




## Article

# ALMA Band 3 Source Counts: A Machine Learning Approach to Contamination Mitigation below 5 Sigma

Ivano Baronchelli <sup>1,2,\*</sup> , Matteo Bonato <sup>1,2</sup> , Gianfranco De Zotti <sup>3,4</sup> , Viviana Casasola <sup>1</sup> , Michele Delli Veneri <sup>5</sup> , Fabrizia Guglielmetti <sup>6</sup> , Elisabetta Liuzzo <sup>1,2</sup> , Rosita Paladino <sup>1,2</sup> , Leonardo Trobbiani <sup>1,7</sup> and Martin Zwaan <sup>6</sup> 

<sup>1</sup> INAF—Istituto di Radioastronomia, Via Gobetti 101, I-40129 Bologna, Italy

<sup>2</sup> Italian ALMA Regional Centre, Via Gobetti 101, I-40129 Bologna, Italy

<sup>3</sup> INAF—Osservatorio Astronomico di Padova, Vicolo dell'Osservatorio 5, I-45122 Padova, Italy

<sup>4</sup> Nicolaus Copernicus Astronomical Center of the Polish Academy of Sciences, ul. Bartycka 1800-716 Warszawa, Poland

<sup>5</sup> INFN—Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, Via Cintia, 1, 80126 Napoli, Italy

<sup>6</sup> ESO, Karl-Schwarzschild-Straße 2, 85748 Garching bei München, Germany

<sup>7</sup> Dipartimento di Fisica e Astronomia “Augusto Righi”, Università di Bologna, Viale Carlo Berti Pichat, 6/2, 40127 Bologna, Italy

\* Correspondence: ivano.baronchelli@inaf.it

**Abstract:** We performed differential number counts down to 4.25 sigma using ALMA Band 3 calibrator images, which are known for their high dynamic range and susceptibility to various types of contamination. Estimating the fraction of contaminants is an intricate process due to correlated non-Gaussian noise, and it is often compounded by the presence of false positives generated during the cleaning phase. In addition, calibrator extensions further complicate the counting of background sources. In order to address these challenges, our strategy employs a machine learning-based approach utilizing the UMLAUT algorithm. UMLAUT assigns a value to each detection, and it considers how likely it is for there to be a genuine background source or a contaminant. With respect to this goal, we provide UMLAUT with eight observational input parameters, each automatically weighted using a gradient descent method. Our methodology significantly improves the precision of differential number counts, thus surpassing conventional techniques, including visual inspection. This study contributes to a better understanding of radio sources, particularly in the challenging sub-5 sigma regime, within the complex context of a high dynamic range of ALMA calibrator images.

**Keywords:** galaxies: photometry; galaxies: active; galaxies: abundances; sub-millimeter: galaxies



**Citation:** Baronchelli, I.; Bonato, M.; De Zotti, G.; Casasola, V.; Delli Veneri, M.; Guglielmetti, F.; Liuzzo, E.; Paladino, R.; Trobbiani, L.; Zwaan, M. ALMA Band 3 Source Counts: A Machine Learning Approach to Contamination Mitigation below 5 Sigma. *Galaxies* **2024**, *12*, 26. <https://doi.org/10.3390/galaxies12030026>

Academic Editor: Oleg Malkov

Received: 29 February 2024

Revised: 30 April 2024

Accepted: 2 May 2024

Published: 20 May 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The high angular resolution of the Atacama Large millimeter/sub-millimeter Array (ALMA) has made it possible to overcome the confusion limits that affect millimeter wave surveys, even those with single-dish telescopes of the 6–10 m class, such as the Atacama Cosmology Telescope (ACT; [1]) and the South Pole telescope (SPT; [2,3]).

Furthermore, high ALMA sensitivity allows us to detect much fainter sources, down to  $\mu\text{Jy}$  flux density levels, where number counts are dominated by dusty, star-forming galaxies (DSFGs) and are extremely steep due to the combined effect of cosmological evolution and the strongly negative K-correction. On the other hand, the ALMA field of view is very small. Its diameter, usually taken as the full width at half maximum (FWHM) of the primary beam, is  $19''/(300/\nu)$  for a 12 m antenna and  $33''/(300/\nu)$  for a 7 m antenna, with  $\nu$  expressed in GHz. This implies that only very small areas can be surveyed in a reasonable amount of time.

To achieve reliable statistics, it is crucial to minimize the detection limit as much as possible. However, dropping below a signal-to-noise ratio (SNR) of around five introduces

several complications: both the catalog completeness and the reliability of detection rapidly decrease. At the same time, the sources detected near the detection threshold are more susceptible to the Eddington bias [4], thereby resulting in overestimated flux densities (i.e., “flux boosting”). This effect is naturally intensified when, because of a low detection limit, noise fluctuations significantly influence the overall flux of the weakest detections in the sample, thereby determining their inclusion or exclusion from the sample itself. This bias is further amplified if the intrinsic counts are particularly steep [5], as is the case for DSFGs.

Thus far, blind ALMA surveys [6–10] have covered only tiny fractions of a square degree. Searches of serendipitous sources have also been carried out using archive data, where the proximity of primary ALMA targets were explored. Indeed, extragalactic surveys using ALMA can be particularly fruitful when analyzing the data from fields that surround the sources observed for other purposes [11–13]. However, once again, such surveys cover very small overall areas. The largest surveys of the latter type have been made possible in the context of the ALMACAL project [14]. These ‘calibrator’ fields, while primarily intended for instrument calibration, offer the valuable opportunity to serendipitously detect faint, unexpected sources.

In ALMACAL, public calibrator data are retrieved from the ALMA archive, where they are then calibrated and imaged. The compatible data of the same calibrator are also combined to potentially enhance the signal-to-noise ratio of the background sources. Calibrator observations represent a significant fraction of each ALMA scheduling block, and several hundreds of calibrators have been repeatedly observed in different bands during the lifetime of ALMA. The total area covered by such observations amounts to a substantial fraction of a square degree. However, the typically high dynamic range of these images (i.e., the SNR of the brightest source) represents a limitation of these data. In fact, these characteristics are commonly associated with the appearance of peculiar noise patterns, the non-Gaussianity of the noise distribution, and false positives, as well as other problems.

We present a novel machine learning approach to address the aforementioned challenges, and the technique is applied to derive extragalactic number counts at 100 GHz (ALMA band 3).

In Section 2, we provide details on how we selected our sample of ALMACAL images (Section 2.1); we describe the source extraction procedure adopted (Section 2.2); the techniques we used to correct for incompleteness (Section 2.3); the contamination that arises from spurious detection (Section 2.4); flux boosting (Section 2.5); and the contamination effects that occur due to other unwanted detections (Section 2.6). The steps toward the derivation of number counts are presented and discussed in Section 3. The main conclusions are summarized in Section 4.

## 2. Data Analysis

### 2.1. Sample and Image Selection

We selected fields observed in band 3 from the latest version of the ALMACAL sample (see [15]). To minimize contamination by galactic sources, we excluded the calibrators at a galactic latitude of  $|b| < 10^\circ$ .

The measurement of the flux of a calibrator is not affected greatly by the nature of the underlying noise. This is because the calibrator’s flux is significantly higher (hundreds of sigmas) than the noise level. Anomalous noise patterns, deviations from Gaussian distribution, and even poor primary beam corrections do not significantly affect the flux estimate of the calibrator at the center of the images. However, when it comes to identifying and measuring the flux of the serendipitous detection in the background, these same issues become a serious challenge.

Some of the peculiar noise patterns described, which cause noise that deviates from the random theoretical distribution, stem from a poor coverage of the uv plane and can introduce significant uncertainties in the flux estimates. Specifically, a phenomenon known

as “clean bias”, which was first observed in the context of VLA surveys [16–18], may arise during the cleaning process applied to images with inadequate uv plane coverage (typically “snapshots”—defined as observations that are too brief for earth rotation to complete the uv coverage). This “clean bias” leads to changes in object fluxes and the apparent levels of image noise.

For the reasons explained above, not all of the images in the ALMACAL sample were suitable for computing number counts. In order to prevent possible biases, we selected a sub-sample of the original complete sample of the calibrator images characterized by Gaussian noise distribution and by the absence of peculiar noise patterns.

To analyze each image in our original complete sample, we automatically calculated the histogram of fluxes measured for all the pixels within the image. We used the following formula to determine the width of each bin:

$$\Delta F = [F_{\text{pix}}(97) - F_{\text{pix}}(3)]/25, \quad (1)$$

where  $F_{\text{pix}}(3)$  and  $F_{\text{pix}}(97)$  represent the 3rd and the 97th percentiles of the pixel fluxes, respectively. All the fluxes considered for this analysis were not primary beam-corrected.

Next, for each bin  $i$ , we calculated the difference between the distribution  $N(F_i)$  (normalized to have a peak of 1) and the value of the best-fitting Gaussian function at the central flux of the same bin  $G(F_i)$  as follows:

$$\Delta \text{bin}(F_i) = N(F_i) - G(F_i) = N_i - G_i. \quad (2)$$

We used the standard deviation of these values as follows:

$$\sigma_{\text{pix}} = \sqrt{\frac{\sum_i (N_i - G_i)^2}{N_{\text{bins}}}}. \quad (3)$$

This was used as an indicator of the overall deviation from the Gaussian fit. After examining the noise patterns in multiple images and the distribution of pixel fluxes, we found a good compromise between the number of eliminated images and the level of distortion from a Gaussian distribution by setting the threshold for acceptance to  $\sigma_{\text{pix}}^{\text{max}} = 0.015$ .

To evaluate the uniformity of the RMS across the field, we compared the best-fitting Gaussian curves obtained from the flux distribution of the innermost and outer pixels of each of the images considered. In this analysis, we considered images with absolute deviations ( $\text{RMS}_{\text{in}} - \text{RMS}_{\text{out}}$ ) exceeding 10% of the overall RMS as problematic and rejected them.

We performed another image quality analysis by randomly placing a certain number of beam-size apertures throughout the image and calculating the distribution of fluxes measured within these apertures. Unlike the procedure previously explained, where the number of pixels (measurements in the histogram) was fixed, in this case, the number of measurements in the distribution depended on the size of the apertures, which, in turn, was determined by the beam size (i.e., by the PSF) of each image.

For each bin  $j$ , we compared the observed distribution with the best-fitting Gaussian curve by measuring the absolute difference in terms of the area under the two curves. This difference, divided by the total area under the Gaussian fit curve ( $A_{\text{Gauss fit}}$ ), represents the fractional deviation of the observed distribution from the Gaussian fit as follows:

$$GD_j = |A_{\text{aper}}^j / A_{\text{Gauss fit}}^j| / A_{\text{Gauss fit}}. \quad (4)$$

We established a threshold for acceptance at  $GD = \sum_j GD_j = 0.2$ , which implies that the difference between the histogram of random aperture fluxes and the Gaussian fit should not exceed 20% of the total area under the histogram.

