ALFA Memo 040702: Confirmation of Marginal Detections and Optimal Sampling Strategy for ALFA Drift Surveys

R. Giovanelli – 02 July 2004

1. Introduction

ALFALFA will produce a catalog of tentative detections of tens of thousands sources. The number of tentative detections per signal–to–noise (S/N) bin will of course increase steeply, and the detection reliability decrease, with diminishing S/N value.

Internal (i.e. within the survey data set) corroboration of tentative detections will rely on (a) comparison of independent polarization samples and (b) comparison of spatially adjacent survey samples. The effectiveness of part (b) will depend on the spatial sampling density and, in the case of multiple drifts through the same region, on the temporal consistency of the data. These comparisons will help exclude marginal detections with lower probability of confirmation.

Post–survey, corroborating observations will be necessary to confirm tentative detections of lower S/N. What S/N limit should be adopted for follow–up observations? Too high a S/N threshold may imply losing many valuable potential detections; too low a threshold may require impractical amounts of telescope time; a haphazard criterion can damage the completeness of acquired samples. Moreover, too sparse a set of follow–up targets may lead to a follow–up observing campaign of low on–off source duty cycle. Careful optimization will be required.

Using the results of survey simulations described elsewhere (Giovanelli 2003) and the detection reliability simulations carried out by Saintonge (2004), the issues described above are quantitatively considered.

2. Candidate Detections

Suppose an automated source–finding algorithm is applied that flags features in each spectrum of the survey, which are identified as detection candidates. Those candidates can be of three classes: (i) real sources, (ii) statistical fluctuations and (iii) spurious signals due to rfi or other instrumental and data analysis causes.

Real Sources. Figure 1 shows the distribution of simulated sources, plotted against mean S/N, as expected to be obtained by a Dec 0 to 36 degree ALFALFA survey (t_{int} = 12 sec), using the Rosenberg & Schneider (2002; RS02) HI Mass function (HIM), $h = 0.7$, and the cosmic density field as obtained from PSCz and a normalization as required by the peculiar velocity field data. The S/N is estimated for each source as the ratio between the peak signal flux and the r.m.s in the spectrum (note that a matched–filter detection algorithm will depend little on the smoothing applied to the

