of the flux, in the Y band. In red, the limit 0.2, in green, the SNR = 5 one (dotted lines). In dark, the maximum of these two curves. We keep every object under the latter.

By performing similar cuts for the magnitude error, we can further filter our dataset to select only reliable sources. In the end, out of the 784'016 objects in the COSMOS catalogue, we keep 382'103. The mean magnitude of our dataset is 25.86. Comparing this with UltraVISTA's 5σ magnitude depth of 25.8 [1], (as stated in Table 1 of [16], or Table 6 of UltraVISTA's DR6 report [1]) confirms the relevance of our conditions.

We have now successfully selected the sources we will be working with, that are precise enough for our study. We know that they can be trusted, and that results obtained with them can be of interest.

[1]The 5σ magnitude depth, or limiting magnitude of a telescope gives an idea of how far it can see. The higher it is, the deeper it sees through the universe, observing fainter targets.

3 Simulating a Euclid-like catalogue

This section describes how the simulated Euclid-like catalogue has been created. It also discusses the various assumptions that have been made as well as their validity, and the results obtained. We follow the method described by [Payerne et al. [10]] (Section 2.3), to degrade the deep imaging from the COSMOS field in order to mimic a shallower one, simulating *Euclid*'s. This will allow us to perform target selection at *Euclid*'s depth, while its data has not yet been published.

3.1 Hypotheses

We will make a set of assumptions, mostly well-justified. We will first consider the flux as a Gaussian random variable in a high signal-to-noise ratio regime, such as magnitude errors can also be considered as Gaussian random variables. Whilst the first assumption is frequent in the literature, the second is justified by our masks, filtering low SNR objects. We will also assume all our objects to be point-sources, not taking into account any morphological parameters.
Then, we separate the noise-to-signal ratio (NSR) between the systematic error (NSR_{sys}) and the random error (NSR_{rand}). The former expresses errors in imperfect modeling of the Point-Spread Function (PSF) for instance, whereas the latter takes into account random processes in photon counts ([10], [7]). As was assumed by [Payerne et al. [10]], we will take the optimistic scenario in which we neglect the systematic uncertainties. From [7] (Section 3.2.1), we know that the systematic error is maintained inferior to 0.005 mag, therefore neglecting it does not seem to be too much of a simplification. The random photometric NSR can be expressed as follows :

$$NSR_{rand}{}^2 = (0.04 - \gamma)x + \gamma x^2 \tag{2}$$

Where γ is related to the photon count for a SNR of 5 (hereby noted C_5), and :

$$x = \frac{C_5}{C} = 10^{(m-m_5)} \tag{3}$$

With C the photon count and m_5 the 5σ depth in a given band. Details of the calculations are given in Appendix B of [3].
We will assume γ to be equal to 0.04, to simplify Equation 2. As shown in Table 2 of [7], the typical value of γ in the Y-band is 0.039. It thus makes sense to round it up to 0.04. However, we could not find any value for the J- and H-band, as it depends heavily on the observation conditions as well. We will assume 0.04 to be a suitable approximation for these bands, considering the different values present in Table 2 of [7] for the other wavelengths.
Finally, we will assume that the input and output band filters are similar. As Figure 2 and Table 1 show, the wavelengths range of UltraVISTA's and *Euclid*'s Y-, J- and H- bands are close, and we will consider them to be akin, to avoid further complications.

Our simulation is therefore optimistic, not including specific features from either the input or the target output datasets. However, it has the advantage of being done relatively easily, and the results are still relevant, useful and useable, as we will see.
We note here that we will use data from the entire COSMOS field, as it has been roughly observed

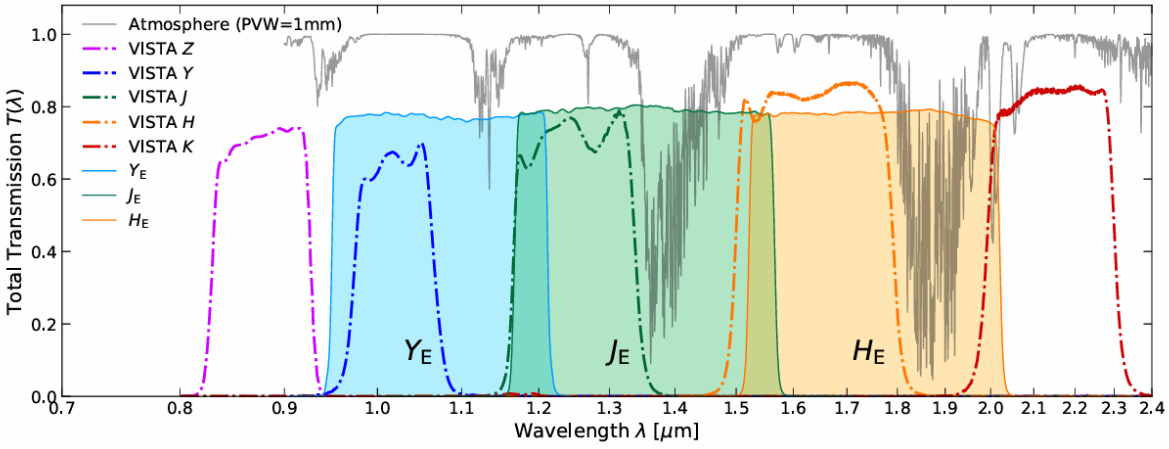


Figure 2: Comparison between NISP's (background) and UltraVISTA's (dotted lines) near-infrared filters. See Figure 1 of [Euclid Collaboration et al. [5]]

	UltraVISTA (nm)	***Euclid*'s NISP (nm)**
Y	926 - 1129	920 - 1146
J	1124.6 - 1403.7	1146 - 1372
H	1432.6 - 1866	1372 - 2000

Table 1: Comparison between UltraVISTA's and *Euclid*'s near-infrared filters. The values come from [13] and [8], respectively.

everywhere at the same depth with UltraVISTA (see [1], the article describing UltraVISTA's sixth data release (DR6)). Figure 3 also gives an estimation of the two-dimensional 5σ magnitude depth map of UltraVISTA's Y-filter over the COSMOS field (see Figure 2 of [10]). As we can see, the entire COSMOS field seems more or less homogeneous, meaning that we can use the same input magnitude depth for every region of it. The resuts are the same in the other bands.

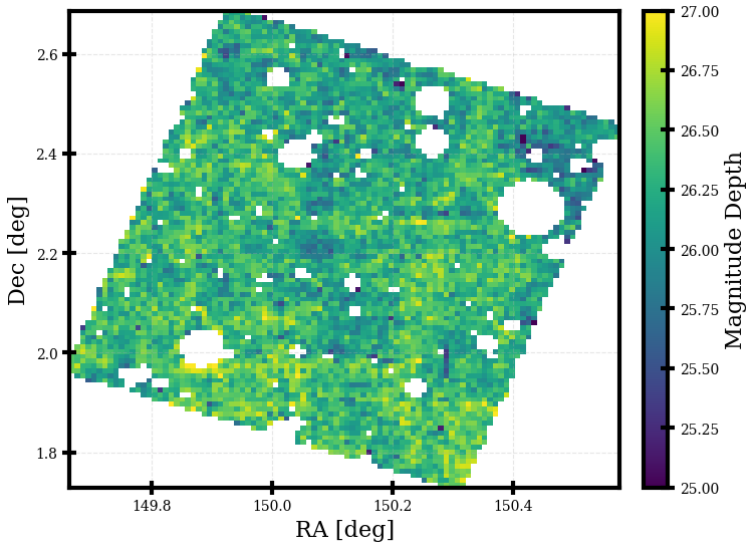


Figure 3: 5σ magnitude depth estimates over the COSMOS field, in UltraVISTA's Y-band.

3.2 Simulation

3.2.1 Formulas

We simulate the shallow flux, $f_{shallow}$, from the deep imaging f_{deep} as a normal distribution, centered around f_{deep}, as follows (See Equation (2.3) of [10]) :

$$f_{shallow} \sim \mathcal{N}(\mu = f_{\text{deep}},\, \sigma^2 = \left[\sigma_f\right]^2_{\text{shallow}} - \left[\sigma_f\right]^2_{\text{deep}}) \tag{4}$$

Where $\left[\sigma_f\right]^2_{\text{shallow}}$ and $\left[\sigma_f\right]^2_{\text{deep}}$ are the flux errors measured with the 'shallow' and 'deep' magnitude depths, i.e respectively 24 and 25.8 (or 25.5 in UltraVISTA's H-band) mag. See Appendix A for the details.
Once we find a value for the 'shallow' flux, we can calculate the 'shallow' magnitude, using the definition of a magnitude:

$$m_{shallow} = 23.9 - 2.5 \cdot log_{10}(f_{shallow}) \tag{5}$$

Where the Zero-Point value has been set to 23.9 because COSMOS' fluxes are given in μJy.
From there, we can find the various uncertainties (see Appendix A for the formulas) and build our catalogue containing every relevant value.

3.2.2 Python function

To do so, we wrote a Python function that converts COSMOS' deep field data into simulated Euclid-like values for the flux, and calculates the magnitude, the errors and the respective SNR. It then builds the Euclid-like catalogue, containing some relevant columns from the initial COSMOS catalogue and adding the new ones. Here is its prototype:

```
def COSMOS_to_Euclid(cat_COSMOS,
    m_depth_shallow = 24,
    output_filename = None,
    additional_columns_photom = None,
    additional_columns_lephare = None,
    mag_min = 17,
    mag_max = 23.5):
```

Its first argument is the COSMOS catalogue, either as a path (in which case it opens it), or as an HDUList. The second parameter is the target 5σ magnitude depth, already initialised at *Euclid*'s. Then comes the output filename. If none is given, the function returns the catalogue as an HDUList; else, if a name (or a path) is given, it writes there the catalogue as a .fits file. The two following arguments are optional additional columns from the COSMOS catalogue that can be added to the simulated one, if needed. The first takes headers from the first column (*Photometry (PSF-homogenized Aperture and SE++ Model-Based)*) of the COSMOS catalogue, the second from the second column (*LePHARE Photometric Redshifts and Physical Parameters*).

There are already some columns that are automatically added to the simulated catalogue. They are listed in Table 2. See their description on the COSMOS-Web DR1 Catalog website ([9]). Finally, the last two arguments concern the detection probability function and will be discussed later, in Section 4.

Photometry	'id', 'ra', 'dec', 'mode', 'ra_model', 'dec_model', 'flag_star', 'flag_blend', 'flag_star_hsc', 'warn_flag', 'mag_model_uvista-y', 'mag_model_uvista-j', 'mag_model_uvista-h', 'mag_err_model_uvista-y', 'mag_err_model_uvista-j', 'mag_err_model_uvista-h', 'flux_model_uvista-y', 'flux_model_uvista-j', 'flux_model_uvista-h', 'flux_err-uncal_model_uvista-y', 'flux_err-uncal_model_uvista-j', 'flux_err-uncal_model_uvista-h', 'flux_err-cal_model_uvista-y', 'flux_err-cal_model_uvista-j', 'flux_err-cal_model_uvista-h'
LePHARE	'zfinal', 'type'

Table 2: Columns from the COSMOS catalogue automatically added to the simulated catalogue. Note that the 'id' column is renamed 'id_COSMOS'.

