Reliability and Completeness for the GLIMPSE Survey

C. Watson, E. Churchwell, R. Indebetouw, M. Meade, B. Babler, B. Whitney

Abstract

This document examines the GLIMPSE observing strategy and criteria used to produce a high reliability ($\geq$99.5%) point-source catalog. Simulations of GLIMPSE observations were made, processed and results compared with input source lists. The simulations used a standard source color and luminosity distribution, along with a source density based on 2MASS observations and a background brightness based on MSX observations. Observing strategies using two and three exposures were compared. It was found that the best criterion for selecting sources for the catalog with reliability $\geq$99.5% was requiring two detections in an IRAC band and at least one detection in an adjacent band. To measure reliability with observations only, an internal truth table criterion was determined which would most closely match the external reliability results. A robust internal truth table was found to be comprised of sources detected at least 8 times (out of 10 possible) in a given band or an adjacent band.

1 Introduction

As comprehensive surveys become more common in astronomy, it is increasingly critical to develop methods for evaluating their completeness and reliability. Furthermore, the requirements for a source to be included in a highly reliable catalog requires extensive investigation. Although these criteria may be investigated to a limited degree through repeated observations of the same area of the sky, such a method is inherently limited by not knowing if a detected source is a true source. Thus, a comprehensive study requires creating catalogs from simulated images using different criteria. By matching these catalogs against the true source list used to create the simulations, a reliability and completeness measurement external to observations is possible. In this study we explore selection criteria for inclusion of highly reliable ($\geq$99.5%) sources in the Galactic Legacy Infrared Mid-Plane Survey Extraordinaire (GLIMPSE) catalog based on realistic simulations with a known truth table.

GLIMPSE, a Spitzer Space Telescope Legacy Science Program, will be a fully sampled, confusion-limited infrared survey of the inner two- thirds of the Galactic disk with a pixel size of 1.2" using the Infrared Array Camera (IRAC) at 3.6, 4.5, 5.8, and 8.0 μm (see Benjamin et al, 2003). The survey will cover Galactic latitudes $|b| < 1$ and longitudes $|l| = 10$ to 65 (both sides of the Galactic center). The survey area contains the outer ends of the Galactic bar, the Galactic molecular ring, and the inner spiral arms.

Completeness and reliability of the GLIMPSE survey were calculated using the simulations described below. Simulated images were created using an external truth table of known point sources against which point sources extracted from those images with the GLIMPSE data processing pipeline could be compared. The main goal was to determine the best criteria to achieve a reliability $\geq$99.5% in the GLIMPSE survey when each point in the survey area is observed only twice in each of the 4 IRAC bands. In actual observations there will be no external truth table from which completeness and reliability can be established. In this document, we show how the simulated data can be used to establish an "internal truth table" from which reliability can be estimated from real data in the absence of an external truth table.

2 Simulations

The GLIMPSE project has obtained a set of early Spitzer Space Telescope observations centered at l =283.4, $b = -0.3$. Observations will cover an area of 0.7 square degrees repeatedly (10 times). The repetition is necessary to determine the quality of data obtained using different survey strategies (e.g. a survey comprising 3 repeated observations instead of the currently planned 2). We now describe simulations of these observations and how we analyze them to calculate the reliability and completeness of the GLIMPSE survey.

The primary engine for creating the simulated IRAC data is Matt Ashby's IRAC Science Data Software (ISDS), which accepts a truth image with resolution at least that of IRAC (1.2 arcsec/pixel) or better and outputs simulated IRAC frames with Poisson noise, pointing error, MUX bleed, saturation (nonlinear response and eventual wraparound), and some other lesser instrumental effects. We add banding to channels 3 and 4 frames using a model based on BRUTUS chamber data obtained through Tom Megeath and the IRAC team. We also attach a world coordinate (WCS) to the images based on the input WCS and the apparent behavior of the ISDS. The ISDS coordinate calculations had errors of 1-2". These errors were corrected in the GLIMPSE pipeline. The simulated images were "flat-fielded" using the same gain maps that were originally used by the ISDS, resulting in images that were photometrically calibrated but have higher noise levels at the frame edges because of the vignetting and decreased sensitivity there. Finally, star lists were created for each frame with x,y positions for comparison with our source extraction routines.

The truth image was created from MSX and 2MASS data with the intent of creating images that are as realistic as possible with variable background, a wide range of stellar densities, and extended emission and absorption features. First, the point sources listed in the MSX catalog were removed from the MSX images. The method is tuned to produce a background free of stellar residuals, not to most accurately preserve the flux of sources in the catalog: the region of the image containing each point source is fitted with a Gaussian whose initial parameters are the MSX PSF width and the flux density of the cataloged source, but the peak flux density from the image can be slightly different from the source in the catalog.

Stars with known positions and flux densities were added to the MSX-derived diffuse background. The base starlist was the 2MASS catalog, with a modified color distribution for the IRAC bands: objects that are red in

J,H,K are red in the simulated IRAC data, but there is no attempt to classify the objects in the 2MASS catalog and place them in the simulated IRAC images with the correct colors for their SED. Thus the color and flux distributions are as correct as possible, but the color of a given object is not necessarily the color it will have when GLIMPSE is performed. Finally, the input starlist is augmented at the faint and bright ends, to fill out the stellar distribution to fit a power law distribution of slope -1. The augmentation of bright sources is necessary because the second incremental release is missing a lot of the brightest stars that produced large diffraction spikes in the 2MASS data. The final 2MASS release is expected to include most of those omitted stars. At the faint end, we wish to have known stars well below our expected confusion limit to fully test our source extraction routines at the faint end and to properly include the background introduced by unresolved faint stars. The inclusion of these undetectable faint stars as well as the diffuse MSX background (which also contains unresolved faint objects) means that the diffuse background in the simulated data is probably a worst case or slight overestimate, but this is acceptable to make a conservative analysis of our photometric routines. The source color and flux distributions were constrained to match those predicted by the SKY program (Cohen, priv. communication).

