PROCEEDINGS OF SCIENCE

Cosmic Ray Shower Profile Track Finding for Telescope Array Fluorescence Detectors

Jon Paul Lundquist[*†]
High Energy Astrophysics Institute and Department of Physics and Astronomy, University of Utah, Salt Lake City, Utah, USA
E-mail: jplundquist@cosmic.utah.edu

for the Telescope Array Project

A simple cosmic ray track finding pattern recognition analysis (PRA) method for fluorescence detectors (FD) has been developed which significantly improves X_{max} resolution and its dependence on energy. Events which have a clear rise and fall in the FD view contain information on X_{max} that can be reliably reconstructed. Shower maximum must be extrapolated for events with X_{max} outside the field of view of the detector, which creates a systematic dependence on the fitting function. The PRA method is a model and detector independent approach to removing these events, by fitting shower profiles to a set of triangles and applying limits on the allowable geometry.

The 34th International Cosmic Ray Conference,
30 July- 6 August, 2015
The Hague, The Netherlands

[*]Speaker.
[†]Full author list and Acknowledgements: http://www.telescopearray.org/images/papers/ICRC2015-authorlist.pdf

PoS(ICRC2015)442

1. Introduction

Cosmic ray extensive air-shower events with the best fluorescence detector (FD) X_{max} resolution will have a clear rise and fall in photon signal flux as a function of atmospheric depth (shower profile), as they contain enough information as to be reliably reconstructed (See Figure 4 for example). The reconstructed shower maximum, for all events with X_{max} out of the field of view (FOV) of the FD, has a systematic dependence on the fitting method and the assumed shower longitudinal distribution function form. The position of the shower maximum must be extrapolated as the profile will have a monotonically increasing or decreasing photon flux. Events on the lower end of the scale of energy for the detector will only be sufficiently bright enough for triggering near the shower maximum. This results in a relatively flat profile (See Figure 3b, for example) which will also have a systematic dependence on the fitting for the X_{max} reconstruction.

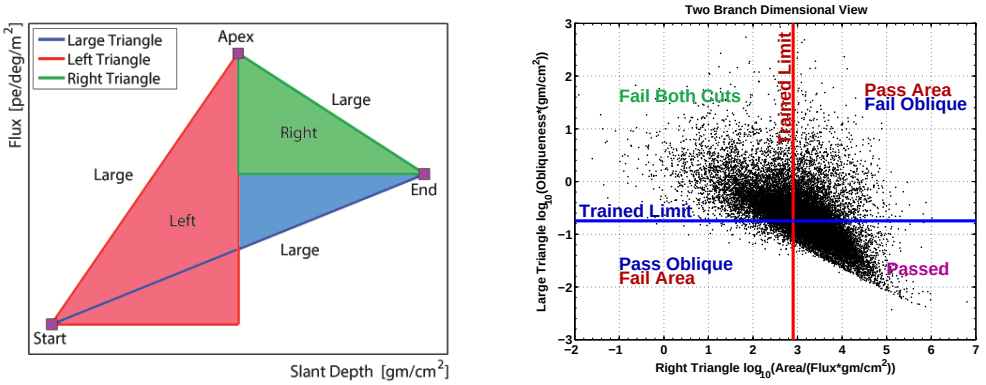
For monotonic profiles lacking a concave curvature fitting the Gaisser-Hillas (GH) function [1], requiring the resulting X_{max} to be within the FD field of view, and requiring a good fit to the GH function, can reduce the effect of these types of events on resolution. The problem of flat profiles cannot be resolved this way as lower energy events have relatively large statistical errors in the signal bins, and goodness of fit tests for the GH profile will often report very good results. Requiring only a goodness of fit limit results in a strong energy dependence of the resolution. A different approach to removing these events is required.

A simple pattern recognition analysis (PRA) method, independent of logitudinal distribution model, has been created which categorizes events as flat, monotonically rising/falling, or sufficiently concave in signal magnitude such that we can be confident X_{max} is in the field of view (Binary PRA). The method, along with Ultra-High Energy Cosmic Ray composition results, has been previously described in [2]. Binary PRA gives a yes/no answer on whether a particular event has sufficient profile curvature. This significantly improves the overall X_{max} resolution, and the energy dependence of the resolution.

