Systematic Errors in "Standard" Background Estimates of Milagro Data

T.J. Haines
LANL, December 14, 1999

Motivation

In thinking about the methods that we have developed over the years to perform our background estimates for source searches, I have been troubled for some time now by the question of how large might the systematic errors be in these background estimates. In fact, since we first began proposing Milagro, we have always known that the signal to background for known sources was expected to be very small, e.g. $\sim10^{-3}$ for the Crab, and as small as $\sim10^{-5}$ for the galactic plane. We knew that the expected data rate from Milagro would *statistically* allow us to reach these levels of sensitivity. I have always been amazed that we have not been questioned about the *systematic* problems with getting a background estimate this small. I guess I should be thankful that no reviewer was smart enough to bother us about this, because we would not have had an answer!

In any event, the purpose of this note is to begin to lay some groundwork for a discussion of systematic errors in our measurements. In the distant past, these were addressed by a variety of means. Regarding systematics in background estimates, various Monte Carlo calculations were performed to show that the method was free of systematic to the necessary level. False signals were inserted into the analysis to study in detail the lost efficiency due to the background method, mostly because of the overestimation of the background due to the "bootstrap" methods used at the time. Similar Monte Carlo or other analyses have not been performed to demonstrate the current analysis methods used on Milagro/Milagrito data are valid. The remarks and examples below may or may not apply to analyses performed to date; I hope they are kept in mind as we design new analysis methods to delve down into the background and mine the signals there.

Description of Some Methods of Background Estimation

The problem of course is that we have no way to "dead reckon" the expected background level in any portion of the over-head sky. Using the rather symbolic language of the source techniques paper [D.E. Alexandreas, NIM A328, 570(1993)], the number of backgorund events expected to fall within a bin (in some coordinate system, here not to be specified) is

$$Nb = \iiint \varepsilon \bullet R(\vartheta,\varphi,t)\,d\vartheta d\varphi dt \qquad \text{(Eq. 1.)}$$

where the integral is taken over the path of the (here not specified) source through the local sky. The local coordinates are obviously θ (zenith angle), φ (azimuth) and t (time).

The variable ε represents the fact that the integral is taken over the path that the celestial source takes through the local sky.

The method I first developed, that wsa used for the "all sky" analysis [D.E. Alexandreas, Astroph. J 383, L53(1991)], performed this integral via a Monte Carlo integration, in particular by assuming that R(θ,φ,t) was separable into R(θ,φ,t) = E(θ,φ) * R'(t). This was the first of various "bootstrap" methods, where we use the observed data are in some way randomized to represent the expected background. Other important assumptions are that E(θ,φ) could be measured by binning the data in local coordinates (i.e. by using the observed data to determine the background rate, with small contribution from the source itself), and it is a constant for a period of time (~12 hrs.) except for overall rate variations that are described by R'(t). The Monte Carlo integration was performed by a random re-assignment of local coordinates to each event, with the new coordinates chosen from E(θ,φ) and the new coordinates and time determining the new celestial coordinates for the event. Each "fake" event was assigned a completely new right-ascension and declination, thus perhaps this method should be termed "coordinate swapping." This procedure was repeated 10 times for each event observed by Cygnus.

Another more generally employed method for background estimation via Monte Carlo integration of eq. 1. was developed later by W.P. Zhang and used in the "daily excess" [D.E. Alexandreas, Astroph. J. 405, 353(1993)] paper and others. In this method, eqn. 1 is better re-expressed as the following:

$$Nb = \iiint \varepsilon \bullet R(ha, \delta, t) d(ha) d(\delta) dt$$

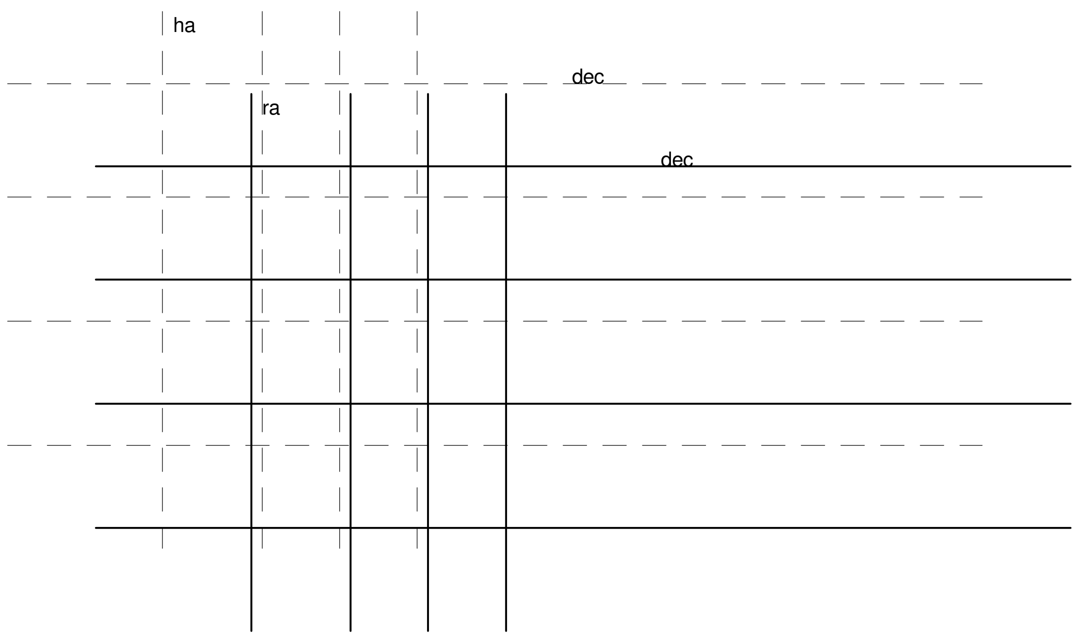
where ha and δ are the local coordinates of hour-angle and declination. R is again assumed to be separable into a part that depends on time and another on local-coordinates. Re-assigning a new time to each event performs the Monte Carlo integration. The "fake" event is transformed only in right-ascension with the declination remaining fixed, thus the "time-swapping" terminology. It is computationally much faster, thus more often employed.

A slight variation on the above is the so-called "direct integration" method described in e.g. Milagro memo #55, "Search for Point Sources and Emission from the Galactic Plane in Milagrito" by Andy Smith. Now instead of performing the integration via a Monte Carlo method, it is integrated via a numerical method. E(ha, δ) is obtained from the data over a long time period (~2 hrs.), and normalized to have unit total area. The numerical integration is performed by performing the "dt" integration each 24 seconds. The number of events observed is calculated for the 24 second period, and the background in celestial coordinates is incremented by the appropriate amount E(ha(ra), dec) times the number of events observed in the 24 seconds.

The Problem Stated

The problem that I address in this note comes from two causes. The first is that we are inherently using two different coordinate systems (local and celestial) in all of our background calculations. The second is that, because our calculations are all performed numerically, the data and/or background estimate are placed into bins in the relevant coordinate system. The problem is that the bins in one system do not perfectly align themselves with the bins in the other system.

This is (poorly) illustrated in the following figure, in which the local coordinate system (in this example, hour angle and dec) is given by the dashed lines, and the celestial one given by the solid lines. This is the situation for a particular instant in time. Obviously the centers of the bins do not coincide; I have drawn it this way purposely to illustrate the point. Given some area in the celestial sky, we want to evaluate the continuous integral of the background distribution over that patch of sky (the source bin). We may approximate this, for example in a method like the "direct integration" method, by breaking up the source bin into many small bins, and evaluating the background in each of these bins by using the value of the background in the corresponding local coordinate bin. As shown, these may or may not align themselves.