3 Definitions

3.1 Externally Determined Reliability

10 different sets of simulated images were created with 10 different realizations of the random noise at the level expected for a 2^s exposure. DAOPHOT was used to extract sources (positions, flux densities, and errors) from each of the ten sets of simulated data. The extracted sources were merged (i.e. cross-identified) in each of the ten "exposures" within a band and across bands. The extracted sources from DAOPHOT were also matched with the truth table.

We define a true source as one that corresponds in position and flux density with one in the external truth table. When the true flux and measured flux are different by more than 1 magnitude, the source is classified as false. This flux discrepancy could result from incorrect photometry or from misidentification of the source in the merging process. Both effects could be

Table 1: Different possible selection criteria for creating the GLIMPSE catalog

Criteria Name	Symbol	Description
2+2:	+	Two detections are required in a single band and another two detections are required in an adjacent band.
2+1:	◊	Two detections are required in a single band and at least one detection is required in one adjacent band.
2+x:	△	Two detections are required in a single band. No requirements are made for adjacent bands
3+2:	□	Three detections are required in a single band and at least two detections are required in one adjacent band.
3+1:	X	Three detections are required in a single band and at least one detection is required in one adjacent band.

present in the real survey, so they are allowed to count against our calculated reliability, making these simulations a conservative prediction of true reliability. Reliability and completeness are defined in the standard way:

$$\begin{aligned} R &= \frac{T_{detected}}{T_{detected} + F_{detected}} \\ C &= \frac{T_{detected}}{T_{all}}, \end{aligned}$$

where $T_{detected}$ is the number of detected true sources, $F_{detected}$ is the number of detected false sources and T_{all} is the total number of true sources. All quantities are for a chosen flux range.

In order to predict the reliability of the GLIMPSE survey, we must first define a criterion for creating a GLIMPSE catalog. We consider five different criteria (see Table 1). The first three criteria assume two exposures in each band, the last two criteria assume three exposures in each band.

Since our simulations and Observing Strategy Verification (OSV) observations consist of ten exposures per band, predicting the GLIMPSE catalog reliability is somewhat complicated. The GLIMPSE catalog reliability is:

$$R(S_v) = \frac{\sum_{S_i > S_v}^{n} P_i \times T_i}{\sum_{S_i > S_v}^{n} P_i}$$

where T_i is 1 for true sources, 0 for false sources, and P_i is the probability that source i will be in the catalog and is a complicated function of the selection criterion (see Appendix A for a full explanation). The reliability is summed over all sources with flux greater than a given flux, S_ν (i.e. a cumulative reliability). For example, a source that is detected only twice in ten possible exposures in a given band has a small chance of being included in the catalog. That is, in the GLIMPSE survey, such a point source is unlikely to be detected at all, much less twice in one band and once in an adjacent band. Thus, such a point source appropriately receives lower weight compared to a point source detected ten times out of ten exposures in each band.

We produced two catalogs, one assuming two observations and one assuming three observations of each point in the sky, to evaluate how much improvement in reliability, and completeness would be gained with three versus two passes. We have also tested several criteria for including a source in the catalog and their impact on the reliability and completeness of the survey for both three passes and two passes. These criteria are described in Table 1. The reliability versus flux density assuming two passes for each of the selection criteria is shown in Figure 1 and the same parameters are shown for three passes in Figure 2. Error bars assume false detections follow Poisson statistics. If false detections are systematic, however, then the errors given will be overestimated. For those fluxes where a small sample size leads to large error bars ($S_v >$ 10-15 mJy), we do not include the errors in Figures 1 and 2. It should be emphasized that, because of the luminosity function used in the simulations, the flux density bin with the largest sample size is near the flux cut-off in each band of the reliability curves. Thus, reliability error estimates are significantly smaller below 10 mJy than above.

The principle results for the external reliability calculations are: 1) all selection criteria achieve a reliability $\geq$99.5% in all bands except band 4 where the 2+0 criterion fails; 2) the cross band selection criteria provides greater robustness than any catalog created based on information in a single band; 3) three passes do not significantly lower the flux density limit at which a reliability of $\geq$99.5% is achieved for any of the tested selection criteria; 4) a reliability $\geq$99.5% is achieved at about the same flux density limits for selection criteria 2+2, 2+1, 3+2, and 3+1 in the absence of saturated sources and strong banding (regions around saturated sources and strong banding were not included in our simulated data).

3.2 Externally Determined Completeness

Completeness is only meaningful using the external truth table. Using an internal truth table produces unrealistically high completeness since systematically undetected sources can obviously never be included in an internal truth table. We calculated the completeness for two and three passes for each selection criterion given in Table 1 as a function of flux density (see Figure 3). The principle results of the external completeness analysis are: 1) the stricter selection criteria have lower completeness values; 2) the completeness is 98% at flux densities greater than the limit at which the reliability is $\geq$99.5% for all the two pass selection criteria; and, 3) the completeness begins to drop below 98% at about 1 mJy in all bands for the 2+1 criterion.