The downside of the Binary PRA is significant loss of statistics. Events can be recovered by applying a technique called Logistic Regression ([3], [4]) which extends the description of events further from a yes/no answer to a scale of the quality of each event. We call this the Quality Factor analysis (QFA).

2. Description

The pattern recognition analysis is applied to reconstructed shower profiles where the general shape of an extensive air-shower track is independent of model and no model is needed to determine if there is an increase and decrease in signal within the FOV. The simplest possible abstraction of a "GH like" (or concave) distribution is a triangle. A set of triangles found from the events profile pattern (see Figure 1a) contain the parameters required to discriminate non-concave events. Based upon an eye scan of a sample (or training) set of events the method finds and sets limits on the allowed shapes of the extracted triangles, and rejects events outside those limits. Only events which contain useful information remain after cuts based on these limits are applied. The result is a non-adaptive track finder similar to those used in particle physics analysis [5].



(a) (b)

Figure 1: (a) Triangles created from the feature construction step. (b) A plot of two attribute cuts. Units given should not be interpreted as physical quantities.

3. Binary Pattern Recognition Analysis (PRA)

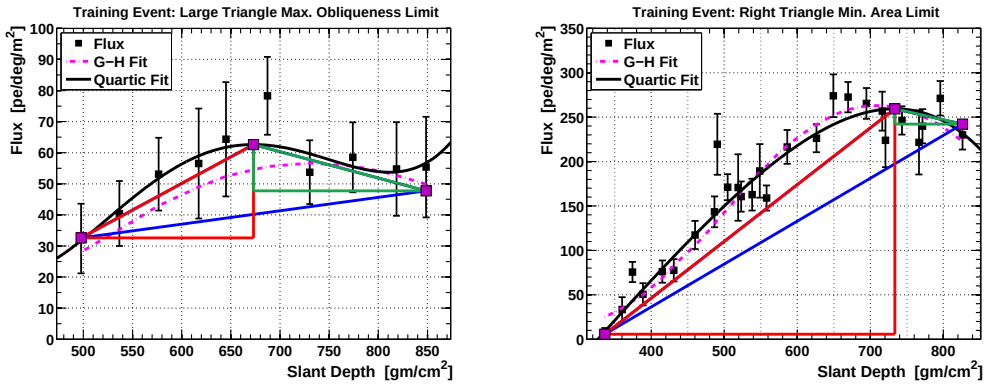
Pattern recognition training involves creating a training set, constructing features, a decision tree (or cut value) population, and feature selection [6].

The Training set is made by scanning by eye a selected subset of data and Monte Carlo (MC) simulated events and categorizing them based on whether sufficient profile concavity can be seen. Data and MC simulation can be treated equally as the training set is used only to find the allowable limits on the geometries of the triangles created from the feature construction step.

For feature construction a fit on the shower profile to a quartic polynomial is done using an iteratively reweighted least-squares minimization using Tukey's biweight function (or bisquare weights fit) which is robust against outliers and bins with large errors [7]. The local maximum of the fit within the track of the shower is the apex of the large triangle (Figure 1a). The other two large triangle vertex points are found using a linear fit using the same fitting method. A weighted average is used for the first or last three bins instead of the linear fit if there are three or less bins on a side of the apex. The quartic fit is not used for these two vertex points as it is unstable due to there being more data points around the apex. These three points are used to form five triangles. The three most useful of these are labeled on Figure 1a. (See also Figures 2 to 4.)