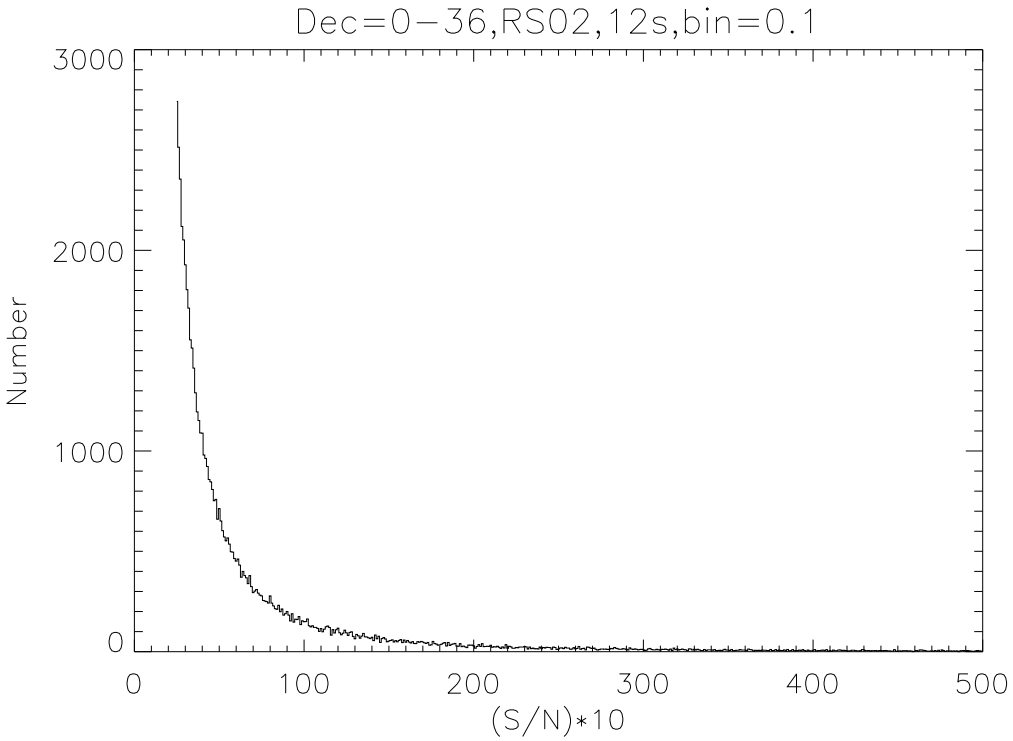

Fig. 1.— S/N histogram of a 12s ALFALFA survey using the RS02 HIM. S/N bins have width 0.1 in S/N. S/N is defined as the peak signal flux to the rms, computed in matched filter mode, over a spectral resolution equal to 1/2 the signal width. Only tentative detections with S/N> 2.5 are plotted.

input spectrum), computed after the spectral resolution has been matched to 1/2 the width of the simulated source. The "crowding" of sources at the low S/N end is apparent. The application of an automated source–finding algorithm will "detect", i.e. deliver detection candidates, with an efficiency that will depend on the nominal S/N. By "nominal" S/N is intended the average S/N of a large number of possible observations of the same source. A single simulated trial of a source of flux S will not necessarily deliver the nominal S/N; a fraction of the time the given trial will deliver a S/N lower than the nominal one, and at low values of S/N the signal extraction algorithm will not detect the source.

Statistical Fluctuations. Assuming Gaussian noise, the probability that a single spectral channel yield a fluctuation of S/N between s_1 and $s_1 + ds_1$ is

$$p_1 ds = \frac{1}{\sqrt{2\pi}} e^{-s_1^2/2} ds_1 \tag{1}$$

where $s_1 = S/\sigma_1$, with S the peak flux density and σ_1 the r.m.s. noise, with single–channel spectral resolution. Similarly, the probability for a n_w channels–wide spectral feature to exhibit a deviation

between s_n and $s_n + ds_n$ is

$$p_n ds = \frac{1}{\sqrt{2\pi}} e^{-s_n^2/2} ds_n \tag{2}$$

where $s_n = s_1/\sqrt{n_w}$.

In a survey of N_{los} line of sight samples, taken with a spectrometer of N_c channels, the number of samples n_w channels wide, exhibiting a S/N between s and $s + ds$ is

$$n_{s,n_w} ds = N_{los} \frac{N_c}{n_w} p_n ds = N_{los} \frac{N_c}{n_w} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds \tag{3}$$

and the total number of statistical fluctuations of that width exhibiting a S/N larger than a threshold s_{th} is

$$N_{s_{th},1} = N_{los} \frac{N_c}{n_w} \frac{1}{\sqrt{2\pi}} \int_{s_{th}}^{\infty} e^{-s^2/2} ds = N_{los} \frac{N_c}{n_w} [F(\infty) - F(s)] \tag{4}$$

where, again, the S/N s is meant to be computed over a spectral resolution of $n_w/2$ channels, and $F(s)$ is the familiar cumulative distribution of the normal error function: $F(-\infty) = 0$, $F(\infty) = 1$ and $F(0) = 0.5$. The total number of purely statistical fluctuations with S/N between s_a and s_b, with widths between n_{w1} and n_{w2}, appearing in the survey will then be

$$N_{[a,b],[1,2]} = \sum_{n_w = n_{w1}, n_{w2}} N_{los} \frac{N_c}{n_w} [F(s_a) - F(s_b)] \tag{5}$$

To first order, and ignoring the fact that in the expression above a high S/N, broad feature gets overcounted as several, lower S/N, narrower ones, we can approximate

$$N_{[a,b],[1,2]} = N_{los} N_c \ \ln \frac{n_{w2}}{n_{w1}} \ [F(s_a) - F(s_b)] \tag{6}$$