Still, the methods described above did not consider asymmetrical deviations from Gaussianity, where positive fluxes may differ from the Gaussian fit differently from negative

fluxes. To address this, we implemented an additional criterion. We rejected images if the absolute difference between the left deviation  $GD_{\text{left}}$  (negative fluxes) and the right deviation  $GD_{\text{right}}$  (positive fluxes) exceeded 40% of the overall deviation from Gaussianity as follows:

$$SD = |GD_{\text{left}} - GD_{\text{right}}| / GD. \quad (5)$$

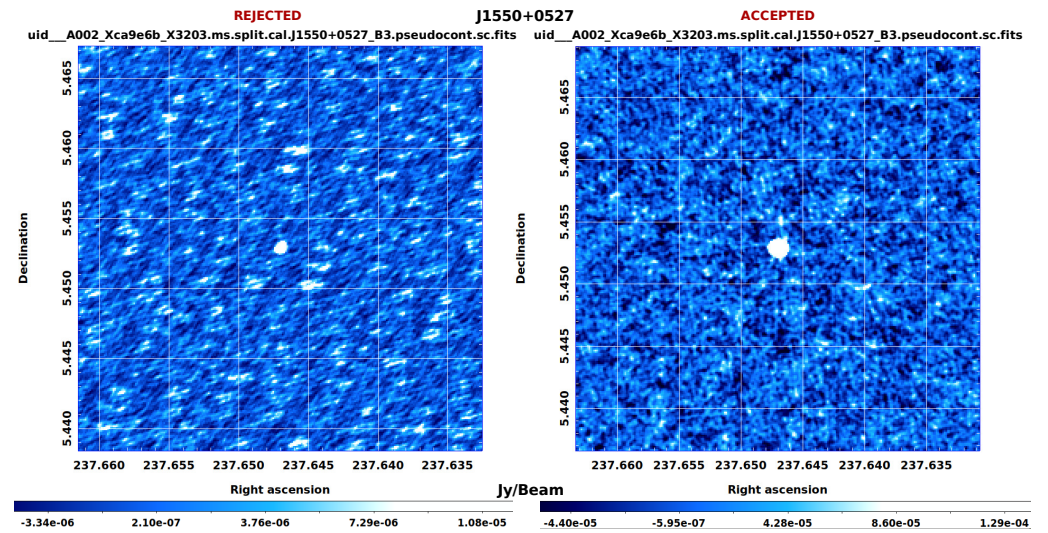
This criterion allowed us to exclude the images characterized by significant asymmetrical distributions of their beam aperture fluxes.

As a last requirement, we excluded images with poor angular resolution from our analysis, namely those with a FWHM larger than 5 arcsec of the synthesized beam.

In Figure 1, we show two different images of the same calibrator (J1550 + 0527), one of which was rejected (left panel), while the other was accepted (right panel). We selected these two examples to clearly show the impact of different noise patterns on the same subject/field. The non-random behavior of the noisy background is clearly visible in the rejected image. For the particular example shown in the figure, we measured the following rejection parameters (listed in the order: rejected image, accepted image, and threshold):

- Galactic latitude ( $b$ ):  $+42.2^\circ$ ,  $+42.2^\circ$ , and  $<\pm 10.0^\circ$ ;
- Gaussianity deviation 1 ( $\sigma_{\text{pix}}$ ): **0.052**, 0.009, and  $<0.015$ ;
- Homogeneity deviation ( $HD$ ): 0.011, 0.08, and  $<0.10$ ;
- Gaussianity deviation 2 ( $GD$ ): **0.25**, 0.13, and  $<0.20$ ;
- Symmetry deviation ( $SD$ ): 0.047, 0.067, and  $<0.40$ ;
- FWHM:  $1.5''$ ,  $1.7''$ , and  $<5.0''$ .

We clearly marked (in bold) which parameters caused the first image to be rejected. We also rejected images even if just one of parameters fell outside the acceptable range (either above or below the limit).



**Figure 1.** Band 3 images of calibrator J1550 + 0527. In the (left panel) is an example of an image excluded from our sample due to surpassing certain thresholds specified for image rejection (the specific values of Galactic latitude,  $\sigma_{\text{pix}}$ ,  $HD$ ,  $GD$ ,  $SD$ , and FWHM are reported in the text for both the images). In the (right panel) is the image included in our sample for the same calibrator. The general shape of the noise patterns was not directly related to any of the selection parameters in particular. However, the same parameters were related to the probability to observe the non-random patterns. The scientific notation “ $aEb$ ” indicates  $a \times 10^b$ . For example,  $1.5E8$  stands for  $1.5 \times 10^8$ .

By applying the criteria mentioned above to the entire sample, we were able to accept multiple images of the same calibrators, all of which met the specified requirements, in many cases. In these cases, in order to select the most suitable image for each calibrator, we opted for the one with the smallest RMS noise, i.e., the deepest (not the most extended) one.

On the other hand, in some instances, none of the available images of a particular calibrator met all the criteria mentioned above. Consequently, 218 of the calibrators out of 824 observed in band 3 had to be rejected. Repeating the analysis for the other bands, we found that our criteria reject 77 out of 345 calibrators in band 4, 31 out of 138 in band 5, 179 out of 788 in band 6, 113 out of 491 in band 7, 39 out of 148 in band 8, 20 out of 75 in band 9, and 2 out of 16 in band 10. The complete list of parameters used for the image rejection phase is listed in Table 1.

**Table 1.** Image rejection criteria <sup>[a]</sup><sub>[b]</sub>.

Parameter Tested	Equation	Threshold for Acceptance
Galactic latitude	-	$ b  < 10^\circ$
Gaussianity deviation 1 <sup>[c]</sup>	$\sigma_{\text{pix}} = \sqrt{\frac{\sum_i (N_i - G_i)^2}{N_{\text{bins}}}}$	$\sigma_{\text{pix}} < 0.015$
Homogeneity deviation <sup>[c]</sup>	$HD =  \sigma_{\text{pix}}^{\text{inner}} - \sigma_{\text{pix}}^{\text{outer}}  / \sigma_{\text{pix}}$	$HD < 0.1$
Gaussianity deviation 2 <sup>[d]</sup>	$GD = \sum_j  A_{\text{aper}}^j - A_{\text{Gauss fit}}^j  / A_{\text{Gauss fit}}$	$GD < 0.2$
Simmetry deviation <sup>[d]</sup>	$SD =  GD_{\text{left}} - GD_{\text{right}}  / GD$	$SD < 0.4$
FWHM		FWHM of <5 arcsec

<sup>[a]</sup> Rejecting calibrator images does not necessarily mean eliminating the corresponding calibrators from our sample. Other images of the same calibrator field may have been accepted following the same criteria. <sup>[b]</sup> The definitions of the parameters mentioned in the equations are given in Section 2.1. <sup>[c]</sup> Computed over the histogram of the pixel fluxes. <sup>[d]</sup> Computed over the histogram of the beam-size aperture fluxes.

## 2.2. Source Extraction

We employed SExtractor [19] for source detection and flux measurements. While originally not intended for radio interferometry, SExtractor has been employed in processing CLEANed images since its initial utilization by [20]. Despite its origins in optical astronomy, its speed and ease of use have also contributed to its adoption in radio astronomy. In [21,22], various algorithms for source extraction were compared when applied to radio images, where minimal differences were found among them (in terms of contamination, completeness, and measured fluxes). This was especially the case when operating on maps that were characterized by being close to theoretical noise conditions (particularly those with a Gaussian noise distribution). Regarding this final aspect, it is worth noting (see Section 2.1) that we deliberately selected a subset of images with a Gaussian noise distribution and no discernible peculiar noise patterns.

Since each calibrator image was characterized by distinct beam size and shape, we allowed our software to automatically determine the most suitable SExtractor parameters. The essential parameters used are summarized in Table 2 (the employed default SExtractor options are not listed).

To perform the initial filtering of the image, we applied a Gaussian filter with a size closely matching two times the FWHM of the specific image's beam. The parameters of BACK\_SIZE and BACK\_FILTERSIZE, which control the filtering intensity during background computation, were set to four and two times the FWHM, respectively.

To consider a source as detected, a specific number of adjacent pixels (DETECT\_MINAREA) must exceed a certain threshold (DETECT\_THRESH). We set this parameter to a value comparable to the beam area:

$$\text{DETECT\_MINAREA} = 2\pi \left( \frac{\text{FWHM}_a}{2.355} \right) \left( \frac{\text{FWHM}_b}{2.355} \right), \quad (6)$$

where  $\text{FWHM}_a$  and  $\text{FWHM}_b$ , expressed in units of [pixels], represent the FWHM of the point spread function (PSF) along the major and minor axis, respectively. We set the parameters DETECT\_THRESH and ANALYSIS\_THRESH at  $1.2\sigma$ .



**Table 2.** Source extraction: the non-default SExtractor parameters used <sup>[a]</sup>.

SExtractor Parameter	Value	Units
Filter	“Gauss” with size $\sim 2 \times \text{FWHM}$	Pixels
BACK_SIZE	$4 \times \text{FWHM}$	Pixels
BACK_FILTERSIZE	2	/
DETECT_MINAREA	$2\pi \left( \frac{\text{FWHM}_a}{2.355} \right) \left( \frac{\text{FWHM}_b}{2.355} \right)$	Pixels
DETECT_THRESH	1.2	$\sigma$
ANALYSIS_THRESH	1.2	$\sigma$
PHOT_AUTOPARAMS (KRON_FACT) <sup>[b]</sup>	2.5	A_IMAGE
PHOT_AUTOPARAMS (rKRON_min) <sup>[c]</sup>	3.5	Pixels

<sup>[a]</sup> Not all the sources extracted were kept in the final catalog. After the SExtractor run, we performed a source selection (see Section 2.2). <sup>[b]</sup> KRON\_FACT represents the size of the elliptical aperture used for “AUTO” photometry in the units of A\_IMAGE and B\_IMAGE, i.e., the semimajor and semiminor axes of the elliptical representation of the source identified. <sup>[c]</sup> rKRON\_min represents, in pixels, the minimum automatic aperture used.

We limited the search for serendipitous sources to the region outside 4 arcsec and within 45.5 arcsec from the central calibrator. The inner region was likely affected by emissions related to the calibrator. The outer region required too heavy ( $\geq 78\%$ ) primary beam corrections.

To estimate the total flux of the sources, we relied on the FLUX\_AUTO parameter. This parameter helps to provide a flux measurement within an elliptical aperture centered on the source. The semi-major axis of the aperture is determined by multiplying the Kron radius [23] by a certain factor (in our specific case, KRON\_FACT = 2.5). In SExtractor, the Kron radius is independently estimated along the two axes of the elliptical representation of the source identified. These two estimates can be outputted as A\_IMAGE and B\_IMAGE and are measured in pixels. FLUX\_AUTO is specifically designed to capture a significant portion of the emitted flux from the source while minimizing any contamination from the neighboring objects or background noise. As reported in the SExtractor manual, for extremely noisy objects, the Kron ellipse may sometimes become too small (even smaller than the isophotal footprint of the object). To address this issue, SExtractor enforces a minimum size for the Kron radius, thereby ensuring it cannot be less than rKron\_min. We set this parameter to 3.5 pixels and verified that, with only one exception, the size of the automatic apertures used is always larger than the beam size.