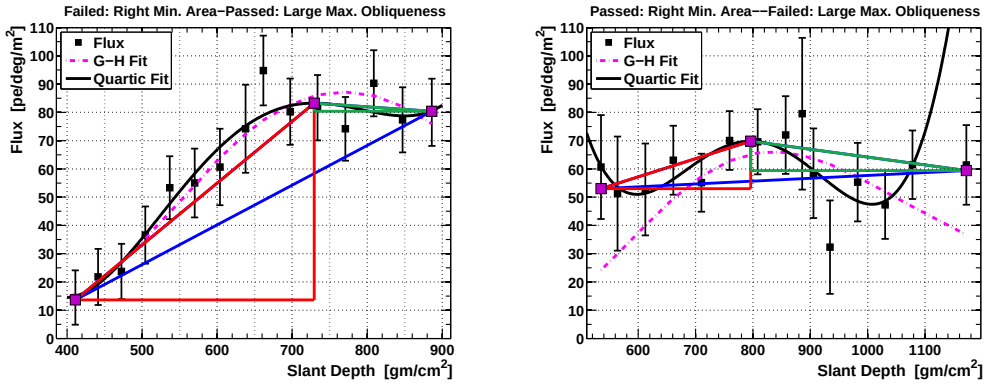
For each training set event, attributes of the shower profile and fitted triangles are calculated, such as the signal mean and standard deviation, size of the large triangle, and the attributes of the signal quartic polynomial fit which was used to find the apex. The largest and smallest values of each of these attributes, for the events which were found to have a clear X_{max} in view by eye scan, give the allowed limits of these attributes. The two cuts that remove the most bad events when applied individually are a maximum limit to the allowed obliqueness (perimeter/area) of the large triangle, and the minimum allowed area of the right triangle. Examples of events are shown in Figures 2-4. The effect of these two cuts is shown in Figure 1b.

For the training set of eye scanned good events, the event in Figure 2a has the maximum value of the obliqueness of the large triangle. This event sets the limit on the maximum allowed obliqueness of passed events. The training set good event in Figure 2b has the minimum value



(a) (b)

Figure 2: (a) The event which set the maximum limit on the obliqueness of the large triangle. (b) The event which set the minimum limit on the area of the right triangle. Bins with large errors have been removed for display purposes.



(a) (b)

Figure 3: (a) An event which passed the large triangle obliqueness test but failed the right triangle area test. (b) An event which passed the right triangle area test but failed the large triangle obliqueness test.

of the right triangle area. This event sets the limit on the minimum allowed right triangle area of passed events. An event which was cut due to the right triangle area being smaller than the allowed limit is shown in Figure 3a. Figure 3b shows a failed event for which the large triangle obliqueness is larger than the allowed limit. Figure 4 shows the best event with the smallest obliqueness and largest right triangle area.

Feature selection is done to decide which of the calculated attributes are necessary to decide whether an event is good or bad. Cuts on attributes for which the extreme values of the good events do not remove any bad events (as determined by eye scan) are removed along with those that remove less than 0.5% of the training set. This minimizes the number of false negatives due to overfitting when the cuts are applied to the whole data set. Cuts (or groups of cuts based on categories) which remove the same events as another cut (or group of cuts) are also not used. The result is the minimum number of parameters needed to categorize events as good or bad. The full

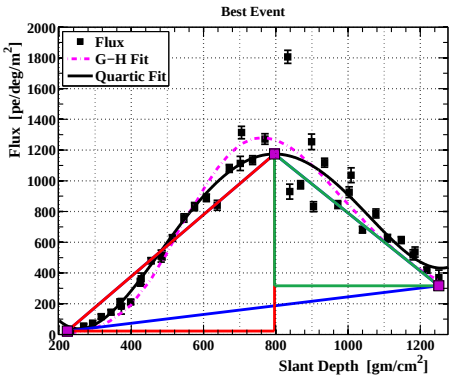


Figure 4: The passed event with the minimum value of large triangle obliqueness, and maximum value of right triangle area of the whole set. This is also the highest energy event at $\log_{10}(\text{Energy}/\text{eV}) = 20.12$. Bins with large errors have been removed for display purposes.

method will be explained in further detail in [8].

The result is an overall accuracy of 97.6% when the cuts are applied to the training set. The total percentage of false positives being 2.4%.