For a survey covering all sky between 0 and 36 degrees, with a beamwidth of 3.5', the number of independent spatial samples is $N_{los} = 1.8 \times 10^7$, the number of useful spectral channels is $N_{ch} \simeq 3700$ and the width limits (30 to 500 km s^{-1}) yield $n_{w1} = 6$ and $n_{w2} = 100$, so that $N = N_{[a,b],[1,2]} \sim 1.9 \times 10^{11} [F(s_a) - F(s_b)]$. This yields $N_{>3} \sim 3.5 \times 10^8$ features with $s > 3$, $N_{>4} = 9.1 \times 10^6$ features with $s > 4$, and $N_{>5} = 9.1 \times 10^4$ features with $s > 5$, with width anywhere between 30 and 500 km s^{-1} .

Rfi or other Spurious Signals. We will ignore this component at this time, although it is clear that the non–Gaussian behavior of the statistics associated with these features should advise for independent observations, if possible at different epochs spaced 3 to 9 months.

Figure 2 shows the relative distribution of real and spurious features, per bin of width 0.1 in S/N. The thick solid line is the distribution of candidates of type (i), the 'real' sources as produced by the simulation (same data as in Figure 1). The other lines identify spurious candidates, features of type (ii) resulting from statistical fluctuations in the data. The highest of the lines yields the

expectations according to Eqn. 6. It is clear that for S/N lower than about 6, spurious signals outnumber real sources. At $S/N \simeq 4$ they do so by 3.5 orders of magnitude.

To some degree, the spurious nature of a feature can be ascertained by comparing the spectrum in which it appears with spectra taken in adjacent positions or at the same position but at different times. Comparison of subsets of the spectrum under consideration, such as independent polarization channels, allows for further discrimination.

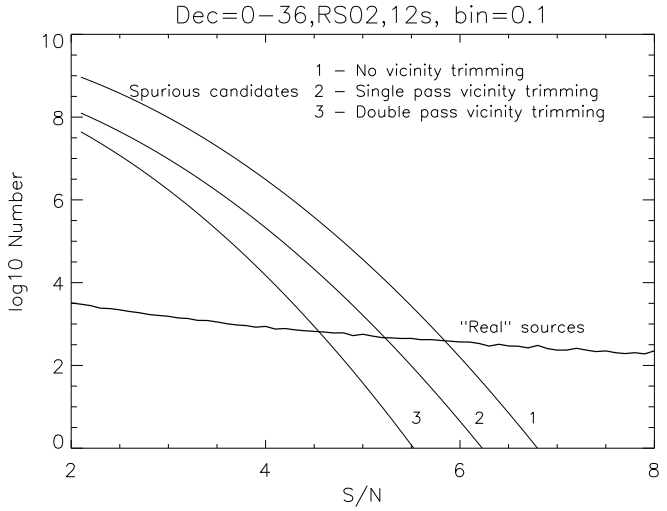

Fig. 2.— .Distribution of candidate detections, as a function of S/N, in bins of 0.1. The thick curve contains the same data as displayed in Figure 1, and it refers to the sources expected in an all–Arecibo sky survey, for a RS02 HIM. The highest of the thin lines, shows the number of expected spurious sources, resulting from statistical fluctuations, for the same survey. The solid curve labelled '2' represents the remaining candidates after comparing adjacent sky tracks in a single-pass survey (spacing 2.1'), and the solid curve labelled'3' represents the remaining candidates after comparing adjacent sky tracks in a 2–pass survey (track spacing 1.05').