3.3 Creating an Internal Truth Table

Since we do not have the benefit of an external truth table for the actual OSV observations, we must determine our reliability using only measured quantities. That is, we must construct a table of true sources using only our observations (i.e. an internal truth table). For reference in constructing such a table, we present a histogram of the detection frequency of all the externally true and false sources above a canonical flux density level of 3 mJy (see Figure 4). We constructed an internal truth table such that the resulting internally determined reliability was as similar to the external reliability as possible. In order to meet this standard, we had to construct the truth table for each band separately. That is, just because a detected source was considered internally true in band 1 did not automatically make it true for band 4. The criterion for constructing an internal truth table that best matched the external reliability results was that a point source had to be detected at least 8 out of 10 times in the band under consideration or an adjacent band. Using this criterion, the internally determined reliability as a function of flux density for two passes and the selection criteria in Table 1 is given in Figure 5. The dips at high flux levels in Figure 5 are due to low sample sizes (e.g. the drop in bands 1, 2 and 3 above 10 mJy are due to 1 false detection and in band 4 is due to 2 false detections).

The main results of the internal reliability analysis are: 1) the internal reliability is $\geq$99.5% down to flux densities of about 1 mJy for bands 1, 2, and 3 and $\sim$10 mJy for band 4 for two passes and the 2+1 selection criterion; 2) the internal reliability falls off at lower flux densities than those for the

external reliability.

4 Discussion

4.1 Two Versus Three Passes

Comparison of Figures 1 and 2 shows that a GLIMPSE survey with either two or three passes can achieve a reliability $\geq$99.5% using selection criteria 2+2, 2+1, 3+2, or 3+1 in all IRAC bands; and, both three passes and two passes achieve a reliability $\geq$99.5% at about the same flux density. This leads us to conclude that there is no advantage to using three passes to achieve reliabilities $\geq$99.5%.

4.2 Selection Criteria

Two critical conclusions can be drawn from the external and internal reliability calculations: 1) The "2+1" criterion allows the GLIMPSE catalog to be $\geq$99.5% reliable at flux levels of $\sim$2-4 mJy in all bands. The "2+x" reliability is noticeably lower at most flux levels, as is expected since it is a less stringent criterion. The "2+2" criterion basically mimics the "2+1" criterion at flux levels $\geq$3 mJy, indicating that most sources are detected twice in each of two adjacent bands. Below 1 mJy the external reliability does not fall below 97-98% simply because below a certain flux limit, no point sources, true or false, meet the selection criterion of "2+1".

Comparison of external and internal reliability (see Figures 1 and 5) shows that the flux density at which the reliability is $\geq$99.5% is significantly lower for internal reliability than for external reliability. The reason for this can be seen from examination of Figure 6 where the frequency of detection in each band for sources with flux density $\geq$1 mJy. Here it is obvious that a small but significant number of false sources are detected 8 or more times and count as true in an internal truth table. The percentage of false sources relative to the true sources detected $\geq$ 8 times in bands 1, 2, 3, and 4 was about 1.4%, 1.3%, 0.5%, and 0.35%, respectively. These false detections at flux densities of a few mJy account for the lower flux density limits between external and internal reliability. The majority of these sources are accounted for by the following effect. If a true flux is found but is different from the mean measured flux by more than 1 magnitude in any band, the match is

considered poor and the source is classified as false. This sifting is necessary because the simulations involve many very faint sources (to ensure good sample size below our flux limits) which are occasionally matched against moderately bright false detections. We cannot discriminate between a false source being incorrectly identified with a faint source in the truth list and a true source whose flux has been incorrectly determined, so the conservative approach calls for classifying them all as false. This effect is strongest below 1 mJy in band 4, although it is present at flux densities between 1 and 3 mJy also. Of course, this disparity check is not possible in creating the internal truth table since no external flux densities are available. Thus, the actual flux density cutoff lies between the external and internal reliability values given here. Again, the most conservative choice of flux density cutoff is the flux density indicated by the external reliability calculation, $\sim$3 mJy.

The difference between the external and internal truth tables, however, does not negate the basic conclusions of this investigation; it only brings into question the flux density limits that should be used for the internal reliability. The conservative limits are those established by the external truth table. The basic conclusions that we believe are robust are: three passes and two passes have about the same flux density limits in all IRAC bands; a reliability $\geq$99.5 can be achieved in all four IRAC bands with two passes and selection criterion of two detections in one band and at least one detection in an adjacent band; a selection criterion of 2+2 does not produce a significantly better reliability than the 2+1 criterion; and, as the selection criteria becomes more stringent the completeness decreases, so one should use the least restrictive selection criterion that can achieve the required reliability in order to achieve the highest completeness compatible with the required high reliability. It should be emphasized that completeness and reliability are not significantly improved with any 3-visit catalog, which would significantly reduce the survey area.

The completeness is $\sim$ 97-98% at flux densities at which reliability is $\geq$99.5%. We cannot meaningfully determine the completeness using the internal truth table since, obviously, the sources that we do not ever detect are precisely the sources that do not make it into the internal truth table. That is, completeness calculated using an internal truth table is, almost by definition, always near 100%.