The function hence calculates the shallow flux and magnitude in each band, their respective errors, the shallow SNR and new flags. Table 3 gives the names of these new columns.
The first flag ('flag_mask') tells whether an object satisfies the general mask applied or not.
The second flag ('flag_NaN_uvista-B', where B is the band (y, j or h)) marks objects that have an undefined value either for the shallow flux or magnitude. Indeed, as the shallow flux depends on the noise to be added to the deeper flux, it can sometimes be undefined. This happens for two reasons:

1. First, if the simulated error happens to be inferior to the deeper one, there is a negative square-root, which is undefined. This occurs only if we take the COSMOS' value for the deep flux error, that can be extremely low. If we follow Equation 24 (in Appendix A), as long as the magnitude depth is fainter in the deep imaging, it never happens.

2. Second, as we calculate the shallow flux using a random draw on the normal distribution, if the noise is negative and in absolute values bigger than the deep flux, we get a negative simulated flux, and hence an undefined magnitude.

As these two conditions show, we want to avoid using undefined values, because they most probably mean that the initial value was unreliable; hence this new flag, allowing us to effectively sort these objects.
Finally, the third flag ('flag_detected_uvista-B', with B being y, j or h) marks objects that are unlikely to be detected by *Euclid*, as we will see in Section 4.

Name	Description
id_COSMOS	Unique id of the object in the COSMOS catalogue
flux_euclid-like_uvista-B	Simulated flux in the band B (B being y, j or h)
mag_euclid-like_uvista-B	Simulated magnitude in the band B (B being y, j or h)
err_flux_euclid-like_uvista-B	Simulated flux error in the band B (B being y, j or h)
err_mag_euclid-like_uvista-B	Simulated magnitude error in the band B (B being y, j or h)
snr_euclid-like_uvista-B	Simulated Signal-to-Noise Ratio in the band B (B being y, j or h)
flag_mask	Flag returning 0 if the object satisfies the conditions and 1 if not (with the Mask 4, see Section **??**)
flag_NaN_uvista-B	Flag returning 0 if no value is undefined and 1 if the simulated flux and/or magnitude is undefined, in the band B (B being y, j or h)
flag_detected_uvista-B	Flag returning 0 if the object is likely to be detected by *Euclid* and 1 if not, in the band B (B being y, j or h) (see Section 4)

Table 3: Names and definitions of the new columns in the simulated catalogue.

3.3 Results

Now that we have a simulated, Euclid-like catalogue, we will analyse it and determine whether it can be reliable and of use.
We will first compare the deep and the shallow magnitude. Figure 4 shows the two distributions, in the Y-band. As expected, the shallow magnitude is, on average, lower than the deeper one, as *Euclid*'s depth is inferioir to UltraVISTA's.

It is also interesting to compare both errors. Figure 5 does so, representing both the shallow and deep errors, in orange and blue respectively. They have been smoothed by a moving average, to ensure a better visibility. We can see that around a SNR of 5, the shallow magnitude is at 24 mag, and the deep one around 26, corresponding to *Euclid*'s and UltraVISTA's Y-band depth, respectively. We can compare this to Figure 4, left panel, of [Payerne et al. [10]].

If now, we only interest ourselves with objects that have a Signal-to-Noise Ratio above 5 (i.e objects under the red-dotted line in Figure 5), Figure 6 shows the magnitude distribution.

As we see, there is a peak around 24 mag for the shallow one, and it then decreases progressively. It follows the same patterns as for the deep magnitude, the only difference being the limiting magnitude. This figure can be compared to Figure 4, right panel, of [Payerne et al. [10]].

In the end, we have about 48'000 objects with a SNR higher than 5 in the simulated catalogue.

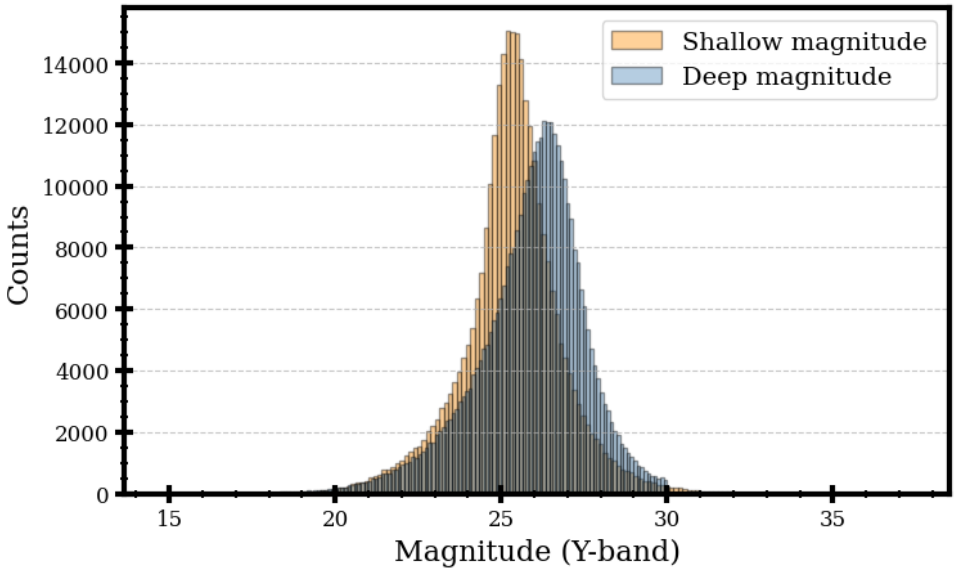


Figure 4: Comparison between the deep (blue) and shallow (orange) magnitudes, in the Y-band.

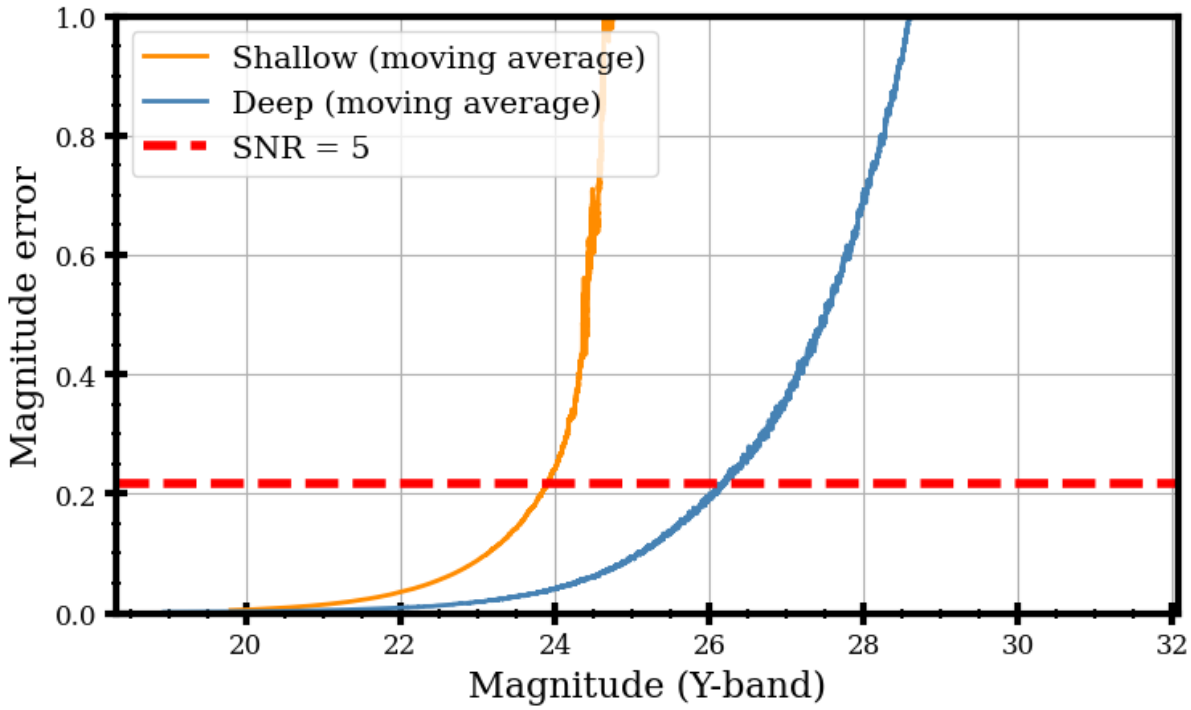


Figure 5: Comparison between both magnitude errors, smoothed by a moving average. The red-dotted line shows a SNR of 5.

4 Detection Probability Function

This section provides a method for determining the probability for an object of being detected by a telescope, given only its magnitude. It uses a linear regression to determine the expected number of targets, and, comparing it with the actual number of objects detected, it fits a specific function (see [Snigula et al. [17]], equation [18]) to the data. Using this method, we hope to filter our catalogue and to keep only objects that can actually be detected by *Euclid*, removing all the

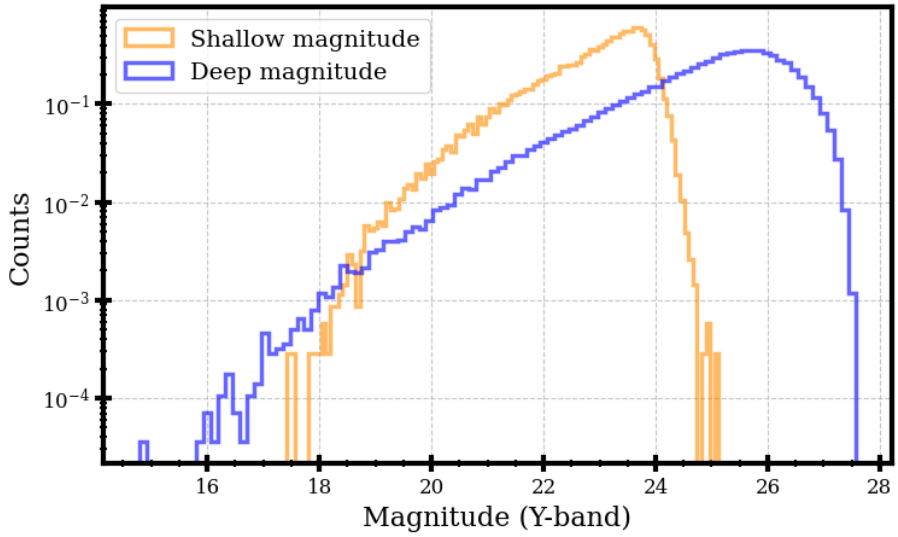


Figure 6: Comparison of shallow (orange) and deep (blue) magnitudes, with a high SNR (> 5), in the Y-band.

sources too faint, or of too low-quality. This reinforces the simulated catalogue's reliability, and allows us to better trust the obtained results. We use the notion of "completeness", which gives the percentage of objects that are detected by an instrument. A 100% completeness means that every existing object at this depth has been detected.