Consider a single–pass ALFA survey, whereby contiguous feed tracks are separated by 2.1' in the sky. A point source swept by one of the feeds will appear also in contiguous tracks, at lower S/N. For a HPFW of the 21cm beam of 3.5', the response 2.1' off the beam center is about 0.37 of that on–beam (if a strategy of 4 ALFA drift sweeps per degree are adopted, every 7–th track the gap increases from 2.1' to 2.4' and the corresponding response 2.4' off–axis is 0.27 of that on–axis). If we exclude all candidate features for which adjacent pixels yield signals of less than some threshold S/N, we can significantly trim the number of spurious signals. We thus ask: what is the probability that, given a spurious signal of $S/N = s$ at a given l.o.s. (call it the 'reference pixel'), a random fluctuation will occur yielding a signal of $S/N \geq s'$ over the same spectral range, at an independently sampled adjacent location? We need to set the S/N threshold. Consider a real source at $S/N = s$ centered on the reference pixel; the S/N at a pixel removed 2.1' from the reference position, on either of two adjacent tracks, will average $0.37s$. Either because the source is not exaclty centered on the Declination of the track with the reference pixel, or due to statistical

fluctuations, we should allow for either adjacent pixel to yield a $S/N < 0.37$. Assume that for a real source the signal ratio between the two adjacent pixels of the reference pixel can vary within up to a factor of 2. We can then exclude all candidate features for which there is no feature with $S/N = s' > 0.17s$ on each of the two adjacent pixels. The probability of such a feature appearing by chance on each of the two adjacent pixels is $[F(\infty) - F(s')]^2$. By excluding candidate features for which adjacent pixels do not *each* exhibit a spectral feature brighter than $S/N = s' > 0.17$, the thin solid expectation labelled '1' in Figure 2 reduces to the thin solid line labelled '2'. The crossover between spurious and real sources now occurs at $S/N \simeq 5.3$.

Suppose the sky is sampled in a two–pass drift strategy, whereby the second pass is shifted 1.05' in Dec. with respect to the first. The off–axis response of a 3.5' Gaussian beam at 1.05' is 0.78 of that on–axis. By the same reasoning followed above, we can exclude all candidate detections of$S/N = s$ which do not exhibit a signal with $S/N = s' > 0.36s$ in each of two adjacent pixels in contiguous tracks. This exclusion reduces the highest of the thin solid lines in Figure 2 to the lowest (labelled '3'). After this exclusion, the crossover between real and spurious sources occurs near $S/N \simeq 4.5$.

Inspection of each polarization channel can further help exclude spurious candidates. This is particularly useful with non–Gaussian noise or rfi. In the case of normally distributed noise, however, the inspection of the two polarization channels (or, for that matter, the inspection of any two independent 'halves' of the total spectrum, such as the first 6-sec half and the second 6-sec half of the drift across a given position) provides little discriminant power between a spurious and a real signal. Given the *prior* that the total signal yields a feature of $S/N = s$, the way to distinguish between a spurious feature and one arising from a real source would be that of checking the ratio of the feature in the two polarization channels. That ratio will average 1 in either case, but it will have a larger dispersion if the feature is spurious. For example, any two independent halves of a feature of $s = 2$ in the total spectrum will have a ratio smaller than a factor 2 about half the time, while a real source with $s = 3$ will exhibit a ratio 2 between two of its halves about 3/4 of the time. As the S/N increases, the difference (between real sources and spurious ones) in the distribution of 'halves ratio' diminishes. The impact of this kind of test is significantly smaller that that between adjacent pixels, and we currently ignore it.

3. Reliability of Candidate Detections and Signal Extraction

Because of the massive bulk of ALFA data sets, signal identification will rely on automated procedures. Amelie Saintonge has coded a matched–filter cross–correlation signal extraction algorithm, described elsewhere (Saintonge 2004), which currently applies to single spectra, but for which an extension to apply to data cubes — so that the matched–filter idea can be applied to the spatial domain — is under development. Simulations have been carried out with the Saintonge algorithm, by injecting in a random manner a signal in a noisy spectrum, and monitoring the ability of the signal extraction algorithm to identify the injected signal. The "detection" probability is

computed as the fraction of all trials that the signal extraction algorithm positively identifies the injected signal. It is monitored as a function of S/N and of signal width. When the S/N is expressed in terms of the noise estimated by averaging over 1/2 the spectral width of the injected signal, the detection probability is largely independent on the signal width. Such a function is shown in Figure 3.