Lastly, artifacts such as cosmic ray contamination, areas around saturated sources, and strong banding in bands 3 and 4 have not been included in the simulations. It may be necessary to eliminate regions around saturated

stars and strongly banded regions in the observed data to achieve a catalog with reliability $\geq$99.5%. Preliminary tests at correcting for banding indicate that these regions can be recovered. The recently discovered decreases in sensitivity in bands 3 and 4 are also not included in the simulations and may modify the flux density limits for completeness and reliability somewhat. Obviously, the "2+1" criterion, because of the requirement that a source must be detected twice in one band and at least once in an adjacent band will eliminate essentially all cosmic ray hits from the GLIMPSE catalog.

5 Appendix A

To calculate the reliability of the GLIMPSE catalog from a simulation of 10 possible observations of each source, we first calculate the probability of a detected source meeting the catalog requirements (for background, see a textbook on introductory combinatorics, e.g. "Applied Combinatorics" by Tucker). Assume a source is detected d_i times out of 10 possible exposures in band i. First we calculate the probability (P_i) of a source being detected in a single exposure in a single band:

$$P_i = \frac{d_i}{10} \tag{1}$$

Next, we specify the different scenarios that will allow a source to be cataloged. Each line of Table 2 corresponds to a different way a source can make it into the GLIMPSE catalog for two different selection criteria (2+1 and 2+x). Furthermore, a source cannot make it into the GLIMPSE catalog without matching one and only one line in the table.

Table 2: Scenarios that lead to inclusion in the catalog

Criterion: 2+1			
	Detections in		
Band 1	Band 2	Band 3	Band 4
Criterion: 2+1			
=2	≥1	any	any
=1	=2	any	any
=0	=2	≥1	any
=≤1	=1	=2	any
=≤1	=0	=2	≥1
=≤1	≤1	=1	=2
Criterion: 2+x			
=2	any	any	any
≤1	=2	any	any
≤1	≤1	=2	any
≤1	≤1	≤1	=2

We can now calculate the probability of each scenario (see Table 3).

Table 3: Probabilities for Selection Criteria 2+1 and 2+x

	Band 1		Band 2		Band 3		Band 4
			Criterion: 2+1				
	P_1^2	$\times$	(1-(1-P_2^2))	$\times$	1	$\times$	1
+	2*P_1 (1-P_2)	$\times$	P_2^2	$\times$	1	$\times$	1
+	$(1-P_1)^2$	$\times$	P_2^2	$\times$	$(1-(1-P_3)^2)$	$\times$	1
+	1-P_1^2	$\times$	2*P_2	$\times$	(1-P_3)	$\times$	1
+	1-P_1^2	$\times$	2*P_2	$\times$	(1-P_3)	$\times$	P_1^2
+	1-P_1^2	$\times$	$(1-P_2)^2$	$\times$	P_3^2	$\times$	(1-(1-P_4^2))
+	1-P_1^2	$\times$	1-P_2^2	$\times$	2*P_3 (1-P_3)	$\times$	P_4^2
			Criterion: 2+x				
	P_1^2	$\times$	1	$\times$	1	$\times$	1
+	1-P_1^2	$\times$	P_2^2	$\times$	1	$\times$	1
+	1-P_1^2	$\times$	1-P_2^2	$\times$	P_3^2	$\times$	1
+	1-P_1^2	$\times$	1-P_2^2	$\times$	1-P_3^2	$\times$	P_4^2

To calculate the probability of a source fulfilling the GLIMPSE catalog criterion, we multiply all the terms in each line, then take the sum of all the lines. The equivalent scenarios and probabilities for the other catalog criteria are given below.

Table 4: Scenarios for Inclusion and Associated Probabilities for 2+2

	Detections in Band 1		Band 2		Band 3		Band 4
			Scenarios				
	=2		=2		any		any
	$\leq$1		=2		=2		any
	any		$\leq$1		=2		=2
			Probabilities				
	P_1^2	$\times$	P_2^2	$\times$	P_3^2	$\times$	1
+	1-$P_1^2\times$	$\times$	$P_2^2\times$	$\times$	P_3^2	$\times$	1
+	1	$\times$	1-P_2^2	$\times$	P_3^2	$\times$	P_4^2

Table 5: Scenarios for Inclusion and Associated Probabilities for 3+1

	Detections in Band 1		Band 2		Band 3		Band 4
			Scenarios				
	$=3$		≥ 1		any		any
	$=1$ or 2		$=3$		any		any
	$=0$		$=3$		≥ 1		any
	≤ 2		$=1$ or 2		$=3$		any
	any		$=0$		$=3$		≥ 1
	≤ 2		≤ 2		$=1$ or 2		$=3$
	$=3$		$=0$		$=1$ or 2		$=3$
			Probabilities				
	P_1^3	$\times$	$1\text{-}(1\text{-}P_1)^3$	$\times$	1	$\times$	1
$+$	$1\text{-}(P_1^3+(1\text{-}P_1)^3)$	$\times$	P_2^3	$\times$	1	$\times$	1
$+$	$(1\text{-}P_1)^3$	$\times$	P_2^3	$\times$	$1\text{-}(1\text{-}P_3)^3$	$\times$	1
$+$	$1\text{-}P_1^3$	$\times$	$1\text{-}(P_2^3+(1\text{-}P_2)^3)$	$\times$	P_3^3	$\times$	1
$+$	1	$\times$	$(1\text{-}P_2)^3$	$\times$	P_3^3	$\times$	$1\text{-}(1\text{-}P_4)^3$
$+$	$1\text{-}P_1^3$	$\times$	$1\text{-}P_2^3$	$\times$	$1\text{-}(P_3^3+(1\text{-}P_3)^3)$	$\times$	P_4^3
$+$	P_1^3	$\times$	$(1\text{-}P_2)^3$	$\times$	$1\text{-}(P_3^3+(1\text{-}P_3)^3)$	$\times$	P_4^3