4.1 Background calculations

Here beneath, we describe the basic calculations that ground our hypothesis. We justify the following relation :

$$log_{10}(N) = \alpha \cdot m + \beta \tag{6}$$

Where N is the number of objects, m is the magnitude, and α and β are parameters to be found. The definition of the magnitude is the following :

$$m = C - 2.5 \cdot log_{10}(f) \tag{7}$$

With C a constant depending on the system we use, and f the flux. Furthermore, we know that the flux depends on the luminosity and the distance :

$$f \propto \frac{L}{4\pi d^2} \tag{8}$$

Now, if we assume that the distribution of objects is homogeneous and that there is no extinction, nor any further corrections to take into account, we can say that the number of objects within a certain sphere of radius d varies with the volume of this sphere, hence :

$$N(\text{objects at a distance} < d) \propto d^3 \tag{9}$$

From equation 8, we see that f and d^{-2} are related, and from equation 7, we have that :

$$f = 10^{-0.4(m-C)} \implies f \propto 10^{-0.4m} \tag{10}$$

Putting all of these conditions together, we get :

$$N \propto d^3 \propto (10^{-0.4m})^{-\frac{3}{2}} = 10^{0.6m} \tag{11}$$

Which finally gives us :

$$log_{10}(N) \approx 0.6m + \beta \tag{12}$$

Here we made some strong assumptions, especially considering that we are dealing with distant galaxies. There will be some extinction, cosmological effects and K-corrections to take into account, resulting in the slope not necessarily being at 0.6. However, the general behaviour for a fixed completeness will still be linear (see Figure 7, left panel for instance).
A perhaps better estimate, taking these details into account, would be the integral of the Schechter luminosity function ([Schechter [15]]), that can also be expressed in terms of magnitude. However, it needs more parameters and is less convenient to use, as it is not a universal function. The parameters depend on the type of objects that are observed. It might be interesting to compare our results with some coming from this alternative method, though.

4.2 Detection probability function

We can know use the fact that the logarithm of the numbers of objects depending on the magnitude should follow a linear function. To determine the slope α and the y-intercept β, we will assume that in a given magnitude interval, all the objects are being detected. Therefore, in this specific interval, we can fit the linear function and find the parameters.
Knowing that the 5σ magnitude depth of *Euclid* in the Y-band is of 24 mag, it seems reasonable to take values for this interval between 20 and 24 mag. However, we did not find a way to precisely determine these values. While the curves do change quite significantly depending on these initial conditions, we noticed that in the end, the number of detected objects is more or less the same. We compared four different intervals of magnitude. To save space, we will only put the results from the one we kept, namely [17 - 23.5]. Indeed, as the brightest object in our catalogue has a magnitude of about 17.4 mag, we will use 17 for the lower threshold, and 23.5 for the upper one, given that *Euclid*'s 5σ depth is of 24 mag; we decided that a slightly lower value would be the most effective (also based on comparisons with other values).

Figures 7 and 8 show the different diagrams and graphs for the final interval. See the other diagrams in Appendix B, as well as tables summarising the results.

4.2.1 Methodology

Using this chosen interval we consider to be perfectly covered, we can fit the best affine function for this set of data. This will be our reference for later, giving us for each magnitude, the total amount of objects that should be detected with a perfect completeness.
We then separate our magnitude range in different bins of the same size (here, we used a 0.1 mag width). For each bin, we can calculate the number of detected sources, and estimate the total amount of objects, using the method described above. Next, we can attribute a probability of

detection value for each bin :

$$P_D(m) = \frac{N_{detected}}{N_{total}} \tag{13}$$

This gives us a set of couples (magnitude, probability), that we will use to best-fit the probability function. Figure 7 shows the repartition of the objects (left) and plots the respective completeness for each bin (right) (see also Appendix B).

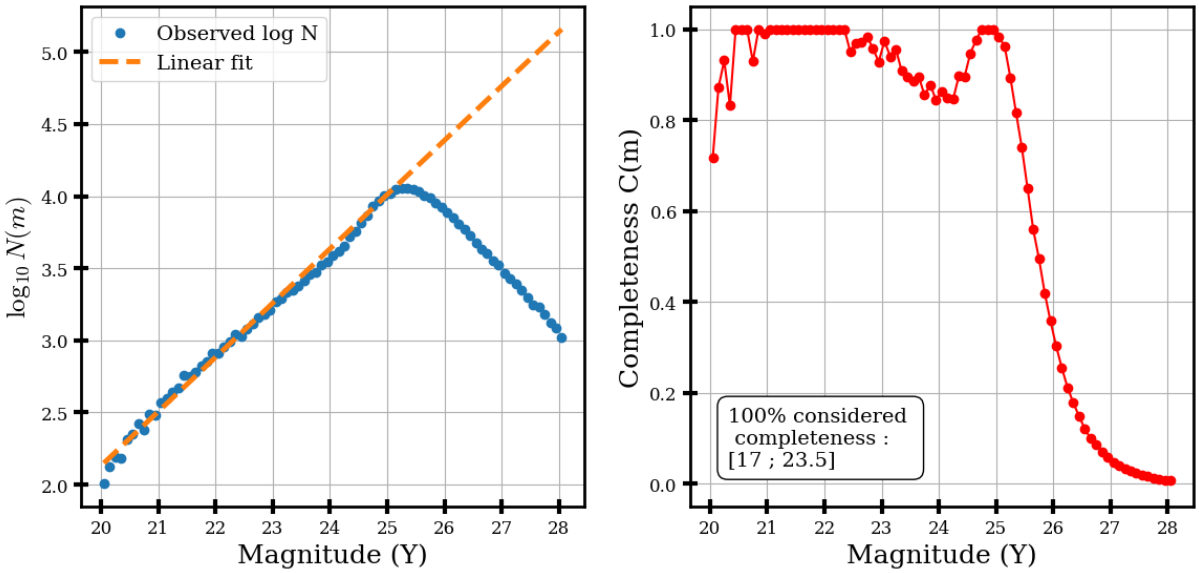


Figure 7: *Left* : log_{10}(N) as a function of the magnitude, in blue, and the linear regression in orange. We see the linearity up to almost 25 mag. *Right* : completeness given by the left figure. Here, the "100% completeness" interval was [17 - 23.5].

Now that we have these couples of points, we can find the best function that fits them, which will give us our detection probability function. The form of this function is given by equation [18] of [17]. It is as follows :

$$P_D(m) = \frac{p_0}{(\frac{m}{m_0})^a + (\frac{m}{m_0})^b} \tag{14}$$

Where p_0, m_0, a and b are the fit parameters. We can understand their meaning as such :

- p_0 gives an estimate of the plateau height, i.e the maximal completeness. The closer it gets to 1, the more effective the telescope is.
- m_0 is the magnitude corresponding to a 50% completeness.
- a gives the slope of the first part of the curve, which is often close to flat. It controls the slope for $m < m_0$.
- b gives the slope for the second part of the curve, which often decreases fast. It controls the slope for $m > m_0$.

Using equation 14 and the couples (magnitude, probability) found before, we have everything to find our detection probability function. We can fit the parameters the best, and it will give us what we need. Figure 8 shows the function for the initial conditions we imposed.

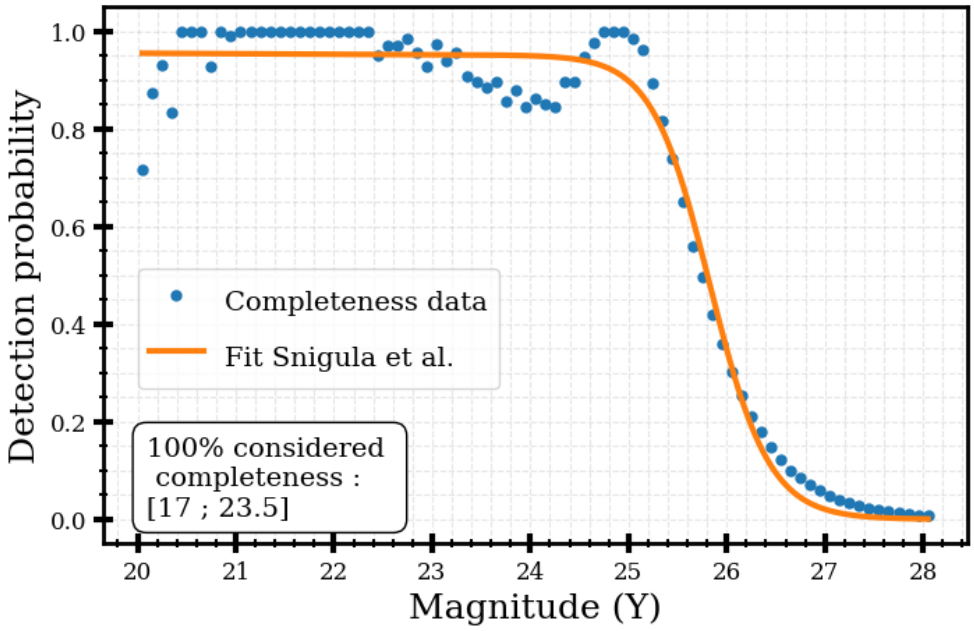


Figure 8: Best-fit of [Snigula et al. [17]] with [17–23.5] interval.

4.3 Discussion

The best-fit parameters for our chosen interval are listed in Table 4. See Appendix B for the other intervals.

Parameters	Value
m_0	25.84
p_0	0.95
a	0.036
b	87.25

Table 4: Best-fit parameters for a 100% considered completeness magnitude interval of [17, 23.5].

As said, we decided to keep this interval as the brightest object in our catalogue has a magnitude of about 17.4 mag, hence using 17 for the lower threshold seems coherent. As for 23.5 for the upper one, it comes from *Euclid*'s 5σ depth of 24 mag. We decided that a slightly lower value would be the most effective, also based on the graphs of Appendix B).
Now that we have this probability function, we can sort our simulated catalogue using a stochastic test. We generate a random number between 0 and 1 for each object, and if the number is lower than the object's probability of detection, the object is kept. This way, the lower an object's probability is, the less likely it is for it to be detected, which corresponds to the reality. This gives us another tool to have a more realistic simulated *Euclid* catalogue. With our interval, 48.51% of the sources would realistically be detected by *Euclid*.
Note that we could also use this method to find other probability functions. For instance, if we wanted to know the probability for an object to be observed at a 5σ precision by *Euclid*, we could use the same methodology, and replace the detected number of sources by the detected number of 5σ sources.

Finally, we used a function coming from [Snigula et al.]'s [17] article (2002). We could use a different type of function, namely a sigmoid function, of the form :

$$P_D(m) = \frac{p_0}{1 + e^{k(m-m_0)}} \tag{15}$$

The resulting curve would be extremely similar to ours, except here the parameter k encompasses both previous parameters a and b. We preferred using [17]'s function, as it is more referenced in the literature and its parameters bear more intuitive meaning.

5 ELGs target selection

With the dataset prepared, we can proceed to the selection of ELGs from the simulated Euclid-like catalogue. We will do so using the colour-colour method[2] and first sorting the redshifts[3] we are interested in. Indeed, as we intend to mimic *Euclid* observations, we will target objects with a redshift between 0.8 and 1.8, as will be done with *Euclid* (see [8] or [2]). More specifically, we will target objects with $0.8 \leq z \leq 1.6$, as spectroscopic measurements are reliable only up to this redshift.
We will also produce a final mask, sorting objects that satisfy the basic conditions (`flag_mask` = 0), have no NaN value (`flag_NaN_uvista-B` = 0 in each band), have a positive photometric redshift and a Signal-to-Noise Ratio above 5 in each band. We will not use the detection probability function, so that we can keep working on every band at the same time, with the same objects, which is more convenient for colour-colour diagrams. Indeed, as the detection probability function involves random processes, it is best to use it separately in each band, which is not convenient for colour-colour diagrams, as we mix bands. We will thus not use it, having already sorted unreliable sources before anyways. This leaves us with 43'876 objects.