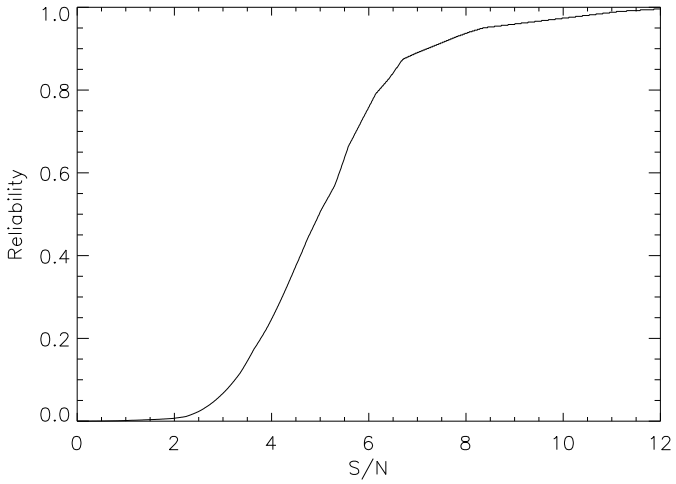

Fig. 3.— Detection reliability of a Gaussian feature, plotted vs. S/N ratio. The noise is computed as the rms at the spectral resolution equal to 1/2 of the HPFW of the Gaussian.

We shall assume that the effectiveness of the algorithm applies equally to real and spurious signal candidates of given S/N.

Thus, the products of the curve in Figure 3 with those in Figure 2 yields the number of detection candidates the survey will deliver. The result is shown in figure 4.

The lower, thick curve (nearly flat) shows the distribution of expected 'real' sources that will be candidate detections, as a function of S/N. Between $S/N = 4$ and $S/N = 6$, there are about 400 sources per 0.1 wide bin. In the case of pure Gaussian noise, absence of rfi and other baseline distortions such as standing waves, the vast majority of detection candidates down to a S/N threshold of 6.5 would be real sources. Through follow–up observations with a single–pixel receiver, other real sources can be recovered. Pushing the detection threshold of an all–Arecibo sky single pass drift survey from 6.5 down to 5.5, for example, would (for an RS02 HIMF) deliver 4000 additional sources, albeit an important fraction of the candidates would be revealed as spurious. At lower S/N, the chase of candidates by follow–up observations would encounter diminishing returns. In the real world of imperfect hardware, standing waves and rfi environment, the S/N limits to which detection can realistically be pushed will be less optimistic.

Figure 5 shows the same data as shown in Figure 4, except that the vertical scale is linear and the curves yield cumulative numbers, i.e. the number of features with S/N larger than the value given along the horizontal axis. Assume an ideal survey where noise is purely Gaussian, covering

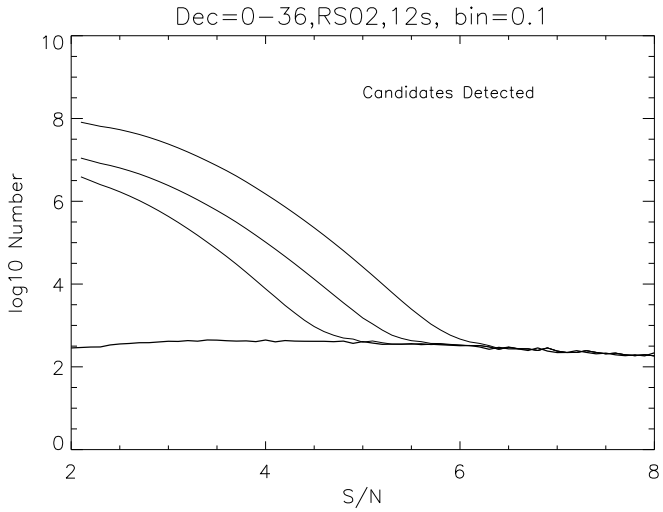

Fig. 4.— Number of candidate detections per 0.1 S/N bin, expected for an all–Arecibo sky drift survey. The lower (nearly flat) line refers to the 'real' sources (assuming an RS02 HIM) while the upper curves refer to the cases described in the caption to Figure 2, after multiplication by the curve in Figure 3.

the whole Arecibo sky (12,000 square degrees) and populated by sources as indicated by the RS02 HIMF. Virtually all features with $S/N > 6$, about 17,000 in number, would be real sources.