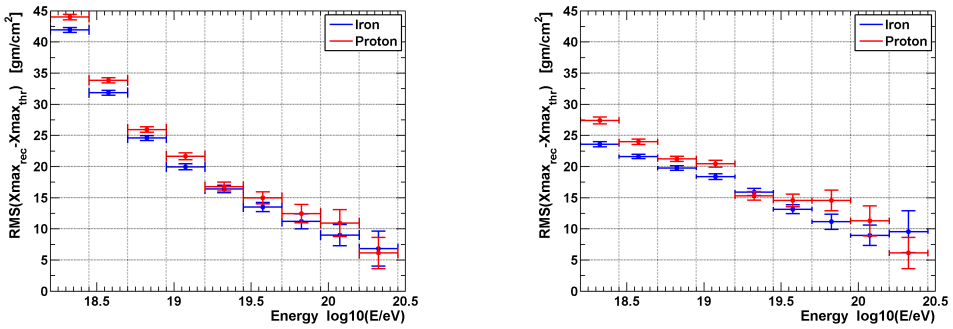
Application involves extracting the features for the set of all data and proton/iron simulated events, calculating the parameters which survived the feature selection process, and applying the cuts. If an event passes all cuts it is considered a good event. The result is a set of events for which the shower profiles have sufficient convex curvature that we can be confident that X_{max} is within the field of view (FOV) of the fluorescence detector.

Random test samples of events from the data and proton Monte Carlo (MC) simulation sets were scanned by eye, and the PRA applied. The result was that the pattern recognition is 96.5% accurate on the data and proton MC test sets. Twice as many random events were chosen for an iron MC test set as the pattern recognition was not trained on iron MC events. The iron MC accuracy is also found to be 96.5%. The overall accuracy including false positives and negatives, when comparing the eye scan and pattern recognition for both training and test sets was 97.2%. On average only events in which X_{max} is not within the FOV (false positives) are detrimental to the resolution. Counting only false positives the accuracy percentage is 99.6%.

One of the benefits of this method can be seen in Figure 5. This figure shows that the energy dependence on resolution, largely caused by events with X_{max} outside the fluorescence detector FOV, is significantly improved by the pattern recognition analysis when compared to the usual shower geometry cuts. This result was published in [2].

4. Quality Factor Analysis (QFA)

The Binary PRA method, while effectively modeling the eye scan determination of events with X_{max} within the field of view of the FD, creates an approximately 50% reduction in the number of passed events compared to the usual type of shower geometry cuts. Instead of cutting events by a binary yes/no answer some of these events can be recovered by determining a scale of quality of events and selecting a cutoff on this value. This can be done with Logistic Regression ([3], [4])



(a) (b)

Figure 5: X_{max} resolution plots showing the energy dependence of the RMS of the difference between QGSJETII-03 MC reconstructed and thrown X_{max}. (a) is geometry cuts only. (b) is the pattern recognition with additional geometry cuts applied.

which is a type of binary classification. Logistic Regression (LR) is used extensively in marketing, finance, the social sciences, biology, and medicine.

The inputs to LR are a binary yes/no, pass/fail response vector and a set of attributes or parameters (predictors) which are intended to predict the response. The coefficients of the LR model that are returned can be used to calculate the probability of a future sample being a "pass" for the yes/no question, given that events values for the same set of predictors.

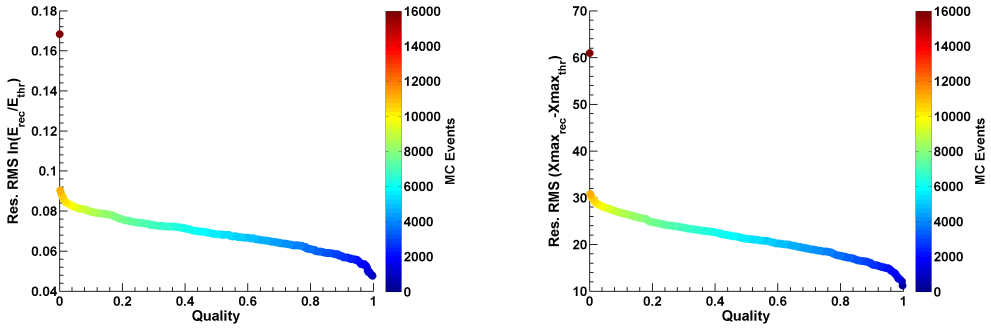
The logistic regression cost function which is minimized to find the coefficents $\vec{\beta}$ is shown in Equation 4.1. This function is derived using maximum likelihood as shown in [3] and [4]. N is the number of events, j is the event index, y_j is the response from PRA (a 0 or 1), $\vec{x_j}$ is the vector of predictor values (a subset of the parameters used in the Binary PRA). Equation 4.2 is the logistic function and takes as input the dot product of the found coefficient vector and the predictor vector and returns the probability that the event is good. The logistic function maps the range $(-\infty, \infty)$ to $[0,1]$.