To assess the significance level of each detection, we employed the method of random apertures to calculate the typical RMS fluctuations of each image. This method involves randomly positioning circular apertures throughout the image and measuring the distribution of the fluxes within them. The RMS was determined by computing the standard deviation ( $\sigma$ ) of the Gaussian curve that best fitted the flux distribution. In each image, the size of these random apertures was chosen to be comparable to the beam size. We also verified that these apertures encompassed more than 90% of the flux emitted by the calibrators.

The fluxes in the pixels of the images were originally expressed in units of [Jy/beam], and they were not primary beam corrected. Therefore, besides applying the primary beam correction, in order to obtain total fluxes in units of [Jy], we need to divide the fluxes computed by SExtractor by the beam area ( $A_{\text{BEAM}}$ ) expressed in [pixels]<sup>1</sup> as follows:

$$A_{\text{BEAM}} = \frac{\pi}{4 \ln 2} \times \text{FWHM}_a \times \text{FWHM}_b, \quad (7)$$

with  $\text{FWHM}_a$  and  $\text{FWHM}_b$  in units of [pixels]. In other words, before integrating over all the pixels inside the apertures, we converted the flux in each pixel as follows:

$$F_{\text{pix}}[\text{Jy/pixel}] = F_{\text{pix}}[\text{Jy/beam}] / A_{\text{BEAM}}[\text{pixels/beam}]. \quad (8)$$

All around the rest of the paper, we refer to aperture-like (or “aper”) fluxes, in all cases in which fluxes are computed inside beam-size apertures, while “total” fluxes or fluxes “AUTO” refer to the actual total fluxes of the sources. Using aperture fluxes is needed in order to assess the significance level of the detected sources, as both the RMS and the image flux must be estimated using apertures with similar sizes. Instead, the total (AUTO) fluxes are computed inside the automatically estimated variable apertures, and they represent the actual total flux of the sources considered.

### 2.3. Completeness

Because of the stochastic nature of noise, not all sources within a field are detected. If we consider a set of sources with expected fluxes precisely matching the detection threshold, and under the assumption of a symmetrical noise distribution, we should expect that half of these sources will be suppressed by the noise, thus making them undetectable. Conversely, the remaining half will experience flux enhancement due to noise, resulting in their detectability. The fraction of sources detected at various flux levels is contingent on the noise distribution. A completeness function can be calculated through numerical simulations.

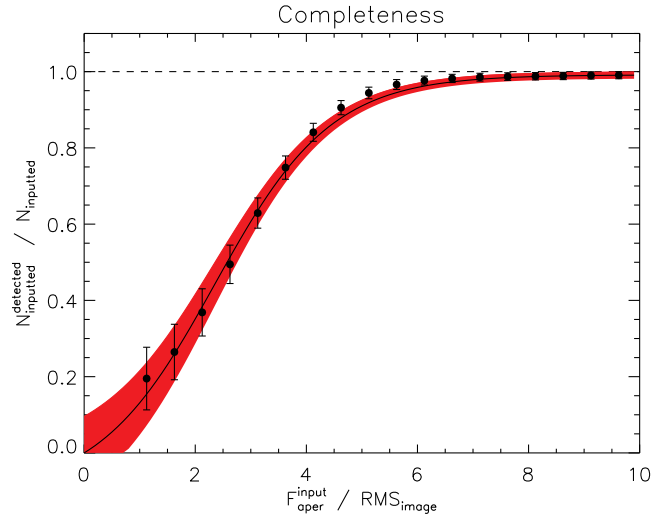
To calculate the completeness correction, we followed a specific procedure. Firstly, we inverted the flux of the original images. Then, we introduced simulated two-dimensional Gaussian sources into the images. These simulated sources had semi-axes similar to those of the point spread function (PSF) and the same position angle. The simulated sources were injected at various significance levels. Next, we applied a source detection procedure identical to the one described in Section 2.2. We considered a source as “recovered” if it was identified above a minimum detection threshold and located within a search radius corresponding to half the size of the synthesized beam from the input position.

The initial completeness curve  $C'(SNR)$ , dependent on the input signal-to-noise ratio (SNR) (i.e., the ratio between the flux of the simulated sources and RMS of the image considered), was obtained by fitting our data with the rational function as follows:

$$C'_i(SNR) = \frac{1}{a_0 a_1^{SNR} + a_2}. \quad (9)$$

The best-fitting parameters that we found were  $a_0 = 9.44$ ,  $a_1 = 0.39$ , and  $a_2 = 1.00$ . Due to the presence of contaminants (see Section 2.4), a certain fraction of injected sources were erroneously considered as recovered even when they were not. This occurred because the contaminants fell within the search circle, thus leading to false detections. Consequently, the original curve did not approach zero at low significance levels (i.e., less than  $3\sigma$ ) as it should have. To address this bias, we subtracted the bias itself ( $C_{\text{bias}} = 0.0957$ , i.e., the value of the best-fitting curve interpolated at  $x = 0$ ), from the best-fitting curve and renormalized it, thereby ensuring that it asymptotically reached the maximum value of completeness observed ( $C = 1$ ). By doing so, we corrected for the effect of contaminants and obtained a more accurate completeness curve. To quantify the associated uncertainty, we measured the difference between the uncorrected data points and the corrected curve. The total uncertainty was calculated as the quadratic sum of this uncertainty and the relatively less significant Poissonian uncertainty.

The corrected data points and completeness curve, together with their associated uncertainties, are shown in Figure 2. Here, the completeness was represented by the ratio between the number of sources injected in our simulations and subsequently detected ( $N_{\text{input}}^{\text{detected}}$ ), to the total number of simulated sources actually injected into the simulated images ( $N_{\text{input}}$ ). This ratio is computed as a function of the significance of the injected sources, i.e., the ratio between the flux of the injected sources computed inside the beam-size apertures ( $F_{\text{aper}}^{\text{input}}$ ) and the RMS of the simulated images (i.e., the  $\text{RMS}_{\text{image}}$ , which is measured with the method of the random apertures explained in Section 2.2, using beam-size apertures).



**Figure 2.** The completeness measurements (the black-filled circles and error bars) and best-fitting curve (the black line and red-filled area). Both the measurements and the fitting curve were corrected for the bias due to the fraction of the spurious sources misidentified as injected sources.

#### 2.4. Spurious Detections

Not all detections represent genuine sources. For instance, assuming a perfect Gaussian noise distribution, roughly 16% of the pixels will exhibit fluxes larger than  $1\sigma$ , 2.3% surpassing  $2\sigma$ , and merely 0.13% exceeding  $3\sigma$ . These seemingly small percentages should not instill confidence in the reader, especially when the goal is to detect rare objects across extensive areas. In a 256-by-256 pixel image, we would anticipate around 10,500, 1500, and 850 pixels exceeding the  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$  thresholds, respectively. Conversely, in a sample of hundreds of similar images, we would only expect to detect a few dozen genuine serendipitous sources. In essence, the likelihood that a single pixel exceeding a  $3\sigma$  threshold represents a false detection is considerably higher than it being a real source.

The fraction of spurious detections can be reduced using several strategies. In our analysis (see Section 2.2), we considered only the sources characterized by a minimum number of adjacent pixels higher than a certain threshold as detected. While sources typically span multiple pixels in the image (the point spread function of point-like sources is typically larger than just a few pixels), the likelihood of two randomly selected adjacent pixels both exceeding the detection threshold is significantly lower than the percentages mentioned earlier. Unfortunately, this strategy has limitations in our specific case: the noise in the ALMA images is correlated, meaning that the flux in one random pixel is influenced by the flux in its adjacent pixels. Specifically, the correlation scale is typically similar to the size of the PSF [24]. As a simpler strategy, higher detection thresholds can also be considered. Nevertheless, diminishing contaminants invariably result in a reduction in completeness. Thus, the selection of a flux limit always necessitates a balanced decision between minimizing contamination and maintaining completeness.

To calculate the contamination resulting from the stochastic noise distribution, we used the same flux-inverted maps employed for completeness determination without injecting any simulated sources. From these inverted maps, we followed the source detection procedure outlined in Section 2.2. At various signal-to-noise ratios, we estimated the contamination as the ratio between the number of spurious sources detected in the flux-inverted maps and the number of sources detected in the non-inverted maps used for subsequent analysis.

We note that the reliability of this technique is based on two assumptions. First, the distribution of the pixel fluxes (i.e., of the noise) must be symmetrical with respect to the peak of the distribution itself (with the few pixels occupied by actual sources deviating from this distribution). Second, we assumed that the shape of the distributions of the pixel fluxes is general and not image-specific. In fact, we did not individually correct each image



for its specific contamination curve: the limited number of real detections would make this approach highly unreliable. Instead, we derived a single curve as a function of the signal-to-noise ratio by averaging the contamination fractions measured across all images.

While these two assumptions may not hold in general, our image selection criteria, as outlined in Section 2.1, ensure that, for the images considered in our sample, the pixel flux distributions are both homogeneous and Gaussian, with good approximations.

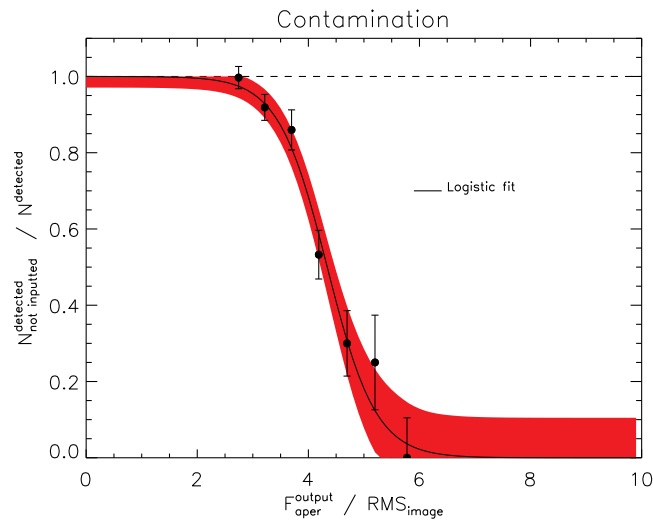
We fit the contamination estimations at various levels of SNR using a logistic curve as follows:

$$f_c^{\text{raw}} = \frac{a}{1 + \exp[b(\text{SNR} + c)]}, \quad (10)$$

where the best-fitting parameters found were  $a = 1.0$ ,  $b = 2.28$ , and  $c = -4.34$ .