5.1 Redshift selection

As shown in Figure 9, the initial photometric redshift distribution, coming from the LePHARE catalogue, contains many objects with photo-z's[4] inferior to 0.8 or superior to 1.6. Our objective is to use photometric measurements and colour-colour diagrams to filter these objects, and keep a homogeneous distribution along our desired range, with fewest possible objects outside of it.

We will follow the method used for Figure 3 of [Richard et al. [12]]. We first have to find a diagram that shows a clear gradient, to separate the redshifts based solely on colour. Then, we will produce a 2-dimensional histogram, fit it best and use these fitted curves to find the most efficient cuts.

[2]The colour-colour method consists of creating diagrams with differences between two bands. This way, we can analyse the received flux in each band better.

[3]The redshift of a galaxy, noted z, is basically the same as its distance to us. It gives an idea of how long its light had to travel to come to us, hence also providing insights on the universe's age when this light was emitted. For instance, at $z = 1$, light traveled for about 7.7 billions years before reaching us, so it shows us how the universe was 7.7 billion years in the past.

[4]A photo-z, or photometric redshift, is an estimate of the redhsift using only photometric data.

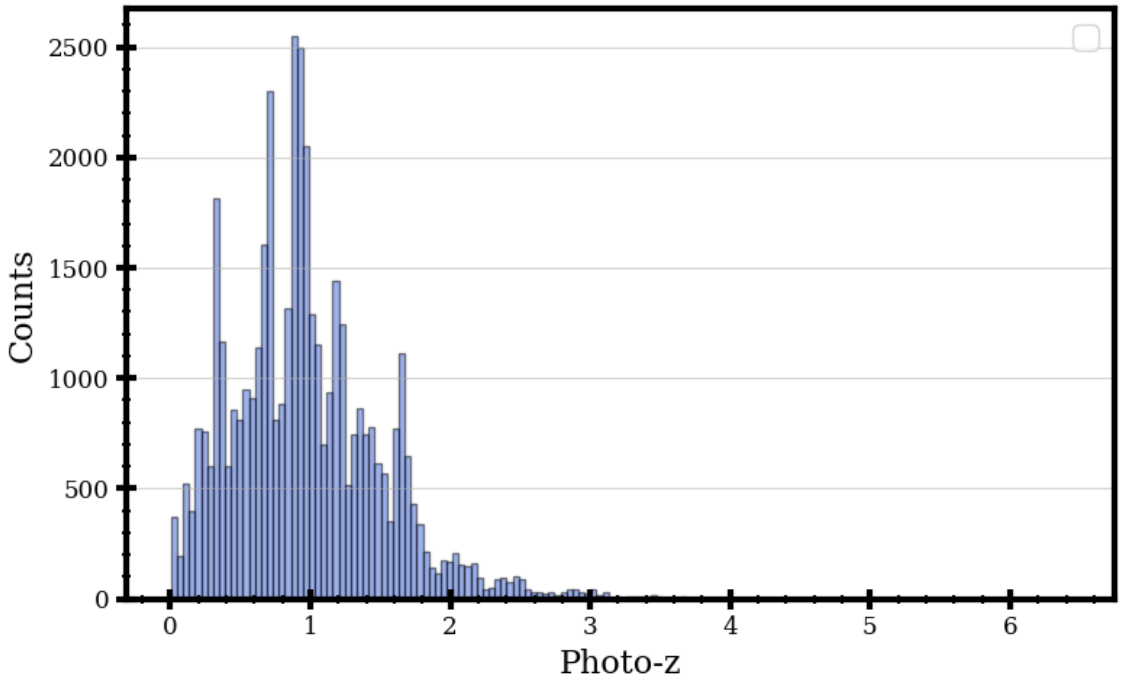


Figure 9: Initial photometric redshift distribution

Unfortunately, using only Y-, J- and H-bands lead nowhere, as we could not find a combination separating redshifts well enough. Appendix C shows the attempts we have made and their corresponding results. In order to continue our work, we had to introduce other bands. Therefore, we used the CH1 and CH2 bands from *Spitzer*'s Infra-Red Array Camera (IRAC) (see [Fazio et al. [6]]). These bands are part of the COSMOS catalogue, and they cover longer wavelengths, offering a complimentary coverage to *Euclid*. They are also similar to WISE's, another telescope surveying a large portion of the sky, hence, if we get results with our simulated *Euclid* and IRAC datasets, it is reasonable to expect that these results will also hold with actual *Euclid* and WISE imaging, although WISE has a brighter limiting magnitude than IRAC.

5.1.1 Colour-colour diagram

We found that the diagram showing best the redshift gradient is J-H vs H-CH2, shown in Figure 10. We can clearly see how the redshift increases the higher we get in the diagram. The reason why this particular diagram works best is probably because the NIR bands (J-H) capture the 4000 Å break, whilst the longer wavelengths (H-CH2) account for the 1.6 μm bump ([14]).

We divide the redshift range into six bins of width 0.4, covering z=0 to z=1.6, including a bin containing all galaxies with $z < 0.01$ and another encompassing galaxies with $z > 1.6$. We will then try to keep galaxies with $0.8 \leq z \leq 1.6$, hence keeping only the two central bins.

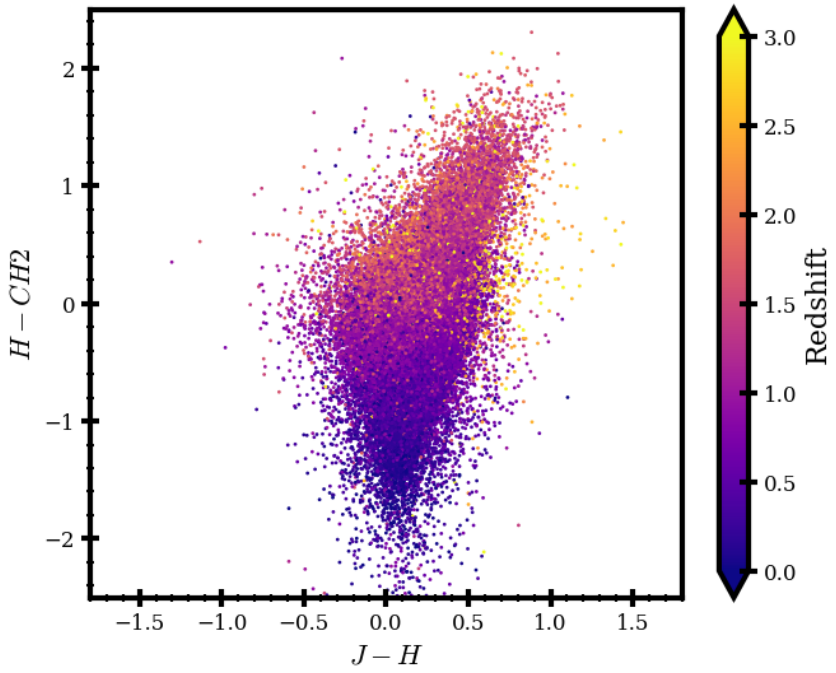


Figure 10: J-H vs H-CH2 colour-colour diagram, showing best the redshift gradient.

5.1.2 2-dimensional histograms

We will now plot a 2-dimensional histogram of the redshift distribution according to each colour, for each bin. This allows us to see how the redshift distribution moves along the diagram, while increasing. We can see these histograms on Figure 11.
Next, we will fit a 2-dimensional Gaussian onto these histograms, to map the redshift distribution. These fits will be later used to determine the most efficient cuts, in order to filter our objects according to our redshift range. Figure 11 shows these Gaussians as grey ellipses on the histograms.

5.1.3 Redshift distribution and cuts

Finally, we can use these Gaussians on the same diagram, to filter the redshifts. Figure 12 shows the fitted redshift distribution, and we can clearly see the difference between each bin (compared to Figure 19, Appendix C).
We will now use this diagram to remove objects that are not within our redshift range, i.e objects with either a redshift below 0.8 or superior to 1.6. In short, we will try to keep only the two middle bins, respectively in light red and orange on Figure 12. The latter shows our final cut, which is the following:

$$(H - CH2) < 2.069 \cdot (J - H) + 0.214$$
$$(H - CH2) > 0.3 \cdot (J - H) - 0.14$$
$$(H - CH2) > 1.9 \cdot (J - H) - 0.62$$

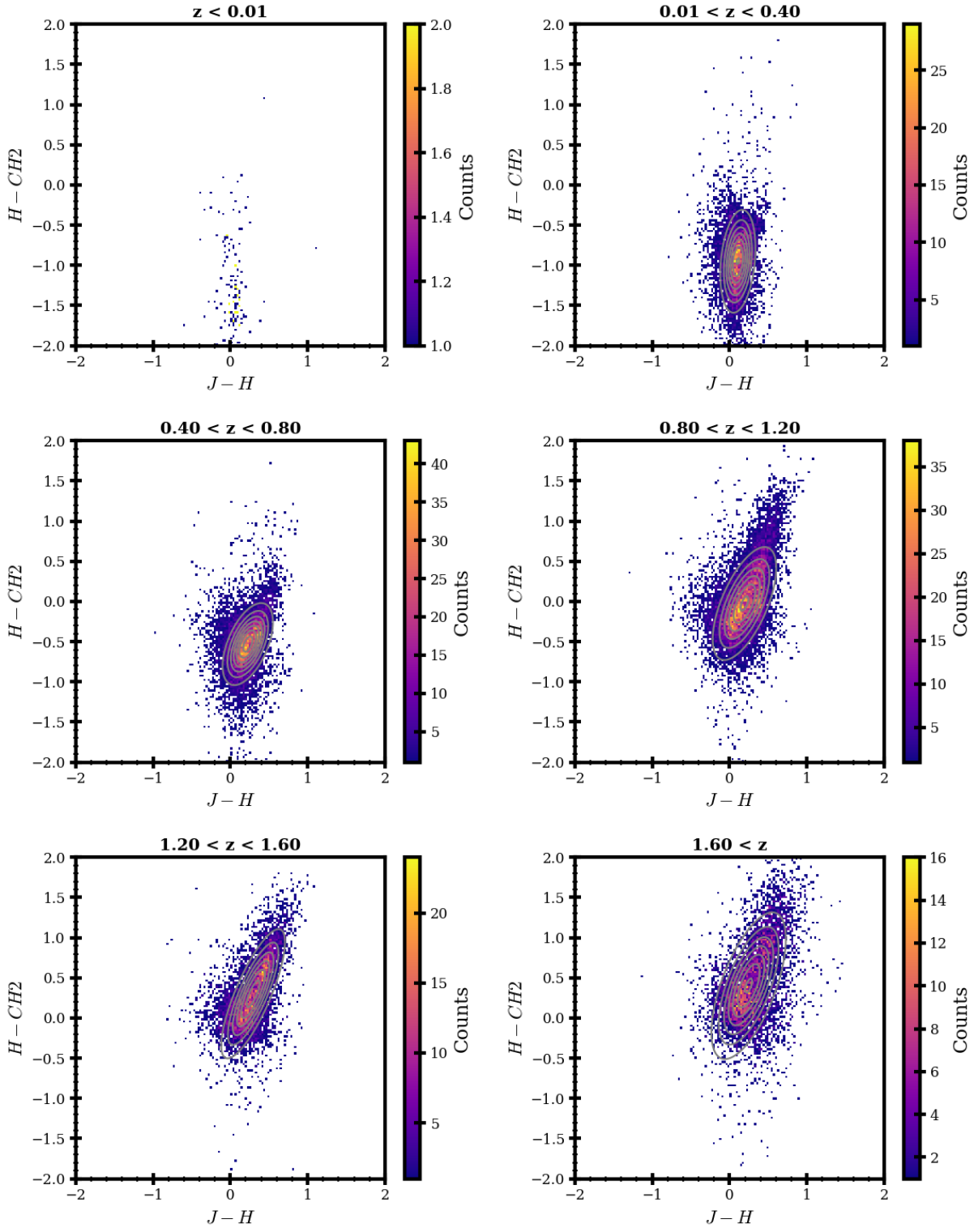


Figure 11: 2-dimensional histogram of the redshift distribution along the J-H vs H-CH2 diagram. In grey ellipses is the best fit for the 2-dimensional Gaussians.