$$min_{\beta} J(\beta) = -\sum_{j=1}^{N} [y_j log F(\vec{x_j}) + (1-y_j) log(1-F(\vec{x_j})] \tag{4.1}$$

$$F_j(\vec{x_j}) = \frac{1}{1+e^{-(\beta_0+\vec{\beta}\cdot\vec{x_j})}} \qquad \vec{\beta}\cdot\vec{x_j} = \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots \tag{4.2}$$

In this case the response vector is the output from the PRA and the predictors are the same shower profile attributes used in the PRA yes/no determination. Only a subset of PRA attribute cuts were used to create the response and predictor vectors. The subsets were chosen to maximize the increase in statistics for any particular resolution and to optimize the correlation between Quality Factor(QF) and resolution.

The resulting QF scale shows a strong correlation with resolutions of a number of cosmic ray shower geometry variables. The energy and X_{max} resolutions with respect to QF cutoff (or integral plot) are shown in Figure 6. There is also an approximately 40% increase in the number of passed events over the PRA for the same resolution.



(a) (b)

Figure 6: QGSJETII-03 Proton Monte Carlo Energy and X_{max} integral RMS resolution plots showing the Quality Factor (QF) cutoff correlation with resolution. (a) is the energy resolution with respect to QF cut. The first data point is with no QF cut and shows an energy resolution of 0.17. (b) is the X_{max} resolution with respect to QF cut. The first data point is with no QF cut and shows an X_{max} resolution of 60.9 g/cm^2. Both figures are for MC events with $\log_{10}(\mathrm{E/eV}) > 18.2$, shower core distance from SD boundary > -1500 meters (negative is outside the array), Zenith angle < 60, and SD/FD shower core position difference < 2500 meters.

5. Conclusion

Pattern recognition analysis is a model independent method that removes events for which X_{max} is not clearly in view, and are therefore generally poorly reconstructed. This results in an improved X_{max} resolution and a decreased energy dependence of X_{max} resolution. While this method causes a large decrease in statistics events can be recovered by extending the method to the Quality Factor (QF) analysis which provides a scale to describe the quality of events. There is a strong correlation between shower parameter resolutions and the QF. A cutoff of allowable quality can be chosen to increase statistics or improve resolution.

References

[1] T.K. Gaisser, A. M. Hillas, *Reliability of the Method of Constant Intensity Cuts for Reconstructing the Average Development of Vertical Showers*, in proceedings of *ICRC*, ICRC Vol. **8** (1977).

[2] R. Abbasi, *et al*, *Study of Ultra-High Energy Cosmic Ray Composition Using Telescope Array's Middle Drum Detector and Surface Array in Hybrid Mode*, *Astroparticle Phys.* **64**, 49 (2014).

[3] D. R. Cox, *The Regression Analysis of Binary Sequences*, *J. Roy. Statist. Soc., B*, **20**, 215 (1958).

[4] D. Collett, *Modelling Binary Data*, 1st Edition, Springer 1991.

[5] A. Strandlie, R. Frühwirth, *Track and vertex reconstruction: From classical to adaptive methods*, *Rev. Mod. Phys.* **82**, 1419 (2010).

[6] I. Guyon, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, 1st Edition, Springer 2006.

[7] P. W. Holland, R. E. Welsch, *Robust regression using iteratively reweighted least-squares*, *Comms. in Stats.-Theory and Methods Vol. 6*, **9**, 813 (1977)

[8] J. P. Lundquist, P. Sokolsky, *Cosmic Ray Shower Track Finding Using Telescope Array Fluorescence Detectors*, *Unpublished manuscript*