In summary, the contamination from spurious detection is computed as the ratio between the number of sources detected without being actually inputted in our simulations ( $N_{\text{not inputted}}^{\text{detected}}$ ) and the total number of sources detected ( $N^{\text{detected}}$ ) in the real images. This ratio is expressed as a function of the significance of the sources detected, i.e., as a function of the ratio between fluxes measured inside beam-size apertures ( $F_{\text{aper}}^{\text{output}}$ ) and the RMS of the images considered (i.e.,  $\text{RMS}_{\text{image}}$ , which is computed with the method described in Section 2.2).

The contamination curve, as determined through the ordinary method previously described and illustrated in Figure 3, exclusively addresses the proportion of false detections originating from the stochastic nature of noise. However, it does not encompass all forms of contamination. We evaluate the impact of these other contaminants in Section 2.6. In Section 3.3, we describe how we mitigate their effects on our number counts using a ML approach.



**Figure 3.** Contamination fraction as a function of the signal-to-noise ratio (black-filled circles and error bars). The best-fitting logistic curve and its associated uncertainty are shown with a black line and a red-filled area, respectively. This type of contamination accounts only for the fraction of false detections due to the stochastic distribution of the pixels fluxes. Other sources of contamination are taken into account separately.

### 2.5. Flux Boosting

Despite accurately determining the average zero-point magnitude, measuring fluxes inevitably introduces errors. These errors result from various factors, including instrumental effects like detector non-linearity, inaccuracies in flat fielding, and vignetting. Additionally, accurately establishing the background near faint sources can be particularly challenging.

Even when the sources of uncertainty mentioned above are minimized, statistical variations still play a crucial role. While all sources have precise intrinsic fluxes, individual

physical measurements inherently exhibit some dispersion around the expected values. Consequently, when applying a detection threshold, only sources with measured fluxes above it are detected, which potentially lead to an average flux boost for sources near the threshold.

A related effect happens when counting the sources within a flux bin, even those well above the detection threshold. Assuming a flat source distribution, an equal number is anticipated to move between adjacent flux bins. Yet, if the intrinsic source distribution is not flat, the number of sources moving between bins depends on the inherent source counts in each considered bin.

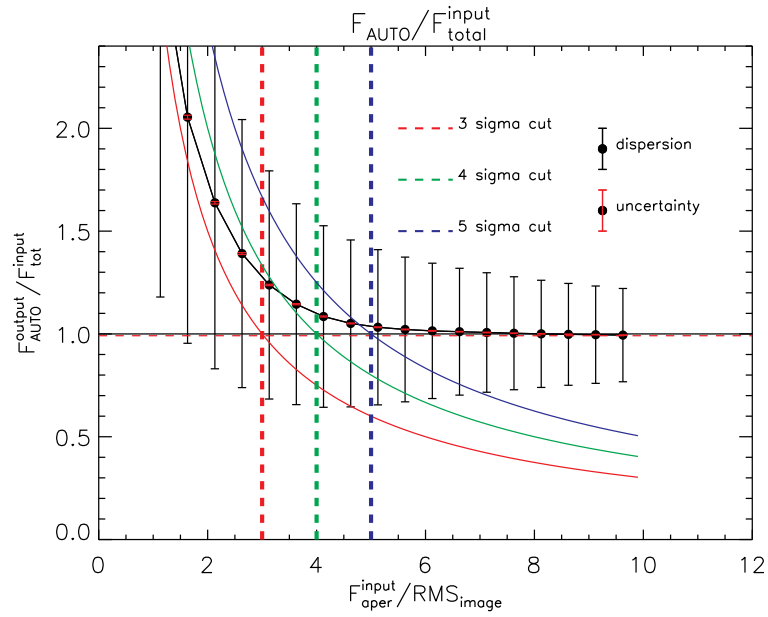
In order to estimate these effects, in each of the images that we simulated (as described in Section 2.3), we injected simulated sources, with PSFs matching the shape and orientation of the PSF specific to the image under consideration (i.e., they were extracted from the header). The quantity of simulated sources differed across images, thereby aiming to use the available space efficiently while preventing overlap. Specifically, we spaced each pair of sources with a distance of eight times the FWHM of the PSF's semi-major axis.

We normalized the injected flux of the simulated sources to match the various levels of SNR, and then we extracted them using the same procedure described in Section 2.2. Finally, we compared the injected flux with the total flux measured through our procedure for each source. Figure 4 illustrates the results of this comparison as a function of the intrinsic SNR (i.e., the ratio between the injected flux and noise). Sources injected at  $3\sigma$  or  $5\sigma$  experienced different degrees of flux boosting. Specifically, at SNR values exceeding  $5\sigma$ , the simulated sources were recovered with nearly the same injected flux (within a few percent deviation on average). On the other hand, at  $3\sigma$ , the recovered flux was, on average, over 20% higher.

It is important to note that adopting a  $5\sigma$  threshold does not eliminate the effects of flux boosting entirely. Despite the higher threshold, many sources injected at an SNR of  $<5\sigma$  are still recovered at higher fluxes, thus contaminating the selected data sample. Even in an ideal scenario where the underlying intrinsic source flux distribution is flat, the dispersion of recovered fluxes (typically larger at a lower SNR) implies that the number of sources intrinsically below the threshold, but still contaminating the sample, is larger than the number of sources with an intrinsic flux being larger than the threshold that will not be detected. The situation described is visualized in Figure 4 using different curves, and it indicates three representative thresholds ( $3\sigma$ ,  $4\sigma$ , and  $5\sigma$ ). For each of those curves, the sources entering in the selected sample are located above the curve itself.

In the computation of source number counts, it is crucial to apply weighting to each source (i.e., each count) based on the specific completeness and contamination estimates at the same flux level (i.e., in the same flux bin) as the detected source. Due to flux boosting, sources that should not be detected end up being counted at higher flux values, thereby contaminating different flux bins and altering the completeness curve. It is important to note that the flux bins are not expected to contain the same number of sources. Consequently, even when having a precise understanding of the fraction of sources lost from one flux bin and gained in another from simulations, we need to know the actual underlying flux distribution of the sources.

However, the accurate determination of the source flux distribution is the ultimate goal of our analysis, thus making a precise correction unfeasible. Nonetheless, we can derive an approximate correction by employing theoretical number counts. It is important to emphasize that this approach does somewhat influence the resulting number counts, but this impact is a second-order effect. Specifically, the correction will be affected by the slope of the theoretical counts curve but not by its normalization.



**Figure 4.** Ratio between the recovered total flux ( $F_{\text{AUTO}}^{\text{output}}$ ) and flux originally injected for the simulated sources ( $F_{\text{tot}}^{\text{input}}$ ) as a function of the intrinsic SNR. Within each SNR bin, the data dispersion are indicated by the black error bars (where the uncertainty in the average measurements was found to be negligible and smaller than the circles used in the plot). Three different SNR thresholds,  $3\sigma$ ,  $4\sigma$ , and  $5\sigma$ , are indicated using dashed vertical lines. It is possible to note that only sources injected below  $4\sigma$  or  $5\sigma$  were substantially affected by flux boosting. This, however, does not indicate that the flux boosting effect can be ignored, even when using these safer thresholds. The continuous curves (with red denoting a  $3\sigma$  threshold, green for  $4\sigma$ , and blue for  $5\sigma$ ) visually indicate the fraction of sources, which were injected at various levels of SNR, entering the sample. The fraction can be estimated comparing the part of dispersion bars above and below the curve considered. For example, when considering a  $5\sigma$  threshold (blue curve), half of the sources injected at  $5\sigma$  will be excluded from the sample, while this fraction decreases to  $\sim 16\%$  when the injected flux is  $7\sigma$ .

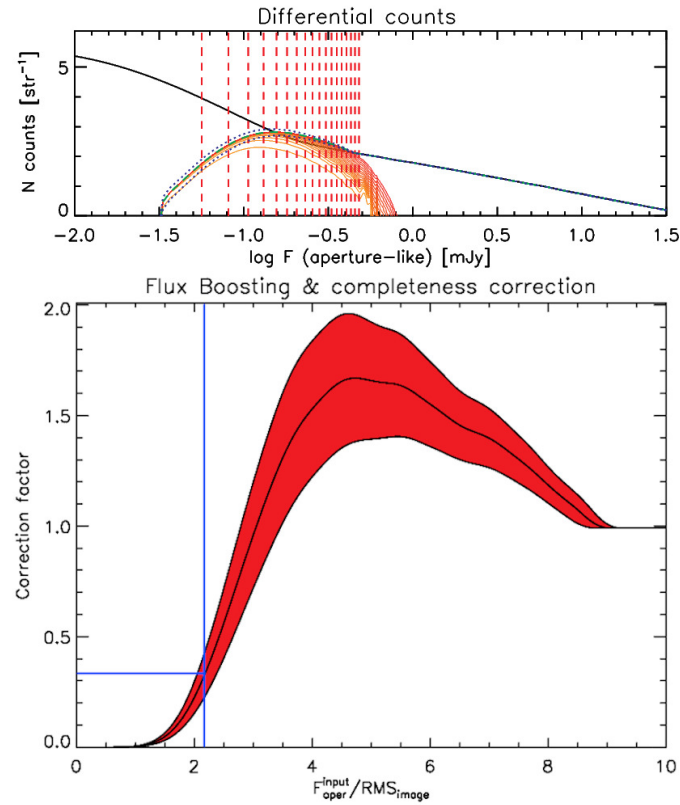
In our analysis, we make use of the predicted number counts from the “C2Ex” model by [25] for radio-loud active galactic nuclei (RL-AGNs) and of the number counts of star-forming galaxies predicted by [26] for star-forming (SF) galaxies.

For all the sources characterized by a specific intrinsic (input) SNR value, we have information about the distribution of recovered SNRs from our simulations (refer to Section 2.3). This distribution already incorporates the completeness achieved at various SNR levels, and, when scaled by the actual value of  $\sigma$  in mJy, it reflects the probability that an input source with flux  $F_i^{\text{in}}$  will be detected at a flux  $F_i^{\text{out}}$ .

Therefore, we proceeded as follows. We divided the theoretical counts curve, which is the sum of the models for RL-AGNs and SF galaxies, into flux bins. Then, we multiplied the average y-axis values within each bin by the corresponding probability function obtained from our simulations. Subsequently, we integrated all of the resulting curves for each bin, thereby extending beyond  $10\sigma$  (Figure 5, upper panel). The ratio between the integrated curve and the theoretical number counts curve provides the correction factor needed to account for both completeness and flux-boosting effects (Figure 5, bottom panel).

It is important to note that this process must be applied individually for each detected source. This is because we need to convert the SNR value from the units of  $\sigma$  to units of mJy, and this conversion is not uniform across the entire image due to primary beam corrections.

In the bottom panel of Figure 5, it becomes evident that, in certain instances, the combined correction (encompassing completeness and flux boosting) can result in a form of over-completeness. This occurs because numerous sources with low flux values can significantly contaminate bins with higher flux values, even when the completeness of these low-flux sources is exceptionally low.