5.1.4 Results

As we can see on Figure 12, these cuts remove efficiently the lower redshifts. Figure 13a compares the initial photometric redshift distribution, in blue, with the one after the cuts, in orange. We

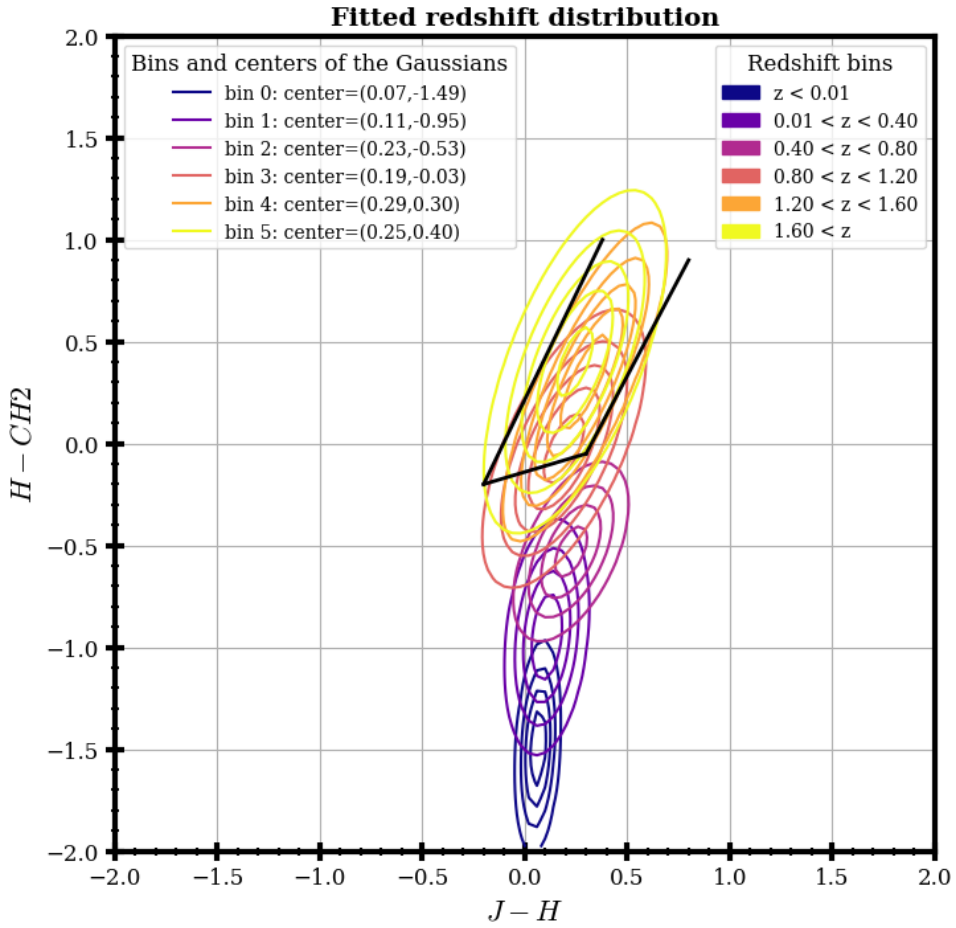
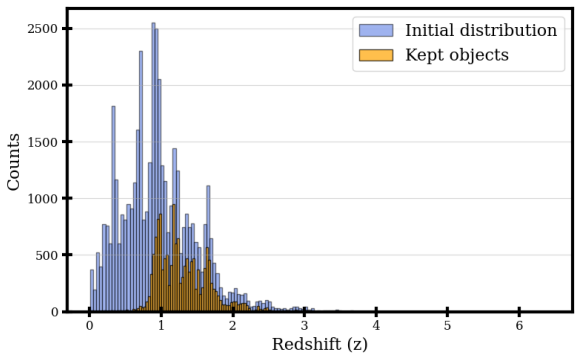
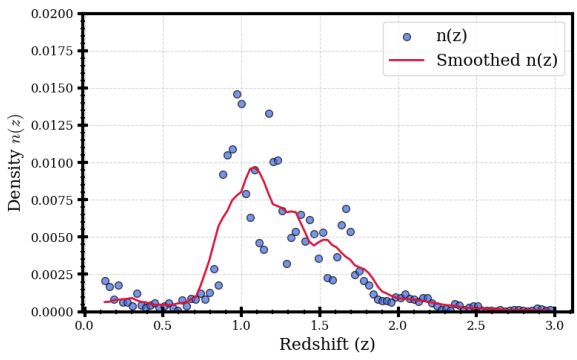


Figure 12: Comparison of the J-H vs H-CH2 redshift distribution and its corresponding cuts, shown in black.

can clearly see that our objective has been reached: very few low redshift objects are kept, and the distribution decreases after z ≈ 1.6 as well. Moreover, within our target range, the redshift distribution seems more or less stable. Indeed, as shown on Figure 13b, the photometric redshift density is roughly flat between 0.8 and 1.6, with only a peak around z ≈ 1.2, although it remains below 0.01.



(a) Photometric redshift distribution, before (blue) and after (orange) the cut.



(b) Photometric redshift density. The red curve is obtained through a moving average.

Figure 13: Comparison of the photometric redshift distribution (left) and density (right).

We have now a selection of about 15'000 objects, within roughly the appropriate redshift range. This constitutes the first step in the ELGs selection, which could be further refined using additional colour-colour diagrams targeting emission lines such as H_α, [O III], H_β and [O II], at more specific redshifts. Using spectroscopic measurements, such as DESI's, we could also confirm these results by analysing the targets' spectra. However, a detailed investigation of these refinements is beyond the scope of the present study.

6 Conclusion

This work explored the feasibility of generating a Euclid-like catalogue from UltraVISTA data using the method described by [Payerne et al. [10]], providing a realistic framework for anticipating the survey's observations. Its results can be used to study the efficiency and reliability of the method in near-infrared bands and on different telescopes than [10]. With the arrival of *Euclid*'s first public data release in October 2026, it will allow us to compare both datasets and have further analyses of the method, sooner than the expected LSST releases from [10]. It also gives us the opportunity to study *Euclid*'s future data beforehand and already find some results.

In order to evaluate the reliability of the simulated dataset, a method was developed to determine the likelihood for an object to be detected by a telescope given only its magnitude. It uses a detection probability function found in [17] and some initial conditions such as a 100% completeness interval. From there, with a stochastic process, each object gets assigned a boolean value, determining whether it would realistically be detected by *Euclid* or not. With this parameter and some other filters, we can ensure a clean, high-quality and reliable simulated catalogue, ready to be used for some analyses.

Using this simulated Euclid-like catalogue, we have then performed an ELGs selection. Having no other choice, we had to use additional bands in the infrared, namely CH1 and CH2 coming from *Spitzer*'s IRAC. With a J-H vs H-CH2 colour-colour diagram, we managed to find efficient cuts and to reduce our dataset to 15'000 objects we assume to be ELGs, within the appropriate $0.8 \leq z \leq 1.6$ redshift range corresponding to *Euclid*'s. IRAC's bands could be replaced by WISE's for a larger scale study, as they are extremely close. They only differ slightly in wavelengths; however, their limiting magnitude is different, WISE's being inferior to IRAC's or *Euclid*'s. These 15'000 potential targets can be confirmed using spectroscopy, which should be possible with DESI's observations: finding them and analysing their spectrum, verifying whether they are ELGs or not.

Acknowledgements

I first want to thank the Laidlaw foundation and the EPFL Laidlaw representatives for making this internship possible. It has been an amazing opportunity, teaching me valuable lessons both in an academic and a personal manner. As merely a first-year student, being able to undertake an internship in such an interesting field is incredible, and I am really grateful for it.

Then, I would like to thank the Laboratory of Astrophysics for accepting me and trusting me with this project. More specifically, I deeply want to thank Aurélien Verdier for his time, patience, engagement and most valuable pieces of advice. Without him, this work would undoubtedly not have been possible. It has been a real pleasure working at his side, and I could not have asked for a better supervisor for my first experience in research.

I also want to thank Constantin Payerne, that significantly helped me understand his (and my) work, allowing me to further it. His experience, thoughts and advice meant a lot to me and definitely helped me in my project. Thank you for your time and reactivity.

Finally, I am grateful to Thomas Delaloye for his support in writing this LaTeX document and giving it a better outlook. I believe his help will be extremely useful for my future.

Glossary

5σ Depth	The faintest magnitude an instrument can reliably detect at 5σ confidence.
5σ Precision	A measure of confidence in detecting a signal, where "5σ" corresponds to a very low chance of a false detection.
Band	A specific range of wavelengths observed by a telescope or instrument. Of interest to our work are the near-infrared bands Y, J and H.
Colour-Colour Diagrams	Plots comparing brightness in two different bands; used to identify types of objects like ELGs.
Completeness	The fraction of objects of a given type or brightness that are successfully detected in a survey.
COSMOS Field	A well-studied patch of the sky observed by many telescopes to provide deep multiwavelength data.
Emission Line	A feature in a spectrum where a galaxy or star emits more light at a specific wavelength, creating a sharp peak.
Emission Line Galaxies (ELG)	Galaxies that shine more brightly at certain wavelengths due to strong emission lines, making them easier to identify.
Euclid	A space telescope launched by the European Space Agency to map the geometry of the Universe and study dark energy using visible and near-infrared observations.
Flux	The amount of light coming from an astronomical object that reaches a telescope; it tells us how bright the object appears.
Gaussian Random Variable	A random variable whose values follow a normal (Gaussian) distribution, often used to model noise in measurements.
Magnitude	A scale for the brightness of an astronomical object; larger numbers mean fainter objects. The unit is written "mag".
MicroJansky (μJy)	A unit of flux density equal to 10^{-6} Jansky, where 1 Jansky = 10^{-26} W m^{-2} Hz^{-1}; thus 1 μJy = 10^{-32} W m^{-2} Hz^{-1}, used to measure very faint astronomical sources.
Near-Infrared (NIR)	Light just beyond visible red light, invisible to the eye but detectable with special instruments.
Photometry	Measuring the brightness of objects in different bands.
Point Spread Function (PSF)	Describes how a point source of light (like a distant star) is spread out by a telescope and atmosphere, showing the blurring effect in images.
Photo-z	A redshift estimated from photometric measurements (brightness in different bands) rather than full spectra.