Table 6: Scenarios for Inclusion and Associated Probabilities for 3+2

	Detections in Band 1		Band 2		Band 3		Band 4
				Scenarios			
	$=3$		≥ 2		any		any
	$=2$		$=3$		any		any
	≤ 1		$=3$		≥ 2		any
	≤ 2		$=2$		$=3$		any
	any		$=\leq 1$		$=3$		$=\geq 2$
	≤ 2		≤ 2		$=2$		$=3$
				Probabilities			
	P_1^3	$\times$	$P_2^3+3*P_3^2*(1-P_3)$	$\times$	1	$\times$	1
$+$	$3*P_1^2*(1-P_1)$	$\times$	P_2^3	$\times$	1	$\times$	1
$+$	$(1-P_1)^3+3*P_1*(1-P_1)^2$	$\times$	P_2^3	$\times$	$P_3^3+3*P_3^2*(1-P_3)$		
$+$	$1-P_1^3$	$\times$	$3*P_2^2*(1-P_2)$	$\times$	P_3^3		
$+$	1	$\times$	$(1-P_2)^3+3*P_2*(1-P_2)^2$	$\times$	P_3^3	$\times$	$P_4^3+3*P_4^2*(1-P_4)$
$+$	$1-P_1^3$	$\times$	$1-P_2^3$	$\times$	$3*P_3^2*(1-P_3)$	$\times$	P_4^3

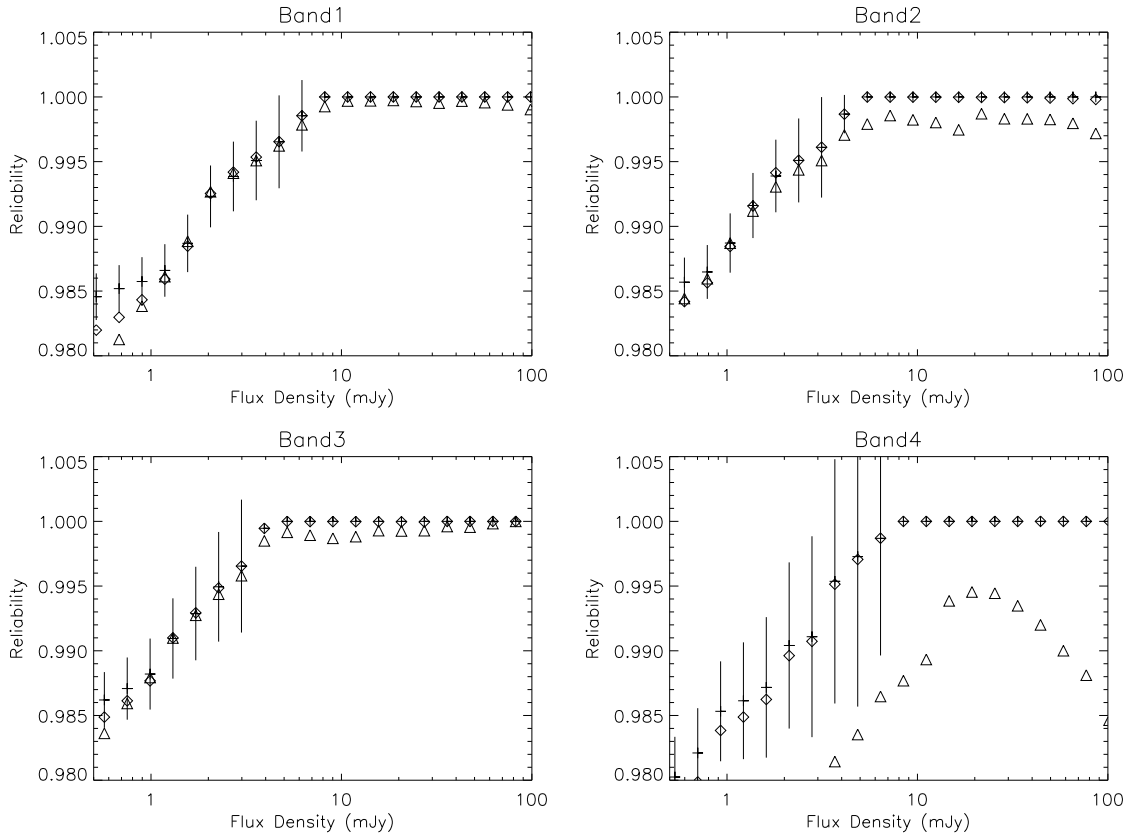

Figure 1: The reliability of a 2-visit GLIMPSE catalog using an external truth table and three different catalog criteria. The catalog criteria are 2+2 (+), 2+1 ($\Diamond$) and 2+x ($\triangle$), as described in Table 1.