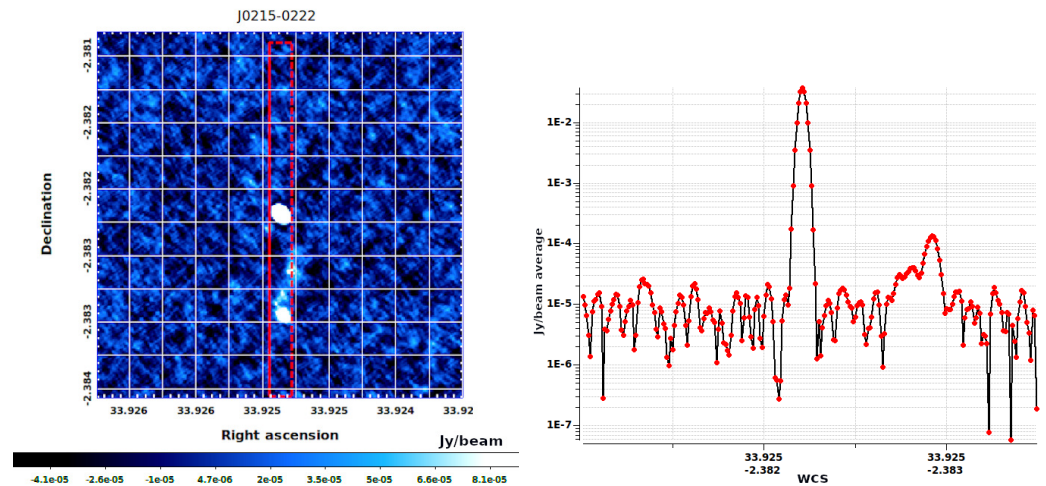
**Figure 5.** Example of the flux-boosting correction curve estimation for a source in our sample. In the (**upper panel**), we segmented the theoretical counts curve (black line) and the sum of the predicted number counts from the “C2Ex” model by [25] for RL-AGNs with the number counts of SF galaxies, as predicted by [26], into logarithmic flux bins. The centers of the bins are denoted by vertical, dashed lines. Given the SNR of each bin, we obtained the distribution of recovered SNRs from our simulations. All these curves are integrated, with results displayed from the lowest yellow curve (representing the first bin) to the highest one (representing the sum over all bins). The ratio between the final integrated curve and the theoretical number counts curve yields the correction curve displayed in the (**bottom panel**). The red-shaded area in the (**bottom panel**) represents the associated uncertainty (which is directly derived from the uncertainty on the integrated curve and is depicted as two dotted lines in the (**upper panel**)). The horizontal blue line represents the correction required for this specific source (which is detected at approximately  $2\sigma$ ). This correction curve comprehensively addresses both the completeness and flux-boosting effects.

## 2.6. Other Unwanted Sources

Our analysis aims to create a collection of the authentic sources detected in the fields of ALMA calibrators without any association to the calibrators themselves. In this context, the spurious detections described in Section 2.4 are not the only contaminants. Additional contamination arises from central calibrators that occasionally exhibit side lobes.

Figure 6 illustrates a typical example of such cases. It is evident that, while the calibrator’s flux is over two orders of magnitude brighter than the lobe, the lobe itself remains more than  $10\sigma$  above the noise level. The “bridge” connecting the calibrator with the lobe is also visible, albeit quantitatively just a few sigma above the noise.

Bright lobes, like the one shown in Figure 6, can be easily identified through a visual inspection of each image in the sample. However, fainter lobes can be misinterpreted as serendipitous detections, particularly when the “bridge” is not immediately detectable and only one of the lobes is visible. Furthermore, calibrators usually reside in over-dense regions. Thus, their fields may contain detectable physical companions, which highly bias the source density.



**Figure 6.** (Left): image of the ALMA calibrator J0215-0222. We measured the average flux along a vertical slit (shown in red), which is positioned to include the positions of both the calibrator and the main lobe. The fluxes, averaged along the x axis, are shown in the plot on the (right). The scientific notation “ $aEb$ ” indicates  $a \times 10^b$ . For example, 1.5E8 stands for  $1.5 \times 10^8$ .

Other sources of contamination related to the presence of the calibrator are false positives, which are often introduced during the imaging process (particularly those in the “cleaning” phase). Calibrators are generally observed at flux levels much higher than the noise. Due to the substantial dynamic range of these images, the cleaning processes might leave behind residuals well above the noise itself. These residuals can be mistaken as genuine sources, and they lead to sample contamination.

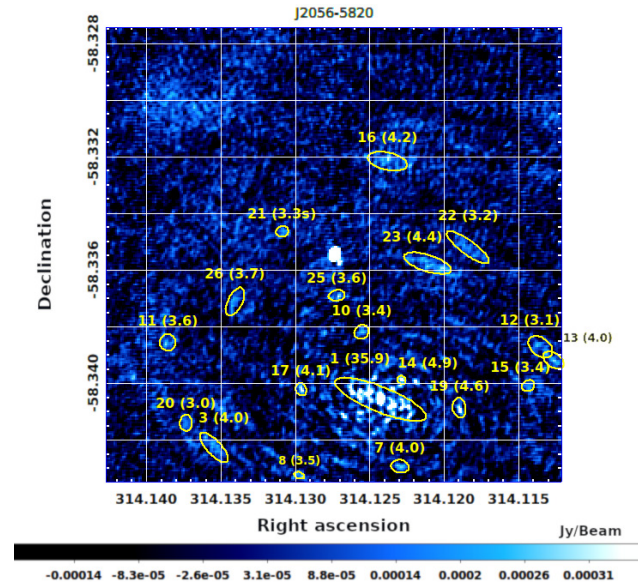
All of this presents a challenge when determining the extent of contamination, as the presence of these undesired sources cannot be predicted from the noise distribution. Statistically, these contaminants do not show the corresponding counterparts at negative fluxes (i.e., their presence cannot be quantified by inverting the fluxes of the maps). Possibly, this type of contamination could be estimated if the actual distribution of the noise and noise patterns were precisely known.

In contrast to the false detections, the contamination related to the calibrator cannot be mitigated by increasing the detection threshold. In fact, we found that the majority of  $\geq 4\text{--}5\sigma$  spurious detections can be attributed to these types of contaminants.

Data reduction and imaging anomalies tend to exhibit discernible spatial patterns, which allow us to visually identify them. For instance, in Figure 7, we present an image featuring numerous false detections with high SNRs, which were arranged along a distinct elliptical spatial pattern. In other cases, however, the patterns are difficult to discern, thus making it challenging to detect the contamination.

Thus far, little attention has been paid, in the literature, to the contaminants discussed in this section. It is usually thought that masking a region within a few arcsecs from the calibrator is enough to get rid of them. Our analysis, however, has demonstrated that they may also come out at much larger angular distances. Evidently, the correction approach in Section 2.4, which exploits inverted maps, cannot be used for these contaminants, which only affect the positive part of the flux distribution. To deal with them, we employed a comprehensive procedure that combines visual inspection and machine learning-based strategies. A detailed description of this procedure is presented in Section 3.





**Figure 7.** Example of an image affected by multiple false detections. We display all the contaminants detected above  $3\sigma$ . For each source, we indicate the ID, which is followed by the SNR in parentheses. It is possible to note that the distribution of contaminants exhibited a distinct spatial pattern characterized by an elliptical shape. The scientific notation “ $aEb$ ” indicates  $a \times 10^b$ . For example,  $1.5E8$  stands for  $1.5 \times 10^8$ .

### 3. Differential Number Counts

The contamination fraction steeply rises at  $\text{RMS} < 5.0\sigma$ , exceeding 95% at  $\text{RMS} < 3.0\sigma$  (Figure 3). At these levels, even a small uncertainty on the estimated contamination (e.g.,  $\sim 5\%$ ) translates into a huge uncertainty on the number counts ( $\sim 100\%$ ). On the other hand, at higher detection limits, the counts in the ALMA bands were extremely steep due to the combined effect of the strong cosmological evolution of dusty, star-forming galaxies and their negative K-correction. Thus, even a small decrease in detection limit resulted in a substantial increase in the number of detections, which would otherwise be very low because of the smallness of the surveyed area.

This has motivated attempts to go down to an  $\text{SNR} = 3.4\text{--}4$  [11,12,27,28]. As a practical compromise, we opted for a threshold corresponding to a contamination level of approximately 50%, which falls between  $4.0\sigma$  and  $4.5\sigma$ . We specifically chose the intermediate value of  $4.25\sigma$ .

It is worth emphasizing that the threshold discussed above primarily addresses the contamination from the spurious detections generated by the random distribution of noise. In contrast, the contamination due to calibrator extensions and cleaning residuals (described in Section 2.6) was relatively insensitive to the choice of the RMS threshold (many of these contaminants can be found well above  $5\sigma$ ). However, the  $4.25\sigma$  threshold roughly corresponds to the minimum SNR for which we are still able to identify these types of contaminants through visual inspection.

In Table 3, we report the effective area covered as a function of the SNR threshold used. It is possible to notice that, in the relevant range of  $-1.0 < \log(F[\text{mJy}]) < 0.0$ , the effective area decreases significantly ( $\sim 11\%$  to  $\sim 43\%$ ), when increasing the threshold from  $4.25\sigma$  to  $5.0\sigma$ .

#### 3.1. Raw Number Counts

As an initial test, we computed the differential number counts by applying corrections only for completeness (Section 2.3), flux boosting (Section 2.5), and for the portion of contaminants due to the stochastic noise distribution (Section 2.4). In other words, in this initial test, we did not correct for the dominant contamination due to calibrator extensions and cleaning residuals (see Section 2.6).

**Table 3.** ALMA band 3 counts—the effective area.

log(F[mJy])	Effective Area [arcmin <sup>2</sup> ]			
	SNR <sub>min</sub> 4.25σ	SNR <sub>min</sub> 4.5σ	SNR <sub>min</sub> 4.75σ	SNR <sub>min</sub> 5.0σ
−1.25	2.38	2.38	2.38	2.38
−1.00	3.18	2.95	2.81	2.72
−0.75	9.07	7.14	6.24	5.69
−0.50	96.1	79.8	66.0	54.3
−0.25	303.5	281.2	260.2	240.6
0.00	524.6	503.9	484.5	465.5
0.25	633.9	628.7	622.3	616.2
0.50	665.2	663.4	662.0	659.7
0.75	674.7	674.3	673.9	673.2
1.00	679.3	679.0	678.7	678.3
1.25	681.8	681.6	681.4	681.2

Therefore, in this initial phase, the contribution from a source  $i$ , which was detected at flux  $F_i$ , to the number of counts, is given by the following:

$$N_i(F_i) = \frac{1 - f_c^{\text{raw}}(\text{SNR}_i)}{D_i C_i A_{\text{eff}}(F_i)}. \quad (11)$$

Here,  $C_i$  represents the completeness estimated at flux  $F_i$ , and  $A_{\text{eff}}(F_i)$  denotes the effective area at the same flux, thereby indicating the total survey area that is sensitive to detecting a source with flux  $F_i$ , given the SNR threshold considered. The raw contamination fraction,  $f_c^{\text{raw}}(\text{SNR}_i)$ , accounts for the contaminants that are due solely to the stochastic distribution of noise, and it is dependent only on the SNR at which source  $i$  is detected, as described in Section 2.4. Lastly,  $D_i$  represents the deboosting correction, which is computed as per Section 2.5. In subsequent analyses, the refined computation of the contamination affected only the term  $f_c^{\text{raw}}$  in the above equation. We derived the uncertainty associated with the counts by combining, in quadrature, the Poissonian uncertainty (which was achieved by considering the small number approximation outlined in [29]), with the overall uncertainty being attributed to the combined effects of completeness, contamination, and flux boosting. These latter uncertainties were addressed using standard error propagation equations.