If the threshold were set at $S/N = 5$, an additional ~ 4200 real sources would be included, but they would have to be separated from spurious features. In the absence of "vicinity triming" as described above, adding those 4200 sources to the catch would require follow–up observations of about 65,000 features, requiring over 100 hours of telescope time (as we discuss in the next Section). If the set of candidate features were trimmed by comparison with adjacent spectra of a single–pass drift survey, half of the follow–up observations would yield confirmation of real signals. In a double–pass drift survey, most follow–up observations would deliver positive detections.

If the threshold were set at $S/N = 4$, an additional 4100 real sources would be added to the catch, for a total near 25,000. However, their confirmation would require the observation of a large fraction of spurious features. Without vicinity trimming, the number of follow–up observations would be unrealistically large (see curve labelled '1' in Figure 4). The same would be true even after vicinity trimming in a single–pass drift survey (tracks separated by 2.1' in Declination). After vicinity trimming in a double–pass (Declination tracks separated by 1.05') drift survey, however, two in seven of the features that would require follow–up observations ($\sim 28,000$) would deliver a positive detection; about 500 hours of follow–up telescope time would yield 8000 additional real sources.

We re–iterate that this is an ideal case, and near the S/N limits discussed above a number of perturbing sources will have a negative impact. Assuming, very coarsely, that those perturbing sources will shift curves in Figure 5 to the right by one unit in S/N, it seems reasonable to expect

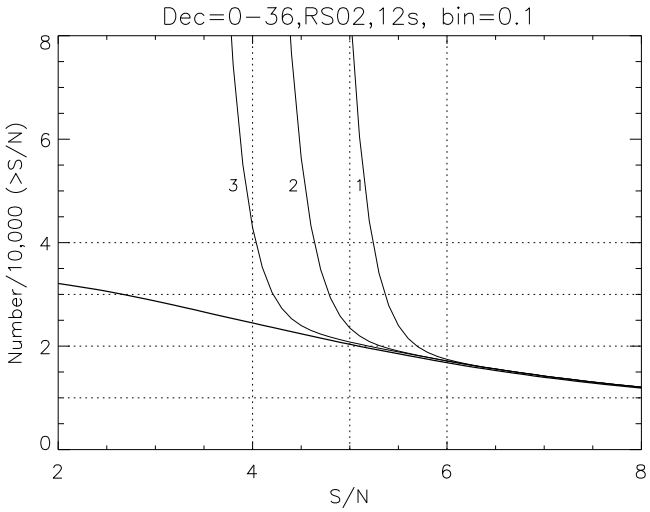

Fig. 5.— Cumulative number of candidate detections as a function of S/N, expected for an all–Arecibo sky drift survey. The lower (nearly flat) line refers to the 'real' sources (assuming an RS02 HIM) while the upper curves refer to the cases described in the caption to Figure 2.

that an all Arecibo sky drift survey of an RS02 HIMF Universe would yield about 14,000 sources, which could increase to approach 20,000 with a relatively modest amount of a few hundred hours of follow–up, single pixel observations. Clearly, follow–up observations should be requested only for candidates of S/N such that the expected return (confirmed sources per unit time of follow–up obervations) is equal to or higher than that of the overall survey.

4. Follow–up Observations

A follow–up, corroborating observation should deliver a S/N that will make the detection reliable. For a tentative detection with $S/N \simeq 5$, the corroborating observation should deliver $S/N > 6.5$, according to Figure 4. In the real world, we may demand a safer figure, say 7 to 8. Thus t_{int} for the corroborating observation should be about 30 sec ($\sqrt{30/12} \sim 1.6$, where 12 sec is a drift survey's single pass t_{int}).