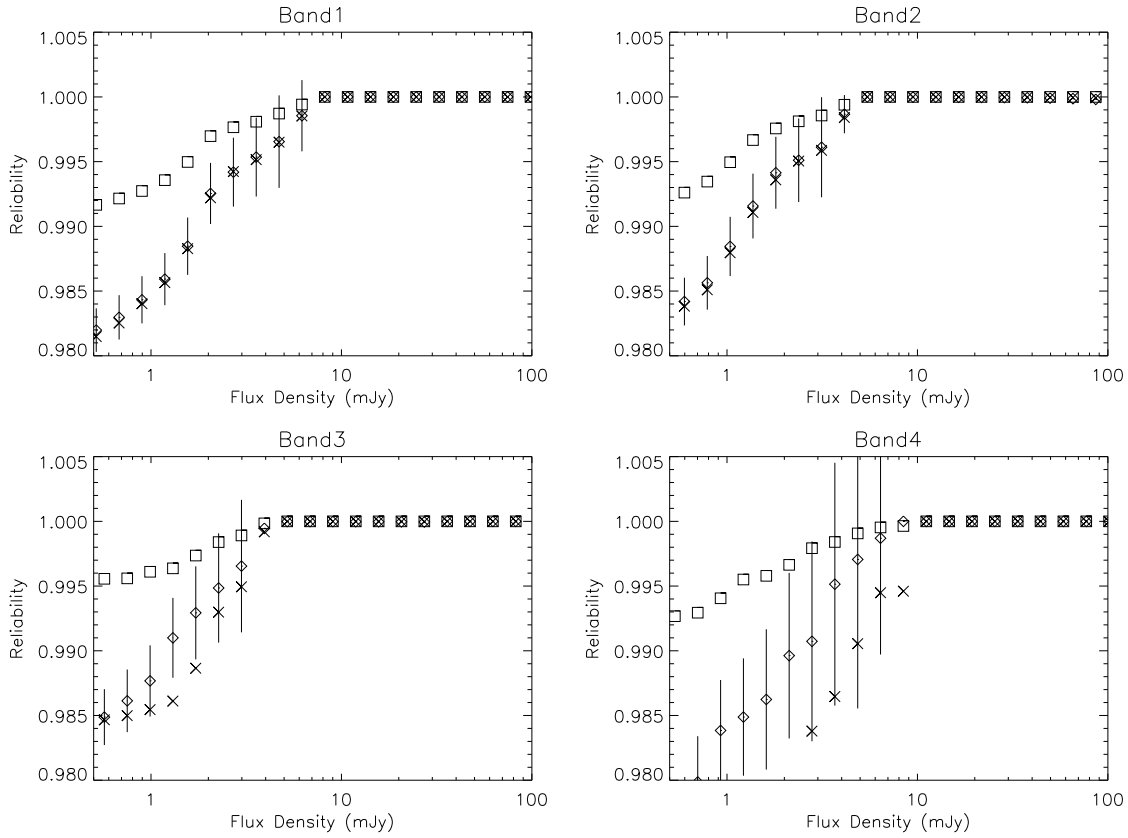

Figure 2: The reliability of a 3-visit GLIMPSE catalog using an external truth table and three different catalog criteria. The catalog criteria are 3+2 (□), 3+1 (X) and 2+1 (◊), as described in Table 1.