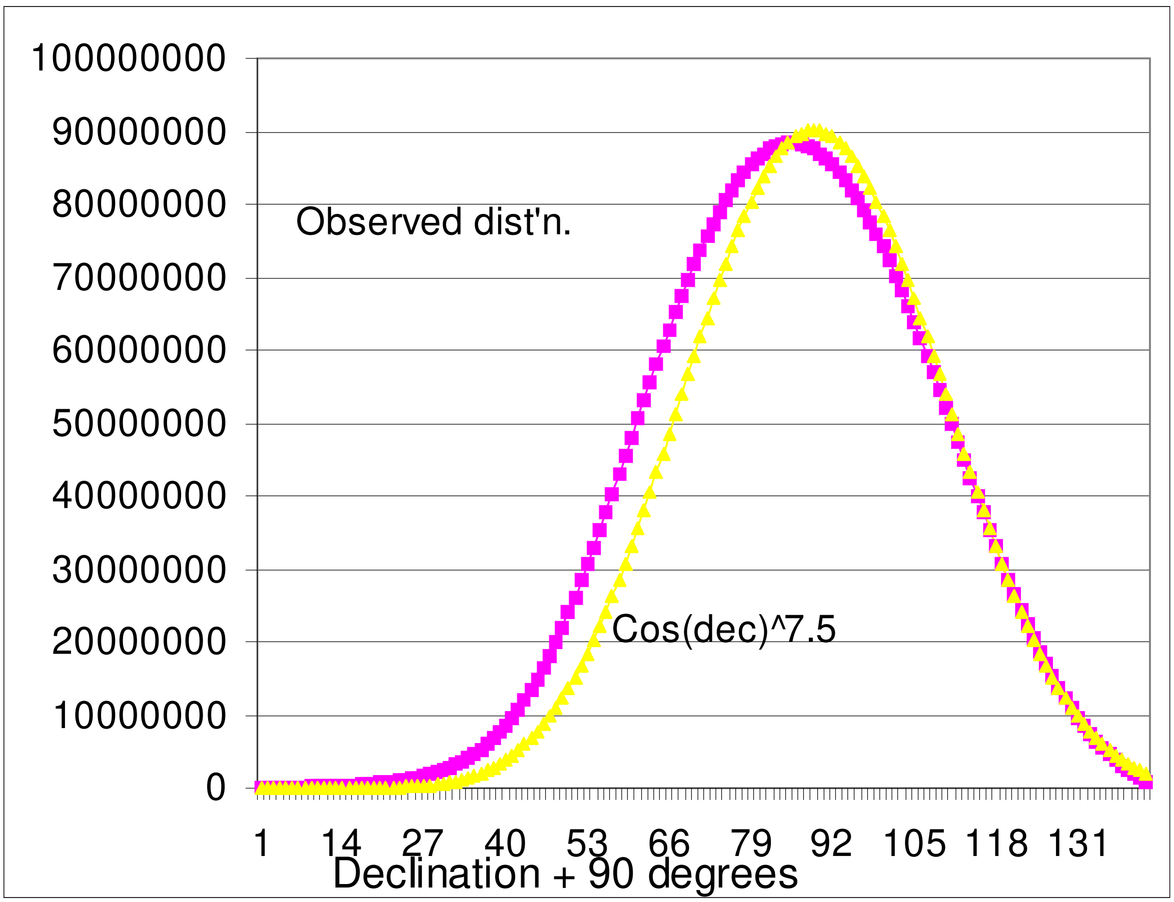


A slightly different situation presents itself when we are performing a background estimate for the galactic plane, or when we are using azimuth and zenith angles to describe the background distribution in local coordinates (as in the "coordinate swapping" method). In this case, the two coordinate systems are locally rotated with respect to each other. This causes an effective shift (similar to above) locally of one coordinate system with respect to another. Again, binning effects cause a systematic error in this situation.

Results from a Toy Model

My toy model of the situation is that the background distribution, in local coordinates, depends only on declination, and is a constant in hour angle; this is appropriate for long averages (>>1 day) of the background distribution. The main reason for this choice is that it gives the lowest value for the systematic error that I am exploring. Said differently, any short time determination of the background E(ha, δ) (or *any* determination of E(θ,φ)) will have a non-linear dependence on both coordinates. This non-linearity, after binning, will show itself as a systematic error after the integration is performed. Thus, I am being the opposite of conservative here, I am estimating a minimum systematic error.