Redshift (z)	The stretching of light from distant galaxies due to the expansion of the Universe; used to estimate distance.
Signal-to-Noise Ratio (SNR)	The strength of a signal (like a galaxy's light) compared to background noise; higher values mean more reliable measurements.
Spectroscopy	Measuring how an object's light is distributed across different wavelengths to study its composition and properties.
Survey	A systematic observational program that collects uniform data over a region of the sky to build catalogues of astronomical objects.
UltraVISTA	A deep near-infrared survey covering part of the COSMOS field, providing photometry in Y, J, H, and K bands.
Wavelength	The distance between successive peaks of a light wave; determines the colour or type of light.

References

[1] S. Rouberol J. Dunlop M. Franx J. Fynbo R. Bowler-B. Milvang-Jensen C. Gonzalez-Fernandez E. Gonzalez-Solares J. Irwin M. Irwin R. Blake N. Cross R. Mann M. Read E. Sutorius A. Moneti H. J. McCracken. *The Sixth UltraVISTA Data Release*. `https://www.eso.org/rm/api/v1/public/releaseDescriptions/221`. Accessed: 2025-07-25. 2024.

[2] Euclid Consortium. *Core science: cosmology*. 2025. URL: `https://www.euclid-ec.org/public/core-science/` (visited on 08/2025).

[3] John Franklin Crenshaw et al. "Probabilistic Forward Modeling of Galaxy Catalogs with Normalizing Flows". In: *The Astronomical Journal* 168.2 (July 2024), p. 80. ISSN: 1538-3881. DOI: `10.3847/1538-3881/ad54bf`. URL: `http://dx.doi.org/10.3847/1538-3881/ad54bf`.

[4] S. Escoffier et al. "The ELG target selection with the BOSS survey". In: *Proceedings of the annual meeting of the French Society of Astronomy & Astrophysics (SF2A 2012)*. Société Française d'Astronomie et d'Astrophysique. Nice, France, June 2012, pp. 427–431. URL: `https://in2p3.hal.science/in2p3-00866822`.

[5] Euclid Collaboration et al. "Euclid preparation - XVIII. The NISP photometric system". In: *AA* 662 (2022), A92. DOI: `10.1051/0004-6361/202142897`. URL: `https://doi.org/10.1051/0004-6361/202142897`.

[6] G. G. Fazio et al. "The Infrared Array Camera (IRAC) for the Spitzer Space Telescope". In: *The Astrophysical Journal Supplement Series* 154.1 (2004), p. 10. DOI: `10.1086/422843`. URL: `https://dx.doi.org/10.1086/422843`.

[7] Željko Ivezić et al. "LSST: From Science Drivers to Reference Design and Anticipated Data Products". In: *The Astrophysical Journal* 873.2 (2019), p. 111. DOI: `10.3847/1538-4357/ab042c`. URL: `https://dx.doi.org/10.3847/1538-4357/ab042c`.

[8] R. Laureijs et al. *Euclid Definition Study Report*. 2011. arXiv: `1110.3193 [astro-ph.CO]`. URL: `https://arxiv.org/abs/1110.3193`.

[9] L. Paquereau C. M. Casey O. Ilbert R. C. Arango-Toro H. J. McCracken M. Franco S. Harish-J. S. Kartaltepe A. M. Koekemoer L. Yang M. Huertas-Company et al. M. Shuntov H. B. Akins. *COSMOS-Web DR1 Catalog*. 2025. URL: `https://cosmos2025:780kgalaxies!@cosmos2025.iap.fr/citation.html` (visited on 08/2025).

[10] C. Payerne et al. *Selection of high-redshift Lyman-Break Galaxies from broadband and wide photometric surveys*. 2025. arXiv: `2410.08062 [astro-ph.CO]`. URL: `https://arxiv.org/abs/2410.08062`.

[11] A. Raichoor et al. "Target Selection and Validation of DESI Emission Line Galaxies". In: *The Astronomical Journal* 165.3 (2023), p. 126. DOI: `10.3847/1538-3881/acb213`. URL: `https://dx.doi.org/10.3847/1538-3881/acb213`.

[12] Johan Richard et al. "4MOST Consortium Survey 8: Cosmology Redshift Survey (CRS)". In: *Published in The Messenger vol. 175* pp. 50-53 (2019), March 2019. DOI: `10.18727/0722-6691/5127`. URL: `http://doi.eso.org/10.18727/0722-6691/5127`.

[13] Carlos Rodrigo. *Filter Profile Service, SVO*. 2025. URL: https://svo2.cab.inta-csic.es/theory/fps/index.php (visited on 08/2025).

[14] Marcin Sawicki. "The 1.6 Micron Bump as a Photometric Redshift Indicator". In: *The Astronomical Journal* 124.6 (2002), p. 3050. DOI: 10.1086/344682. URL: https://dx.doi.org/10.1086/344682.

[15] P Schechter. "An analytic expression for the luminosity function for galaxies". In: *Astrophys. J.; (United States)* 203:2 (Jan. 1976). ISSN: ISSN ASJOA. DOI: 10.1086/154079. URL: https://www.osti.gov/biblio/7285770.

[16] Marko Shuntov et al. *COSMOS2025: The COSMOS-Web galaxy catalog of photometry, morphology, redshifts, and physical parameters from JWST, HST, and ground-based imaging*. 2025. arXiv: 2506.03243 [astro-ph.GA]. URL: https://arxiv.org/abs/2506.03243.

[17] J. Snigula et al. "The Munich Near-Infrared Cluster Survey – IV. Biases in the completeness of near-infrared imaging data". In: *Monthly Notices of the Royal Astronomical Society* 336.4 (Nov. 2002), pp. 1329–1341. ISSN: 0035-8711. DOI: 10.1046/j.1365-8711.2002.05869.x. eprint: https://academic.oup.com/mnras/article-pdf/336/4/1329/3046564/336-4-1329.pdf. URL: https://doi.org/10.1046/j.1365-8711.2002.05869.x.

A Details of Equation 4

We first write Equation 4 again:

$$f_{shallow} \sim \mathcal{N}(\mu = f_{\text{deep}},\, \sigma^2 = \left[\sigma_f\right]^2_{\text{shallow}} - \left[\sigma_f\right]^2_{\text{deep}}) \tag{16}$$

This means that we can express $f_{shallow}$ as:

$$f_{shallow} = f_{deep} + \text{'noise'} \tag{17}$$

This noise is given by $\sqrt{\left[\sigma_f\right]^2_{\text{shallow}} - \left[\sigma_f\right]^2_{\text{deep}}}$, i.e the standard deviation of the normal distribution. In order to find the flux errors, we start with the magnitude. By definition :

$$m = ZP - 2.5 \cdot log_{10}(f) \tag{18}$$

Hence, by error propagation (and because we assumed a high Signal-to-Noise Ratio):

$$\sigma_m = \frac{2.5}{ln(10)} \cdot \frac{\sigma_f}{f} = s \cdot \frac{\sigma_f}{f} = s \cdot NSR \tag{19}$$

Hence:

$$\sigma_f = \frac{f}{s} \cdot m \tag{20}$$

However, we also have (see Equation 2) :

$$NSR_{rand}{}^2 = (0.04 - \gamma)x + \gamma x^2 \tag{21}$$

With our definitions of γ and x, Equation 19 becomes:

$$\sigma_m = 0.2 \cdot s \cdot 10^{0.4 \cdot (m - m_{depth})} \tag{22}$$

Where m_{depth} corresponds to the 5σ magnitude depth, in the shallow or deep imaging. All in all, this gives us the following relation for the general error on the flux:

$$\sigma_f = 0.2 \cdot f \cdot 10^{0.4 \cdot (m - m_{depth})} \tag{23}$$

As we can see, for a given flux (and hence magnitude), it only depends on the magnitude depth. Thus, by considering the 'deep' flux as the *true* flux, we can write:

$$\left[\sigma_f\right]_{shallow} = 0.2 \cdot f_{deep} \cdot 10^{0.4 \cdot \left(m_{deep} - m_{depth_{shallow}}\right)} \tag{24}$$

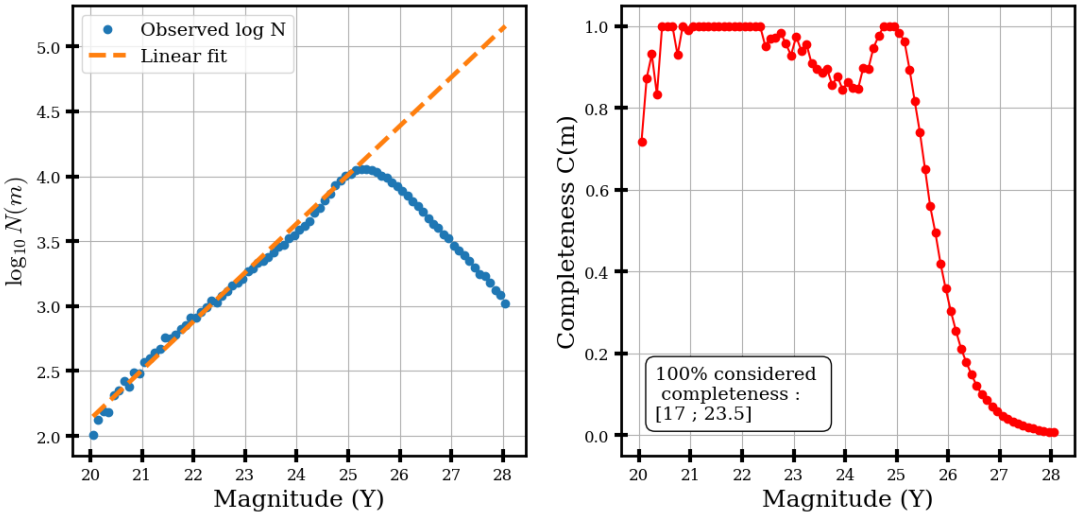
The deep flux error is already given in the COSMOS catalogue and we will use this value. We could however also calculate it using the same method, only replacing $m_{depth_{shallow}}$ by $m_{depth_{deep}}$. This gives us everything we need to find the standard deviation of the 'shallow' flux's distribution. We can then simulate the latter as explained in Equation 17, where the noise is given by a random draw following the normal distribution $\mathcal{N}(\mu = 0,\, \sigma^2 = \left[\sigma_f\right]^2_{\text{shallow}} - \left[\sigma_f\right]^2_{\text{deep}})$.
Note that the same expression for the standard deviation can be found using only considerations on the magnitude and flux errors, added on quadrature. From:

$$\left[\sigma_m\right]_{\text{shallow}} = \sqrt{\left[\sigma_m\right]^2_{\text{deep}} + \Delta(\sigma_m)^2} \tag{25}$$