Figure 8 shows the results of our initial analysis. Due to the contamination from calibrator extensions and false positives, our counts clearly departed from the theoretical expectations, especially those above approximately  $\log(F[\text{mJy}]) \sim -0.75$ .

### 3.2. Number Counts Corrected Through Visual Inspection

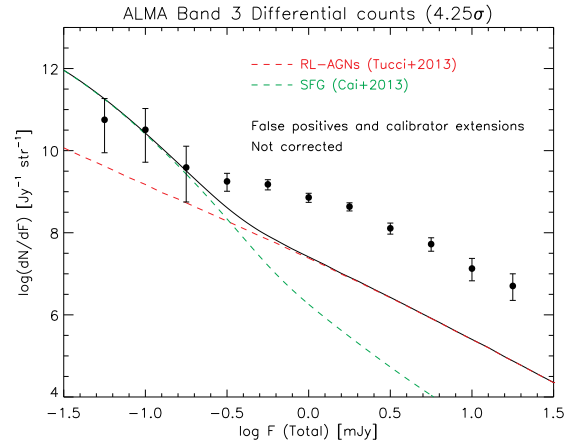
Figure 9 displays the number counts after eliminating the calibrator extensions and false positives identified through a visual inspection of the images. The same results are reported in the second column of Table 4. While achieving absolute security in such a classification is not feasible, we classified anything we judged more likely to be a contaminant than a genuine background source as a contaminant.

Although these results represented an improvement over the uncorrected counts, they still exceeded the theoretical predictions. We indeed expected that this correction was insufficient since our ability to identify these contaminants significantly diminished below SNRs of approximately 4.5. In fact, we achieved a better agreement with the theoretical counts when employing a higher SNR threshold (4.5 or 5).

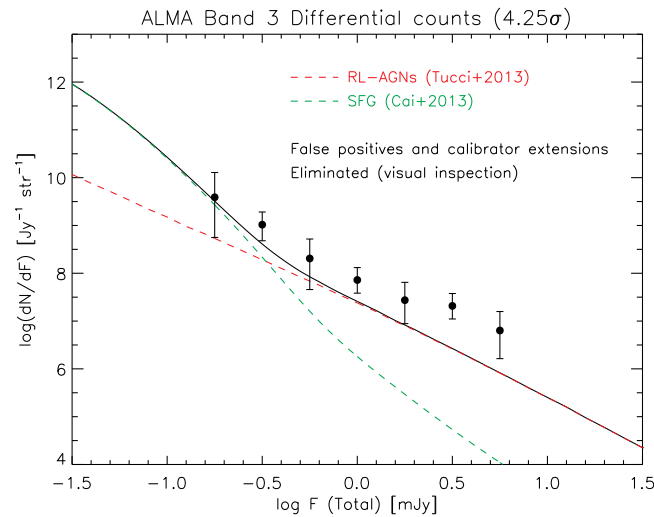
### 3.3. Number Counts Corrected Using Machine Learning

Like any classification based on visual inspection, our identification of calibrator extensions and cleaning residuals carries a degree of subjectivity and may lack reproducibility.

Some of the detections we labeled as contaminants might be considered genuine sources by others, and vice versa.



**Figure 8.** Differential number counts obtained by correcting for completeness, flux boosting, and for the contamination by noise fluctuation. Conversely, the contamination by calibrator extensions and false positives, which substantially affected our number counts above  $\sim 0.3$  mJy, was not corrected. The model predictions by [26] for dusty, star-forming galaxies and by [25] for radio sources are shown by the green and red dashed lines, respectively.



**Figure 9.** Differential number counts obtained by eliminating the contaminants that were caused due to the calibrator extensions and cleaning residuals identified through visual inspection. The theoretical models of [25,26] are shown by green and red dashed lines, respectively.

To ensure a more objective and reproducible identification of contaminants, we used a machine learning (ML) approach by employing a modified version of the K-nearest neighbor (K-NN) algorithm (e.g., [30]), which was implemented through the *UMLAUT* software [31]. Although our personal classification was used as the initial training set for the algorithm, the ultimate identification outcomes stemmed from a more objective assessment grounded in selected features and in the generalization of rules learned from our visual inspection. This method minimizes the inherent subjectivity of visual inspection, and it offers a more standardized approach to identifying contaminants.

Even more importantly, *UMLAUT* provides confidence scores that we used to differentiate between the detections. Higher confidence scores indicate a higher probability that a detection is either a genuine background source or an “ordinary” contaminant (for which corrections were already applied, as detailed in Sections 2.4 and 2.5). Conversely, lower

confidence scores suggest a greater probability that the detections are calibrator extensions or false positives.

*UMLAUT*'s core assumption is that when data samples exhibit similarity across  $N$  dimensions, this similarity extends to an extra  $N + 1$  dimension. Following this principle, it initially determines the position of the data point under analysis within a ranked parameter space. This ranked space has  $N$  dimensions, corresponding to the  $N$  parameters considered, and it was constructed ranking the data points in the training set. Then, it estimates the expected value of the analysis data point along the  $(N + 1)$ -th dimension using the nearest data points of the training set. This may involve, for example, averaging the values of the  $(N + 1)$ -th parameter over the nearest data points.

The process described requires that all the values across all  $N + 1$  dimensions are known for the data points in the reference sample. *UMLAUT* further enhances its assessment using a gradient descent method, where each ranked variable is iteratively weighted (i.e., stretched or compressed) to minimize the gap between the predicted and actual outcomes for the data points of the training sample. This is especially important because users do not need to pre-determine which input parameters are most influential in determining the output; *UMLAUT* dynamically adapts them through this weighting process.

In the specific case of our analysis, we are interested in determining the probability  $P_R^i$  of a given source  $i$  to be a contaminant due to calibrator extensions or to false positives. To achieve this, we used a training set consisting of sources that underwent visual inspection. We assigned  $P_R^j = 1$  when the  $j$ -th source was identified as one of these contaminants and  $P_R^j = 0$  in the opposite case (i.e., the spurious detections that occur due to the stochastic distribution of noise and actual background sources).

This assumption is based on the idea that if, for instance, roughly half of the training sources located near the analysis data point in the  $N$ -dimensional space were classified as calibrator extensions or cleaning residuals, then the average estimated probability for the analysis data point should be  $P_R^i = 0.5$ .

We selected eight input parameters ( $N = 8$ ) to compare the sources under analysis with those in the training set. These parameters encompass the characteristics of both the sources under analysis and of the images in which they were detected. Specifically, we considered the following:

- The SNR of the source, which is the ratio between the source's flux and the noise of the image (both computed using beam-sized apertures);
- The aperture flux of the source, which was corrected for the primary beam;
- The distance of the detection from the image center, which was measured in units of FWHM;
- The FWHM of the image, which was measured in arc-seconds;
- The RMS ratio of the image, which is the ratio between the RMS of the image computed inside the beam-sized apertures and the RMS computed over the pixels;
- The apparent size of the source, which is the ratio between the semi-major axis of the source and the FWHM of the image's PSF;
- The aperture-to-total flux ratio of the source;
- The ellipticity of the source, which is the ratio between the semi-major and semi-minor axes of the source (assuming it is elliptical in shape).

To ensure a robust training set, we restricted it to detections above  $4.5 \sigma$ , where our visual classification demonstrated higher accuracy. In accordance with the "leave-one-out cross-validation" method [32–34], all the sources under analysis were also incorporated into the visually checked training set. However, they were included at different stages of the process. In essence, when *UMLAUT* is tasked with classifying a particular data point within the dataset, it assesses the data point's position in relation to all other data points in the same dataset, excluding the point under analysis. This systematic approach safeguards against the risk of over-fitting. This aspect has been extensively substantiated, particularly for the case of *UMLAUT*, as demonstrated in ref. [31] (see Appendix A therein).

To mitigate the influence of *UMLAUT*'s results on the specific configuration chosen, we adopted a strategy of averaging the outputs from various reasonable configurations. We ran *UMLAUT* multiple times, each time varying the number of closest data points considered. In the initial set of runs, we allowed the parameter  $k$  to span from  $k_{\min} = 2$  to  $k_{\max} = 8$ . In this approach, we calculated the output probability  $P_R^j(k)$  as the average value derived from the  $k$  closest data points (default option AVERAGE="mean" in *UMLAUT*):

$$P_R^i(k) = \frac{1}{k} \sum_{j=0}^k P_R^j. \quad (12)$$

Subsequently, we calculated an alternative probability denoted as  $P_R^{ii}(k)$ . This involved taking an average weighted by the distances between the data points (option AVERAGE = "weighted" in *UMLAUT*). Unlike the previous approach, where only the  $k$  closest data points were considered, in this case, we incorporated all the data points in the dataset. However, each data point was weighted using a Gaussian-shaped weighting function with  $k$  as the  $\sigma$  of the Gaussian.

The combined result was computed by averaging all the solutions obtained across the various  $k$  values and the two distinct methods as follows:

$$P_R^i = \frac{1}{(k_{\max} - k_{\min})} \sum_{k_{\min}}^{k_{\max}} \frac{P_R^i(k) + P_R^{ii}(k)}{2}. \quad (13)$$

The uncertainty associated with these probabilities was calculated as the dispersion among all the values of  $P_R^j(k)$  or  $P_R^{ii}(k)$  that were obtained over the  $N_{test} = 2(k_{\max} - k_{\min})$  tests performed <sup>2</sup>.

It is important to note that the reliability of the outputs generated by *UMLAUT* varies depending on whether we aggregate the probabilities associated with the individual detections inside the bins or if we analyze them individually.

In our analyses, we opted to utilize the *UMLAUT* outputs exclusively for the sources with fluxes exceeding 0.3 mJy. This decision stemmed from the observation that the three bins below this threshold contained only one detection each. In such cases, these individual detections cannot be aggregated, thus making the probabilities associated with single detection lacking in meaningful interpretation.

Furthermore, upon visual inspection, we found that two of these three sources were extensions of the central calibrators. For these detections, *UMLAUT* appropriately assigned the low probabilities to be real background sources. However, this directly results in extremely low number counts in the estimations as there are no genuine background sources in the same bins that can statistically compensate for these low probabilities. For the reasons explained above, for detections below 0.3 mJy, we relied on our own visual inspection.