The declination distribution as observed from Milagrito data (for Nfit >40) is shown in the following figure

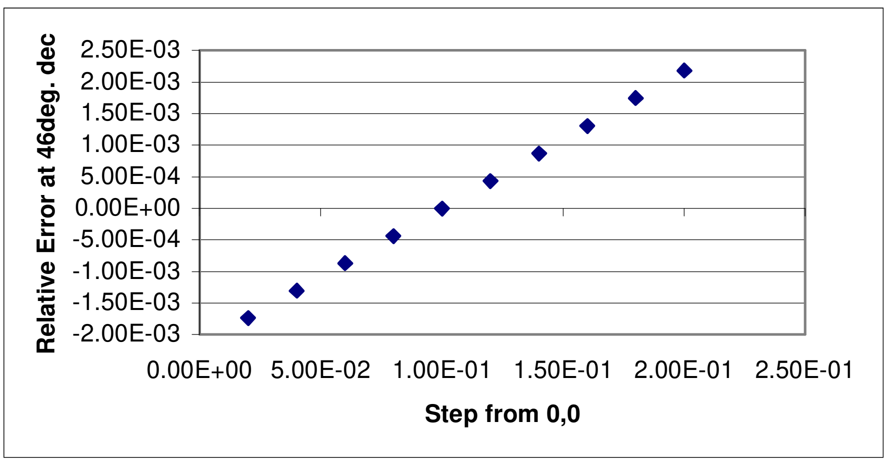


Where I have included a crude $\cos^{7.5}(\delta)$ model for comparison. You can see that this isn't such a bad model of the data observed to the north of our latitude, but does not represent the data particularly well to the south; much of this difference is due to solid-angle effects. I will use a $\cos^{7}(\delta)$ model in the calculations presented here.

For simplicity, I have chosen the background to be a constant in hour angle. This removes one variable (time) from the problem. (As stated above, this unrealistic simplification also results in an underestimation of the systematic error). So, to summarize, I assume the true background distribution of events to be E(ha, δ) = const * $\cos^7(\delta)$, and I integrate (numerically) this distribution over the corresponding background region of interest to determine the true background level.

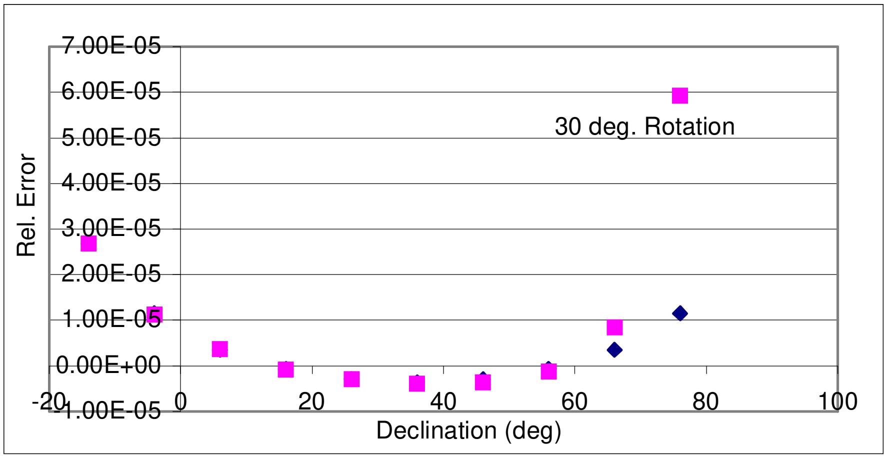
The "direct integration" method of background estimation is easiest for me to model in a semi-analytic way, and I do so by the following procedure. At some point in time, t, I am interested in determining the background for a region at (ra, δ) of width δ(ra) by δ(δ) wide; I assume that δ(ra) = δ(δ)/cos(δ), as usual. The background density at (ra, δ) is determined by the background density in the bin at the corrsponding local coordinates ((ha, δ)). The estimated background level in the bin is then the background density at the corresponding (ha, δ), times the size of the bin. I define the systematic error to be the fractional deviation of the estimated background from the "true" background (i.e. error = (true – estimate)/true).

As shown above, in general the bins in (ra, δ) are not perfectly aligned with those in (ha, δ). To illustrate the size of this effect, this figure shows the relative fractional error in the background estimate as a function of the misalignment in declination (in degrees) for an area of 0.2 by 0.2 degrees.