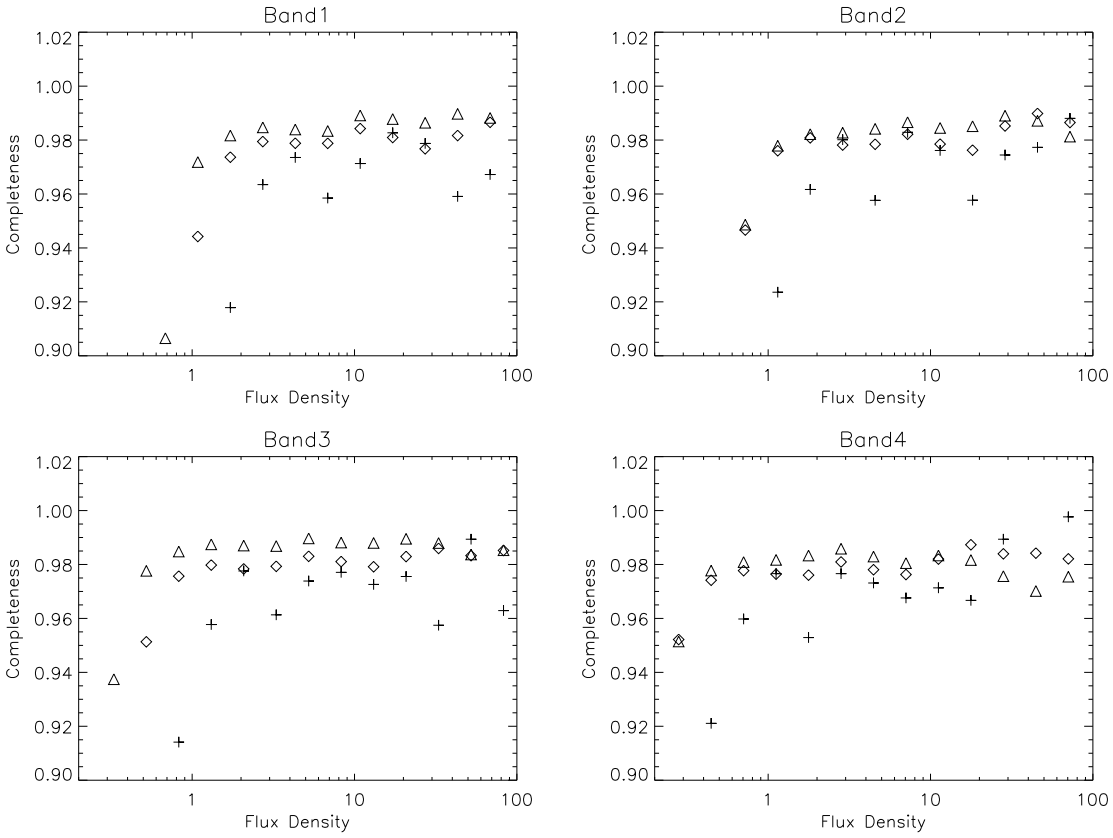

Figure 3: The completeness of a GLIMPSE catalog using an external truth table and three different catalog criteria. The catalog criteria are 2+2 (+), 2+1 ($\Diamond$) and 2+x ($\triangle$), as described in Table 1.

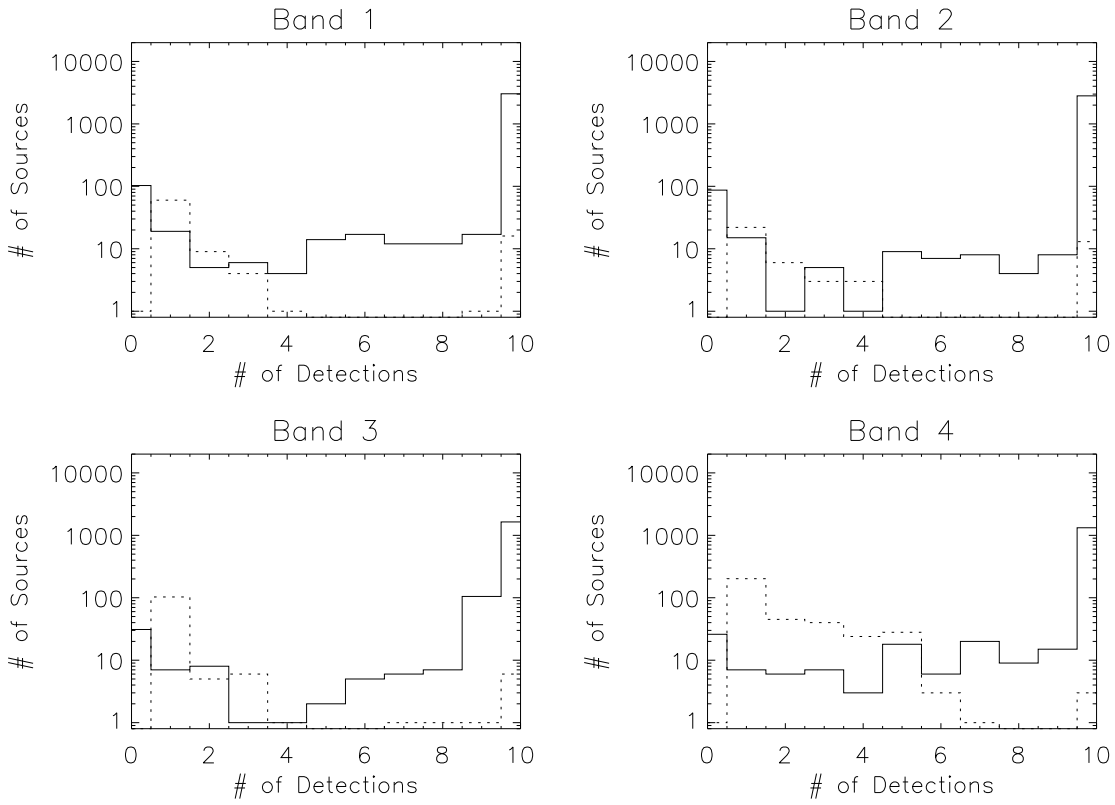

Figure 4: The frequency of source detection for true (solid) and false (dotted) sources. Frequency was normalized to be out of 10 possible detections, as is common for most of the simulated area. All sources with S_ν >3 mJy are included. *Note the log scale; the vast majority of true sources are detected 10/10 times*