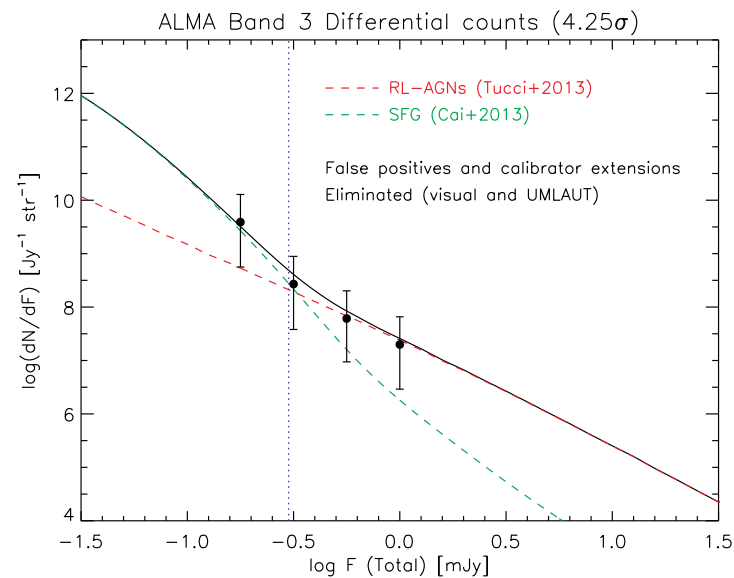
In our initial *UMLAUT* test, we employed a classification criterion that removed detections identified as calibrator extensions or cleaning residuals by *UMLAUT* with a probability  $P_R^i > 0.5$ .

The results of this test are displayed in Figure 10, and they are also reported in the third column of Table 4. It is evident that the bins above  $F = 1$  mJy appear to be completely empty. However, this does not necessarily mean that all the sources in these bins are contaminants. To clarify, let us consider an example: suppose there are 10 detections in a particular bin and each of them has a probability  $P^i \sim 0.1$  to be an actual background source. In this case, we would expect that one of them is a genuine background source. However, since all the individual probabilities  $P^i$  were below 0.5, all of these sources were excluded from the count, which made the bin appear as it were empty. Once again, this explanatory example highlights the importance of aggregating *UMLAUT* outputs.

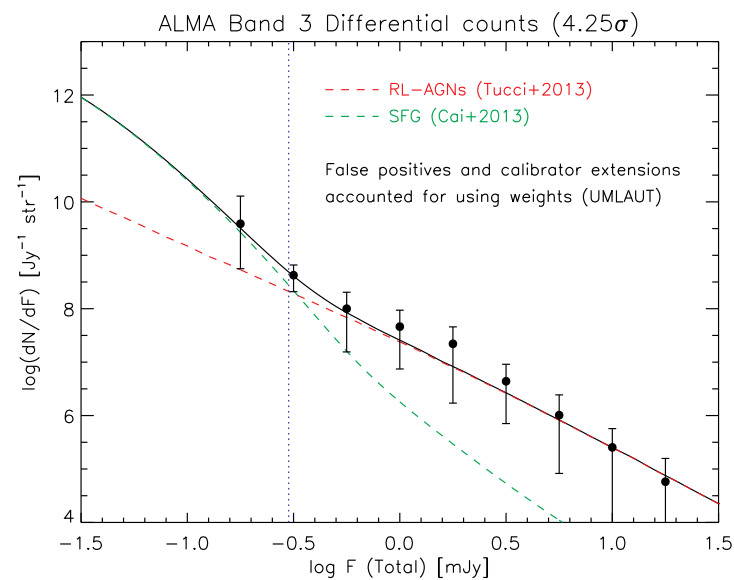
To overcome the limitation described above, in our second test, we employed the probabilities calculated by *UMLAUT* to assign weights to each of the detections. Specifically,



while still weighting each detection for completeness, flux boosting, and contamination (for the part that was exclusively due to the stochastic distribution of noise), as we did in our other tests, in this case we additionally applied a weight computed as  $w^i = 1 - P_R^i$ . The results of this computation are shown in Figure 11, and they are also reported in the third column of Table 4.



**Figure 10.** The differential number counts derived by excluding the detections classified (with a probability exceeding 50%) as calibrator extensions or cleaning residuals by *UMLAUT*. For cases below  $\sim 0.3$  mJy (vertical dotted line), we relied on our own visual inspection. The theoretical models of [25,26] are shown by the green and red dashed lines, respectively.



**Figure 11.** The differential number counts derived by weighting the detections using the probabilities, which were computed by *UMLAUT*, of being calibrator extensions or cleaning residuals. The weights were computed only for sources above  $\sim 0.3$  mJy (vertical dotted line), and we relied on our visual classification for the few detections that occurred below this threshold. The theoretical models of [25,26] are shown by the green and red dashed lines, respectively.

**Table 4.** ALMA band 3—the differential number counts.

$\log(F[\text{mJy}])$	$\log(\text{counts} [\text{Jy}^{-1} \text{sr}^{-1}])$ Visual Inspection	$\log(\text{counts} [\text{Jy}^{-1} \text{sr}^{-1}])$ UMLAUT Eliminated	$\log(\text{counts} [\text{Jy}^{-1} \text{sr}^{-1}])$ UMLAUT Weighted
−0.75	9.59 <sup>10.11</sup> <sub>8.75</sub>	9.59 <sup>10.11</sup> <sub>8.75</sub> <sup>[a]</sup>	9.59 <sup>10.11</sup> <sub>8.75</sub> <sup>[a]</sup>
−0.50	9.02 <sup>9.28</sup> <sub>8.68</sub>	8.43 <sup>8.94</sup> <sub>7.58</sub>	8.63 <sup>8.82</sup> <sub>8.32</sub>
−0.25	8.31 <sup>8.72</sup> <sub>7.66</sub>	7.79 <sup>8.30</sup> <sub>6.97</sub>	8.00 <sup>8.31</sup> <sub>7.19</sub>
0.00	7.86 <sup>8.12</sup> <sub>7.58</sub>	7.30 <sup>7.82</sup> <sub>6.46</sub>	7.66 <sup>7.97</sup> <sub>6.87</sub>
0.25	7.44 <sup>7.81</sup> <sub>6.95</sub>	-	7.34 <sup>7.66</sup> <sub>6.23</sub>
0.50	7.32 <sup>7.58</sup> <sub>7.04</sub>	-	6.64 <sup>6.96</sup> <sub>5.85</sub>
0.75	6.80 <sup>7.20</sup> <sub>6.21</sub>	-	6.01 <sup>6.96</sup> <sub>5.85</sub>
1.00	-	-	5.40 <sup>5.76</sup> <sub>-</sub>
1.25	-	-	4.76 <sup>5.20</sup> <sub>-</sub>

<sup>[a]</sup> Due to the limited number of detections populating the bins below 0.3 mJy (mostly contaminants), only visual inspection was employed for the sources detected below this threshold.

#### 4. Conclusions

The struggle to resolve, as far as possible, the millimeter and sub-millimeter extragalactic backgrounds faces strong challenges. Even the largest single-dish telescopes, of the 6–10 m class, suffer from severe confusion limits. ALMA overcomes such limits thanks to its excellent sensitivity and angular resolution, but it also has a tiny field of view.

The most effective ALMA exploitation for extragalactic surveys takes advantage of fields around the sources observed for other purposes, particularly those around calibrators. Such an approach offers the following advantages:

- Increased survey area: Given the large number of calibrators available, these fields provide observations that can be used for survey purposes;
- Diverse target selection: These fields may contain a wider variety of source types that are not originally targeted;
- Free observation time: Since ALMA is already pointed at in the calibrator field, observing additional targets does not require additional observation time.

The large number of calibrator observations, which were retrieved and calibrated by the ALMACAL project, has made it possible to gather a total surveyed area in band 3 of about 0.2 deg<sup>2</sup>.

This area, although much larger than that covered by blind ALMA surveys, is nevertheless too small to achieve a good statistical determination of the number counts if we restrict ourselves to sources above the canonical  $5\sigma$  detection limit. Going somewhat fainter substantially increases the number of detections since, at the depth of ALMA fields, the number counts, which are dominated by high- $z$ , dusty, and star-forming galaxies, are extremely steep. The drawback is that the necessary corrections for incompleteness, contamination by noise fluctuations, and flux boosting rapidly increase with decreasing SNRs below 5. Procedures to compute these corrections have been widely discussed in the literature, and they were carefully dealt with in this paper.

We conducted a thorough investigation into contaminants. On the one hand, when using traditional simulation techniques (e.g., inverting the map fluxes and randomly placing simulated sources all around the field), it is possible to estimate the fraction of spurious detections due to the stochastic nature of noise. On the other hand, the same techniques cannot reveal the presence of real foreground sources or other false positives associated with the calibrator itself. The presence of such contaminants is evident when visually inspecting the images and when comparing the number counts with the theoretical expectations. Previous analyses have often overlooked this issue, wherein they merely masked a region within a few arc-seconds from the calibrator, or by discarding entire images when visibly contaminated by such sources.

Our machine learning-based approach, which exploits the UMLAUT algorithm, has underscored these kind of contaminants that are spread throughout the field. The algo-

rithm assigns to each source the probability of being a contaminant. Together with the corrections for incompleteness, spurious detections, and flux boosting (Sections 2.3–2.5), these combined probabilities have allowed us to obtain a solid estimate of 100 GHz number counts of extragalactic sources over about 2.5 orders of magnitude in flux density. An analysis of the scientific results of this research is presented by [35].

**Author Contributions:** Conceptualization, I.B.; formal analysis, I.B.; investigation, I.B.; methodology, I.B.; software, I.B.; validation, I.B.; writing—original draft, I.B.; writing—review and editing, I.B., M.B., G.D.Z., V.C., M.D.V., F.G., E.L., R.P., L.T. and M.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding

**Data Availability Statement:** The ALMACAL images used in this paper are publicly available on the ALMA science archive (<https://almascience.eso.org/alma-data>, accessed on 29 April 2024).

**Acknowledgments:** We wish to thank Vincenzo Galluzzi for his insightful remarks and for helping us with the revision of the manuscript. We also wish to thank the anonymous referees for their valuable comments, which improved the quality of the paper.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Notes

<sup>1</sup> See <https://science.nrao.edu/facilities/vla/proposing/TBconvforadetaileddescription> (accessed on 29 April 2024).