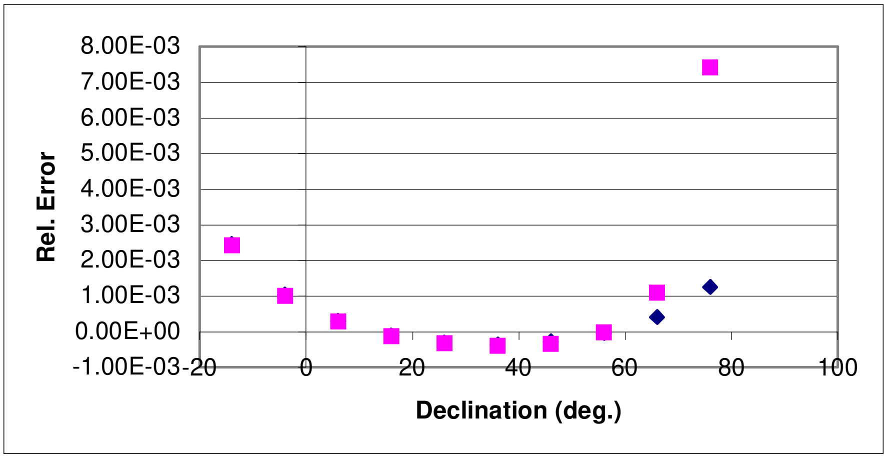
As one might expect from considerations of Taylor's theorem, the "best" estimate of the integral is the area of the bin times the density at the center of the bin, and this is borne out by the figure. However, in general this will not be the misalignment, a more typical value being ~1/4 of a bin corresponding to a relative error of $\sim 10^{-3}$.

Taking the center of the bin to represent the best one can hope to do with an approximation to the true integral, this figure shows the relative error as a function of declination:



where I have plotted the results for two situations, one (diamonds) where the two coordinate systems are aligned, and the other (squares) where they are both displaced and one is rotated by 30 degrees degrees with respect to the other.

One might also be interested in the situation when the local coordinate system is "binned" much more coarsely. The same plot as above, where this time the integration is made over an area of 2 deg. by 2 deg/cos(dec), is shown here:



where the diamonds and squares represent the same as above. You can see that the magnitude of the effects are ~100 times larger, where the bin is 100 times larger, just as one would expect.

(BTW—even if the coordinate systems are aligned, the effects discussed here are still relevant as a continuous integral over a variable x is only approximated by the sum over finite intervals of width dx, such as the 24 sec. integration steps made in the "direct" method.).

Conclusion

Several different situations are examined that can cause systematic errors in our background estimation methods. The level of the systematic examined here is from ~10^{-5} to 10^{-3}, depending on the situation. Even the smallest of these values is important enough to merit much more serious consideration for our present and future analyses.

I can make no statement about the precise value of the systematic errors made in each analysis so far performed on Milagro or Milagrito data. The precise value of the systematic mistake will be determined by details of the background estimation program that I do not know, and that have not been described in memos or meetings. However, the systematics discussed above may be important for some of the analyses (e.g. the galactic plane, and perhaps the point source searches for Mrk and the Crab). I have shown that the effects are large, possibly as large as the statistical precision of our data. In any event, it is time for us to seriously study our systematic background uncertainties.

There may be other sources of error that we should consider, including the statistical uncertainties in our background map ("$E(\theta,\varphi)$") because this map is determined by a statistical sampling of the true background distribution. This statistical uncertainty can show itself as a systematic background error if the data is not handled properly. Perhaps much more important is the assumption that the true background rate ("$R(\theta,\varphi,t)$") is separable into a time-dependent, and a time-independent part. I would wager that it is not separable at the required level of accuracy.

I hope that these toy-model results are useful at least to raise our collective consciousness in that we are exploring a new regime in sensitivity. In the past, many studies were performed that showed the methods were of sufficient accuracy, as merited by the data, via e.g. Monte Carlo simulations of fake data sets, and analysis of said fake data. As we push our sensitivity down, we must verify that our methods are unbiased and free of systematics to the required level, or develop and verify new methods.