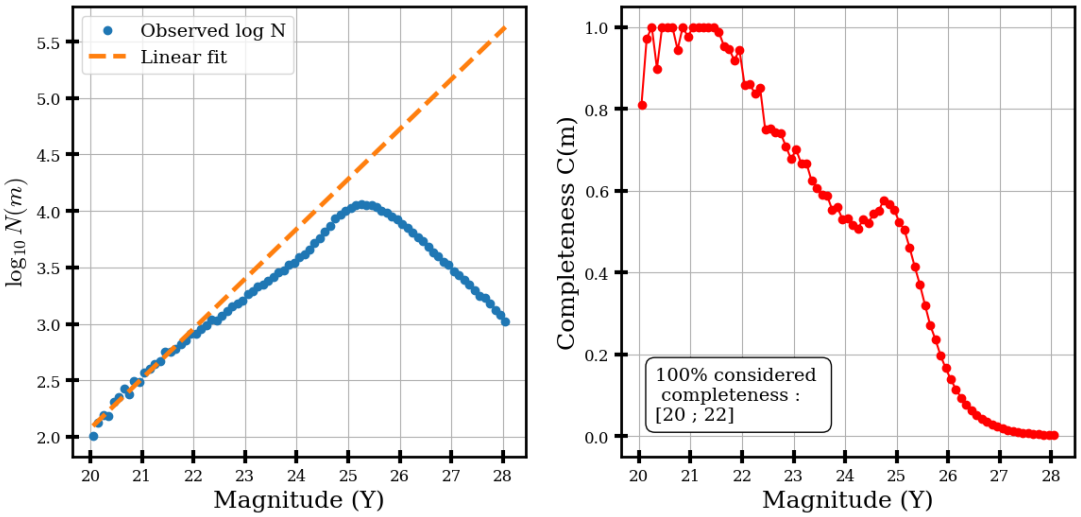
We can isolate $\Delta(\sigma_m)$, and find the difference between the two flux errors using Equation 20.

B Different 100% completeness intervals for the detection probability function

See the different graphs and diagrams for the four 100% completeness intervals we tested. As we can see, the results differ depending on the initial conditions, mostly if we stop too soon, where we lose a lot of targets. Table 5 and Table 6 summarise the results with our different intervals, containing the percentage of detected sources and the fit-parameters' values.

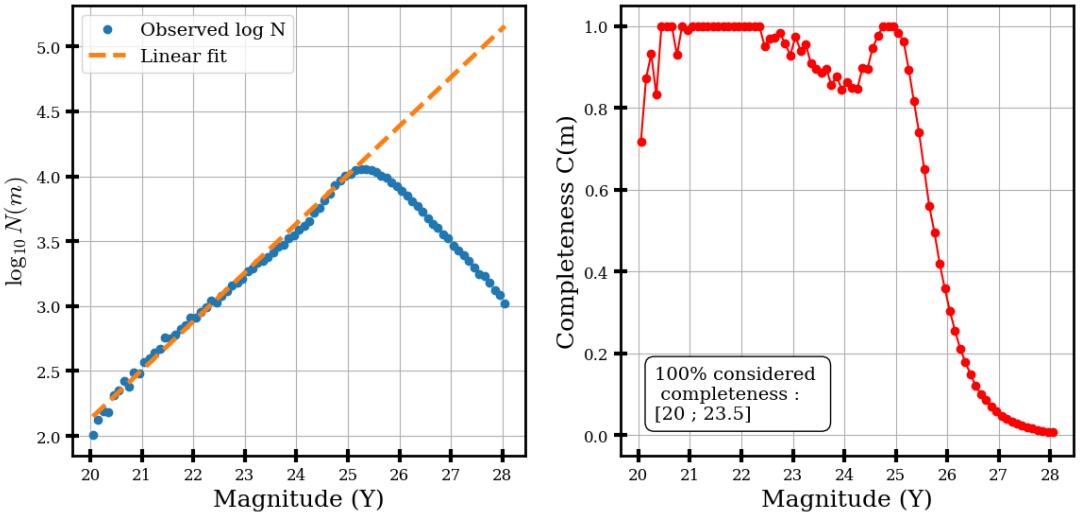


(a) *Left* : log_{10}(N) as a function of the magnitude, in blue, and the linear regression in orange. We see the linearity up to almost 25 mag. *Right* : completeness given by the left figure. Here, the "100% completeness" interval was [17 - 23.5].

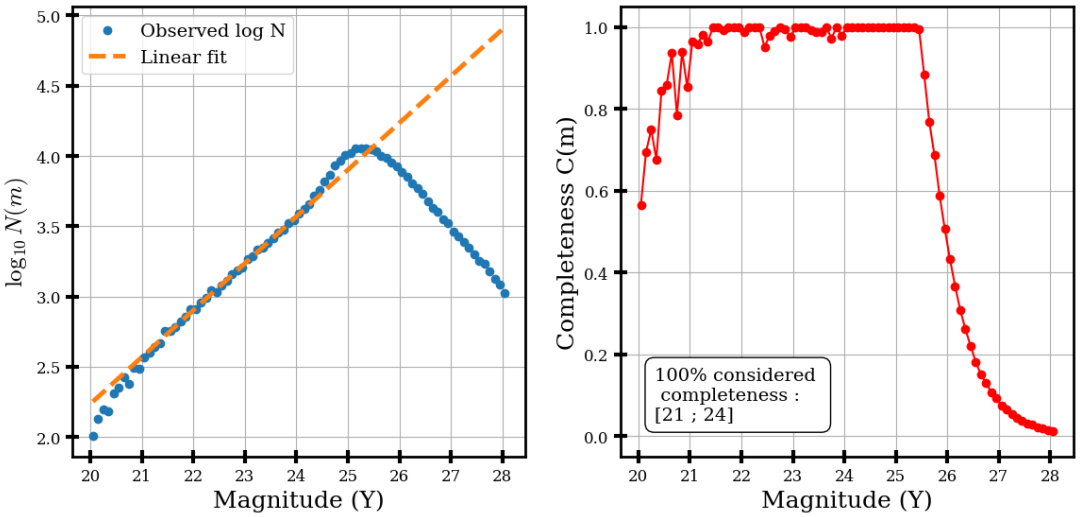


(b) *Left* : log_{10}(N) as a function of the magnitude, in blue, and the linear regression in orange. We see the linearity up to almost 25 mag, though the fit is not as good here. *Right* : completeness given by the left figure. Here, the "100% completeness" interval was [20 - 22].

Figure 14: Comparison of log_{10}(N) and completeness for different "100% completeness" intervals (part 1).

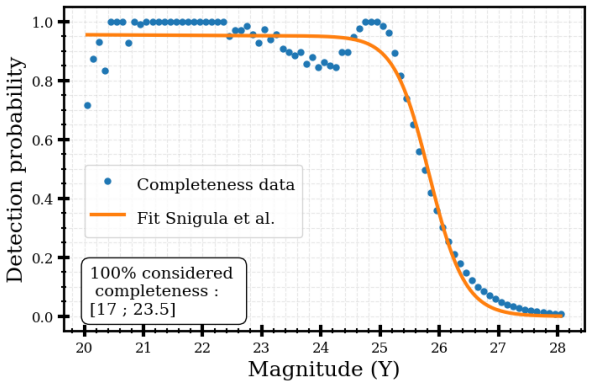


(a) *Left* : log_{10}(N) as a function of the magnitude, in blue, and the linear regression in orange. We see the linearity up to almost 25 mag. *Right* : completeness given by the left figure. Here, the "100% completeness" interval was [20 - 23.5].

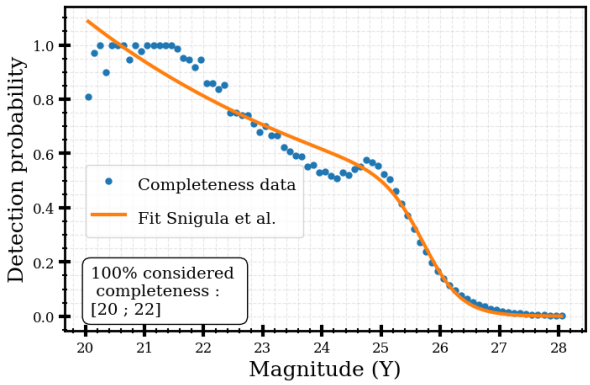


(b) *Left* : log_{10}(N) as a function of the magnitude, in blue, and the linear regression in orange. We see the linearity up to almost 25 mag. *Right* : completeness given by the left figure. Here, the "100% completeness" interval was [21 - 24].

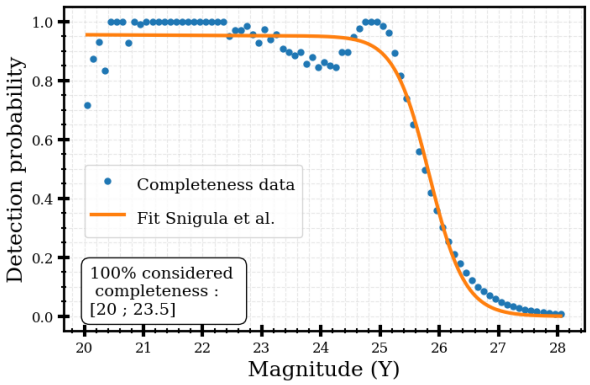
Figure 15: Comparison of log_{10}(N) and completeness for different "100% completeness" intervals (part 2).



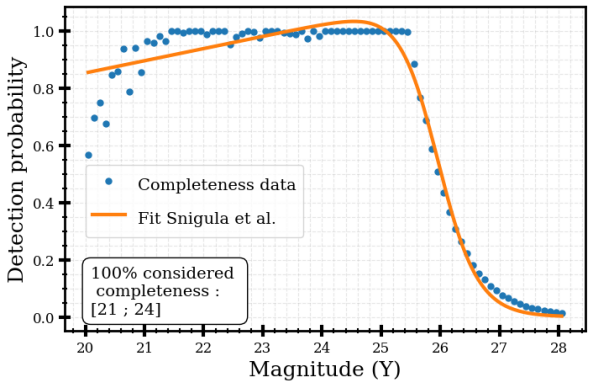
(a) Best-fit of [Snigula et al. [17]] with [17–23.5] interval.



(b) Best-fit of [Snigula et al. [17]] with [20–22] interval.



(c) Best-fit of [Snigula et al. [17]] with [20–23.5] interval.



(d) Best-fit of [Snigula et al. [17]] with [21–24] interval.

Figure 16: Comparison of best-fits of [Snigula et al. [17]]'s function for different completeness intervals.

100% completeness magnitude (mag)	**Percentage of detected sources**
17 - 23.5	48.51%
20 - 22	29.18%
20 - 23.5	48.34%
21 - 24	55.21%

Table 5: Detection percentages according to the different initial conditions.