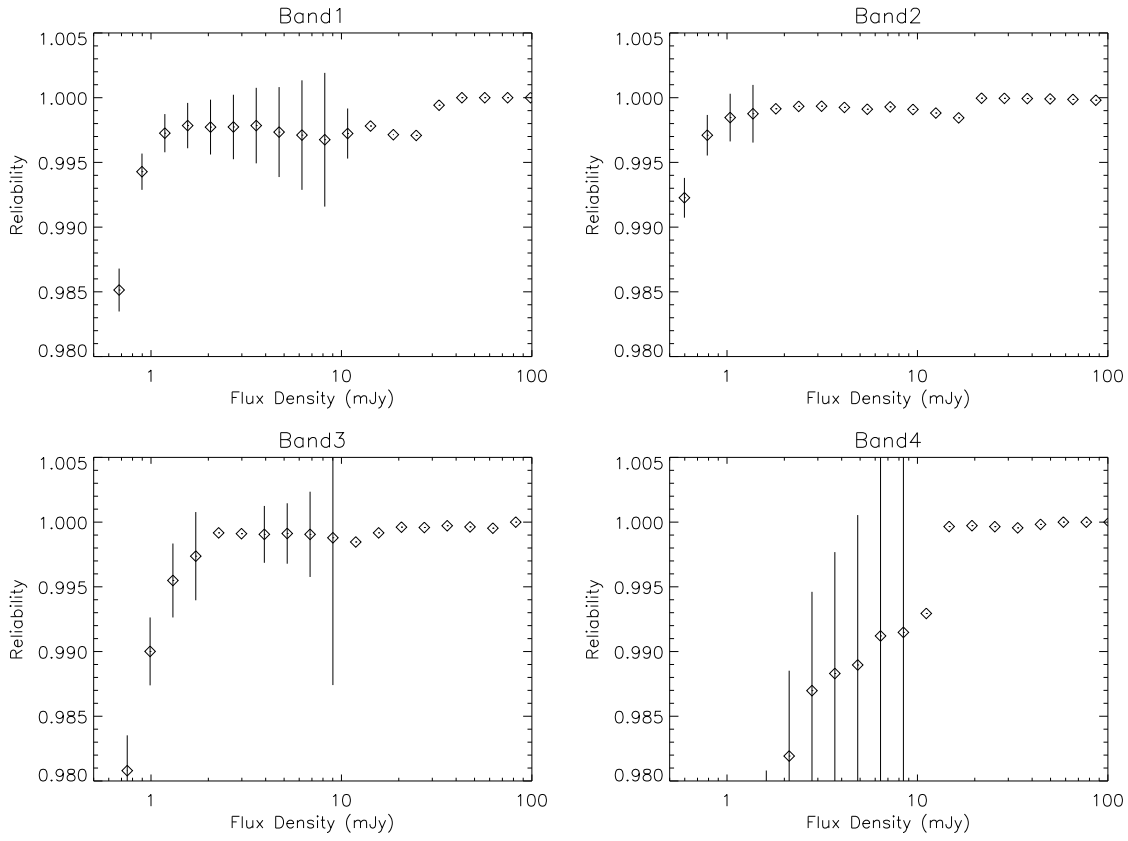

Figure 5: The reliability of a GLIMPSE catalog using an internal truth table. The internal truth table was constructed band-by-band and contained all sources detected at least 8/10 times in the current band or at least 8/10 times in an adjacent band.

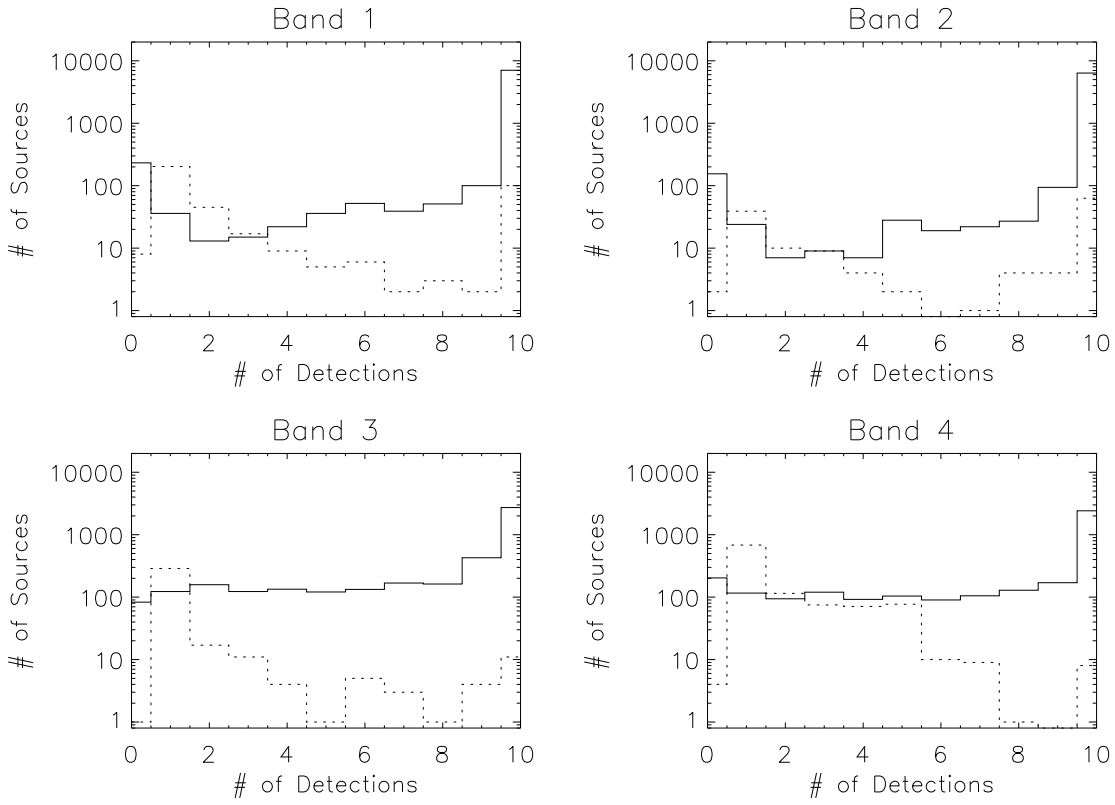

Figure 6: The frequency of source detection for true (solid) and false (dotted) sources. Frequency was normalized to be out of 10 possible detections, as is common for most of the simulated area. All sources with S_ν >1 mJy are included.