Short observations in staring mode are impractical, given the slew overhead. The total time required by follow–up, corroborating observations will depend not only on the t_{int} required for each observation, but also on the sky density of the candidates. If the tentative detections to be checked are very sparse — say one every several square degrees —, slew times will be very substantial and bandpass–correcting observations will be required for each candidate source, more than doubling the required telescope time. In that case, the on–source t_{int} of ~ 0.5 minute becomes a small fraction of the overall time required to observe each source. The Arecibo telescope slew times are respectively 0.4° s^{-1} in azimuth and $0.04\ ^\circ$ s^{-1} in elevation. A 1° change in elevation will require 25 seconds. It will thus be observationally advantageous if the sky density of tentative sources to

be corroborated is high, e.g. on order of one per square degree or higher. Not only that will reduce the overhead of slew motions and settle time, but it will also allow for a running mean bandpass to be accumulated over a few contiguous staring observations, as the telescope configuration would change little between adjacent source candidates. In that case, allowing for slew and settle time, a single source follow–up observation will require on the order of one minute of telescope time.

5. Single Pass or Double Pass Drift Survey?

A single–pass drift survey, with ALFA at the optimal angle to equalize separation between contiguous beams, will produce constant Declination tracks spaced 2.1'. About four ALFA drifts, i.e. 28 tracks, will be required to cover one degree. Thirty degrees in Declination will require about 2900 hours of telescope time. Since the low latitude regions will be done commensaly with pulsar and other galactic surveys, a single pass drift survey will require somewhat less, about 2500 hours. A double pass survey will double the required time, if the sky area is maintained constant. Such a request is not very reasonable. Reduction of the sky area to be covered would be required. Let's consider pros and cons.

The volume sampled for any given HI mass is proportional to the solid angle covered and to the integration time per sample to the 3/4 power. For a fixed amount of telescope time allocated to the survey, doubling the sampling time per unit area and reducing the solid angle covered by 2, reduces the survey volume by $2^{-1/4} \simeq 0.84$, if the unit slid angle is covered with identical efficiency. Thus, in principle multiple passes are not desirable, unless a specified minimum mass limit needs to be reached.

There are however other considerations that make a double pass survey worth considering. First and perhaps foremost, it provides the ideal check against a variety of spurious sources, especially of instrumental or rfi origin. It also makes the detection of continuum source transients possible, allowing useful commensal options. It does, in addition, provide for better "vicinity trimming", as discussed in previous sections of this report. The 2.1' Declination track separation of a single pass is sub–Nyquist, and thus yields a "scallopped" sensitivity coverage of the sky, in the Declination direction. The amplitude of the scallopping is however not large: measured in terms of the S/N that a source of given flux would yield, if ideally swept across the Declination coverage of the one–pass survey, the amplitude of the variations would be less than 2%, peak–to–peak, with a spatial period of 2.1' (the track separation). This amplitude is significantly smaller than the variations in gain between the central and the peripheral feeds of ALFA. This reason alone would not be sufficient to justify a double–pass strategy.

A double–pass survey does, moreover, provide much better discrimination of real from spurious signals at low S/N, as discussed earlier in this report, allowing to push down the S/N threshold for follow–up observations by about 0.5, potentially yielding an increase of $\simeq 10\%$ in the number of real detections that can be extracted per unit volume surveyed. This offsets by more than half the

$2^{-1/4}$ loss of volume surveyed.

In summary, the pros and cons of a double–pass strategy are:

- + Increased ability to discriminate between spurious and real signals at low S/N levels
- + Increased ability to centroid, to map extended sources
- + Increased ability to discriminate the effects of rfi and instrumental instabilities
- + Opens up possibility to identify transients; commensality will play a big role in allocation of survey time
- – Loss of a few percent in the total number of detections
- – Loss of solid angle coverage: hard to expect full sky coverage to be possible

A double–pass strategy appears appealing.