<sup>2</sup> The uncertainty cannot be obtained by dividing the dispersion for  $\sqrt{N_{test} - 1}$ , as the results derived using different configurations (but with the same training set) are not independent of each other

## References

1. Gralla, M.B.; Marriage, T.A.; Addison, G.; Baker, A.J.; Bond, J.R.; Crichton, D.; Datta, R.; Devlin, M.J.; Dunkley, J.; Dünner, R.; et al. Atacama Cosmology Telescope: Dusty Star-forming Galaxies and Active Galactic Nuclei in the Equatorial Survey. *Astrophys. J.* **2020**, *893*, 104. [CrossRef]
2. Mocanu, L.M.; Crawford, T.M.; Vieira, J.D.; Aird, K.A.; Aravena, M.; Ausermann, J.E.; Benson, B.A.; Béthermin, M.; Bleem, L.E.; Bothwell, M.; et al. Extragalactic Millimeter-wave Point-source Catalog, Number Counts and Statistics from 771 deg<sup>2</sup> of the SPT-SZ Survey. *Astrophys. J.* **2013**, *779*, 61. [CrossRef]
3. Everett, W.B.; Zhang, L.; Crawford, T.M.; Vieira, J.D.; Aravena, M.; Archipley, M.A.; Ausermann, J.E.; Benson, B.A.; Bleem, L.E.; Carlstrom, J.E.; et al. Millimeter-wave Point Sources from the 2500 Square Degree SPT-SZ Survey: Catalog and Population Statistics. *Astrophys. J.* **2020**, *900*, 55. [CrossRef]
4. Eddington, A.S. On a formula for correcting statistics for the effects of a known error of observation. *Mon. Not. R. Astron. Soc.* **1913**, *73*, 359–360. [CrossRef]
5. Hogg, D.W.; Turner, E.L. A Maximum Likelihood Method to Improve Faint-Source Flux and Color Estimates. *Publ. Astron. Soc. Pac.* **1998**, *110*, 727–731. [CrossRef]
6. Walter, F.; Decarli, R.; Aravena, M.; Carilli, C.; Bouwens, R.; da Cunha, E.; Daddi, E.; Ivison, R.J.; Riechers, D.; Smail, I.; et al. ALMA Spectroscopic Survey in the Hubble Ultra Deep Field: Survey Description. *Astrophys. J.* **2016**, *833*, 67. [CrossRef]
7. Hatsukade, B.; Kohno, K.; Yamaguchi, Y.; Umehata, H.; Ao, Y.; Aretxaga, I.; Caputi, K.I.; Dunlop, J.S.; Egami, E.; Espada, D.; et al. ALMA twenty-six arcmin<sup>2</sup> survey of GOODS-S at one millimeter (ASAGAO): Source catalog and number counts. *Publ. Astron. Soc. Jpn.* **2018**, *70*, 105. [CrossRef]
8. González-López, J.; Decarli, R.; Pavesi, R.; Walter, F.; Aravena, M.; Carilli, C.; Boogaard, L.; Popping, G.; Weiss, A.; Assef, R.J.; et al. The Atacama Large Millimeter/submillimeter Array Spectroscopic Survey in the Hubble Ultra Deep Field: CO Emission Lines and 3 mm Continuum Sources. *Astrophys. J.* **2019**, *882*, 139. [CrossRef]
9. González-López, J.; Novak, M.; Decarli, R.; Walter, F.; Aravena, M.; Carilli, C.; Boogaard, L.; Popping, G.; Weiss, A.; Assef, R.J.; et al. The ALMA Spectroscopic Survey in the HUDF: Deep 1.2 mm Continuum Number Counts. *Astrophys. J.* **2020**, *897*, 91. [CrossRef]
10. Gómez-Guijarro, C.; Elbaz, D.; Xiao, M.; Béthermin, M.; Franco, M.; Magnelli, B.; Daddi, E.; Dickinson, M.; Demarco, R.; Inami, H.; et al. GOODS-ALMA 2.0: Source catalog, number counts, and prevailing compact sizes in 1.1 mm galaxies. *Astron. Astrophys.* **2022**, *658*, A43. [CrossRef]
11. Hatsukade, B.; Ohta, K.; Seko, A.; Yabe, K.; Akiyama, M. Faint End of 1.3 mm Number Counts Revealed by ALMA. *Astrophys. J. Lett.* **2013**, *769*, L27. [CrossRef]
12. Carniani, S.; Maiolino, R.; De Zotti, G.; Negrello, M.; Marconi, A.; Bothwell, M.S.; Capak, P.; Carilli, C.; Castellano, M.; Cristiani, S.; et al. ALMA constraints on the faint millimetre source number counts and their contribution to the cosmic infrared background. *Astron. Astrophys.* **2015**, *584*, A78. [CrossRef]

13. Kohno, K.; Fujimoto, S.; Tsujita, A.; Kokorev, V.; Brammer, G.; Magdis, G.E.; Valentino, F.; Laporte, N.; Sun, F.; Egami, E.; et al. Unbiased surveys of dust-enshrouded galaxies using ALMA. *arXiv* **2023**, arXiv:2305.15126. [[CrossRef](#)]
14. Oteo, I.; Zwaan, M.A.; Ivison, R.J.; Smail, I.; Biggs, A.D. ALMACAL I: First Dual-band Number Counts from a Deep and Wide ALMA Submillimeter Survey, Free from Cosmic Variance. *Astrophys. J.* **2016**, *822*, 36. [[CrossRef](#)]
15. Chen, J.; Ivison, R.J.; Zwaan, M.A.; Smail, I.; Klitsch, A.; Péroux, C.; Popping, G.; Biggs, A.D.; Szakacs, R.; Hamanowicz, A.; et al. ALMACAL IX: Multiband ALMA survey for dusty star-forming galaxies and the resolved fractions of the cosmic infrared background. *Mon. Not. R. Astron. Soc.* **2023**, *518*, 1378–1397. [[CrossRef](#)]
16. Condon, J.J.; Cotton, W.D.; Greisen, E.W.; Yin, Q.F.; Perley, R.A.; Broderick, J.J. The NRAO VLA Sky Survey. In *Astronomical Data Analysis Software and Systems III*; Astronomical Society of the Pacific Conference Series; Crabtree, D.R., Hanisch, R.J., Barnes, J., Eds.; Astronomical Society of the Pacific: San Francisco, CA, USA, 1994; Volume 61, p. 155.
17. White, R.L.; Becker, R.H.; Helfand, D.J.; Gregg, M.D. A catalog of 1.4 GHz radio sources from the FIRST survey. *Astrophys. J.* **1997**, *475*, 479. [[CrossRef](#)]
18. Condon, J.J.; Cotton, W.D.; Greisen, E.W.; Yin, Q.F.; Perley, R.A.; Taylor, G.B.; Broderick, J.J. The NRAO VLA Sky Survey. *Astron. J.* **1998**, *115*, 1693–1716. [[CrossRef](#)]
19. Bertin, E.; Arnouts, S. SExtractor: Software for source extraction. *Astron. Astrophys. Suppl. Ser.* **1996**, *117*, 393–404. [[CrossRef](#)]
20. Bondi, M.; Ciliegi, P.; Zamorani, G.; Gregorini, L.; Vettolani, G.; Parma, P.; de Ruiter, H.; Le Fevre, O.; Arnaboldi, M.; Guzzo, L.; et al. The VLA-VIRMOS Deep Field. I. Radio observations probing the  $\mu$  Jy source population. *Astron. Astrophys.* **2003**, *403*, 857–867. [[CrossRef](#)]
21. Huynh, M.T.; Hopkins, A.; Norris, R.; Hancock, P.; Murphy, T.; Jurek, R.; Whiting, M. The Completeness and Reliability of Threshold and False-discovery Rate Source Extraction Algorithms for Compact Continuum Sources. *Publ. Astron. Soc. Aust.* **2012**, *29*, 229–243. [[CrossRef](#)]
22. Hancock, P.J.; Murphy, T.; Gaensler, B.M.; Hopkins, A.; Curran, J.R. Compact continuum source finding for next generation radio surveys. *Mon. Not. R. Astron. Soc.* **2012**, *422*, 1812–1824. [[CrossRef](#)]
23. Kron, R.G. Photometry of a complete sample of faint galaxies. *Astrophys. J. Suppl. Ser.* **1980**, *43*, 305–325. [[CrossRef](#)]
24. Tsukui, T.; Iguchi, S.; Mitsuhashi, I.; Tadaki, K. Proper evaluation of spatially correlated noise in interferometric images. In *Proceedings of the Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy XI; 2022; Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Montreal, QC, Canada, 31 August 2022*; Zmuidzinas, J., Gao, J.R., Eds.; SPIE: Cergy Pontoise, France; Volume 12190, p. 121901C. [[CrossRef](#)]
25. Tucci, M.; Toffolatti, L.; de Zotti, G.; Martínez-González, E. High-frequency predictions for number counts and spectral properties of extragalactic radio sources. New evidence of a break at mm wavelengths in spectra of bright blazar sources. *Astron. Astrophys.* **2011**, *533*, A57. [[CrossRef](#)]
26. Cai, Z.Y.; Lapi, A.; Xia, J.Q.; De Zotti, G.; Negrello, M.; Gruppioni, C.; Rigby, E.; Castex, G.; Delabrouille, J.; Danese, L. A Hybrid Model for the Evolution of Galaxies and Active Galactic Nuclei in the Infrared. *Astrophys. J.* **2013**, *768*, 21. [[CrossRef](#)]
27. Ono, Y.; Ouchi, M.; Kurono, Y.; Momose, R. Faint Submillimeter Galaxies Revealed by Multifield Deep ALMA Observations: Number Counts, Spatial Clustering, and a Dark Submillimeter Line Emitter. *Astrophys. J.* **2014**, *795*, 5. [[CrossRef](#)]
28. Fujimoto, S.; Ouchi, M.; Ono, Y.; Shibuya, T.; Ishigaki, M.; Nagai, H.; Momose, R. ALMA Census of Faint 1.2 mm Sources Down to  $\sim 0.02$  mJy: Extragalactic Background Light and Dust-poor, High- $z$  Galaxies. *Astrophys. J. Suppl. Ser.* **2016**, *222*, 1. [[CrossRef](#)]
29. Gehrels, N. Confidence Limits for Small Numbers of Events in Astrophysical Data. *Astrophys. J.* **1986**, *303*, 336. [[CrossRef](#)]
30. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175–185. [[CrossRef](#)]
31. Baronchelli, I.; Scarlata, C.M.; Rodríguez-Muñoz, L.; Bonato, M.; Morselli, L.; Vaccari, M.; Carraro, R.; Barrufet, L.; Henry, A.; Mehta, V.; et al. Identification of Single Spectral Lines in Large Spectroscopic Surveys Using UMLAUT: An Unsupervised Machine-learning Algorithm Based on Unbiased Topology. *Astrophys. J. Suppl. Ser.* **2021**, *257*, 67. [[CrossRef](#)]
32. Allen, D.M. The Relationship between Variable Selection and Data Augmentation and a Method for Prediction. *Technometrics* **1974**, *16*, 125–127. [[CrossRef](#)]
33. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1974**, *36*, 111–133. [[CrossRef](#)]
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2001.
35. Bonato, M. 100 GHz ALMA number counts. INAF—Istituto di Radioastronomia, Via Gobetti 101, Bologna (I-40129), Italy. 2024, manuscript in preparation.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.