100% completeness magnitude (mag)	50% limiting magnitude (m_0)	p_0	a	b
17 - 23.5	25.84	0.95	0.036	87.25
20 - 22	25.77	0.50	3.09	83.66
20 - 23.5	25.84	0.95	0.036	87.25
21 - 24	25.95	1.11	-1.00	77.04

Table 6: Fit parameters for [Snigula et al. [17]]'s function.

C Redshift selection with Y-, J- and H-bands only

Here beneath, we show the different diagrams and results we found using only Y-, J- and H-bands to filter the redshifts. We justify the use of external CH1 and CH2 bands, from IRAC.
Figure 17 shows the main three colour-colour diagrams using only Y-, J- and H-bands. Even though some gradients can be seen, it is not absolute and there is always a patch of low-redshift objects going through the entire diagram. As we will see, this prevents us from filtering redshifts using these diagrams only.

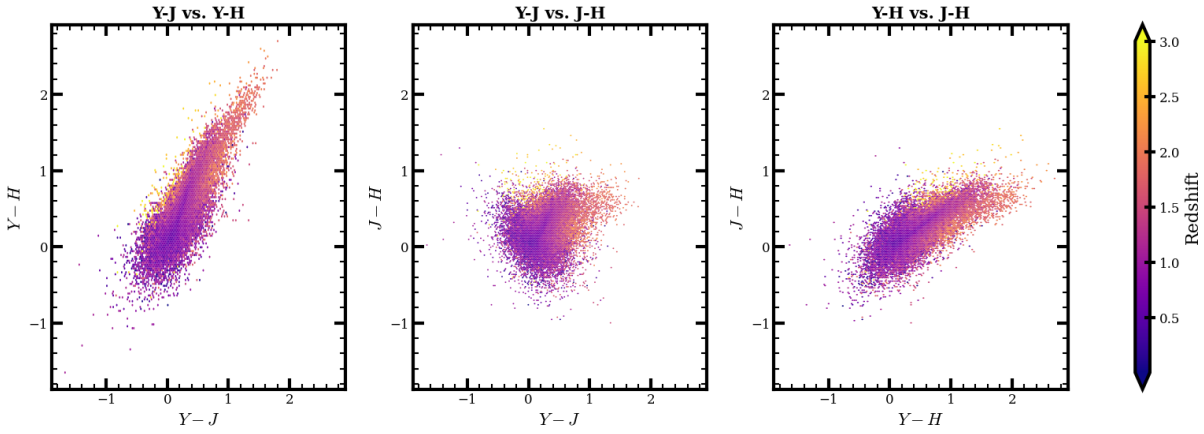


Figure 17: Major diagrams using the Y-, J- and H-bands. *Left*: Y-J vs Y-H; *center*: Y-J vs J-H; *right*: Y-H vs J-H.

To save space, we will only show the results coming from the Y-J vs Y-H diagram (the left one on Figure 17), but they are similar with the other ones. We follow the same method as described in Section 5.1.
First, we produce the 2-dimensional histograms of the redshift distribution, with respect to the two colours (see Figure 18. We then find the best Gaussian fit to describe this distribution (in grey ellipses, on Figure 18), and put these fits on the same diagram, hoping to finding a redshift-cut. However, as shown on Figure 19, the resulting diagram is not very helpful. All the Gaussians overlap, not allowing us to properly filter the redshifts. The same happens with the other diagrams using only Y-, J- and H-bands, leaving us no choice but to use other filters, in our case CH1 and CH2.

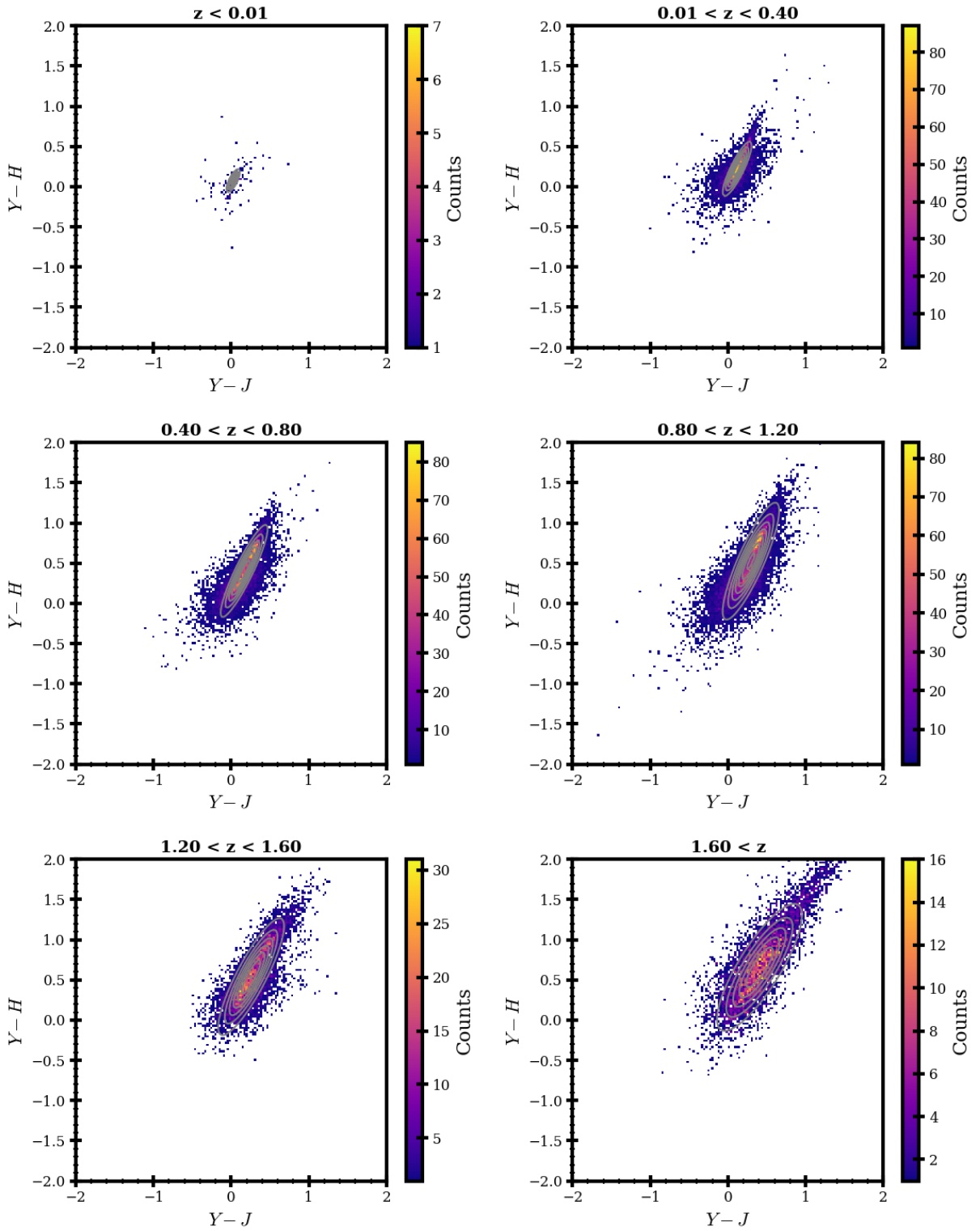


Figure 18: 2-dimensional histograms of the redshift distribution according to the Y-J and J-H colours. In grey ellipses, there is the best-fit of the 2-dimensional Gaussians.

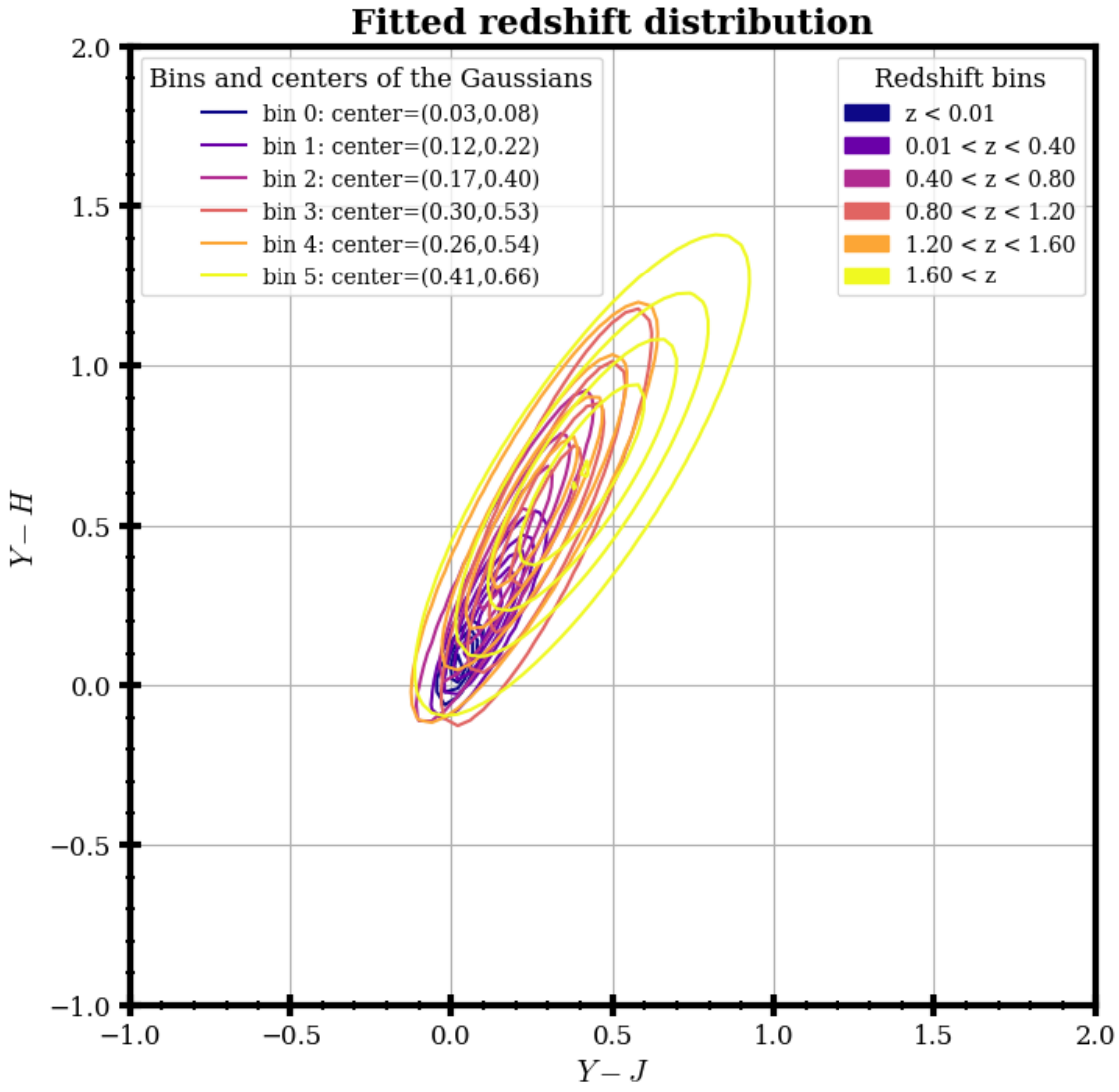


Figure 19: Redshift distribution map from the Y-J vs Y-H diagram.