A Logistic Regression Analysis of 1FGL Unassociated Sources

D. Salvetti[1,2,*], P. A. Caraveo[2], P. M. Saz Parkinson[3], A. de Luca[2], A. Belfiore[1,2]
on behalf of the Fermi Large Area Telescope Collaboration
[1]Dipartimento di Fisica Nucleare e Teorica, Università di Pavia, 27100 Pavia, Italy [2]INAF-IASF, I-20133 Milano, Italy
[3]Santa Cruz Institute for Particle Physics, University of California at Santa Cruz, Santa Cruz, CA 95064
*E-mail address: salvetti@iasf-milano.inaf.it

630 sources in the first catalog of *Fermi*-LAT γ-ray sources have no plausible counterpart. A logistic regression analysis allows us to quantify the probability for an unassociated source to be either an AGN or a pulsar.

The Fermi Large Area Telescope First Source Catalog (1FGL) lists positional, spectral, and temporal properties for 1451 γ-ray sources. 630 of these sources (~40%) remain "unassociated", i.e. do not have any plausible known counterpart. We report on the use of Logistic Regression as a very promising method to understand the nature of Fermi-LAT unassociated sources. Logistic Regression allows us to quantify the probability for an unidentified source to be either an AGN or a pulsar (the two most abundant classes of objects in the Fermi catalog), using parameters describing the γ-ray properties of the source. We will discuss applications of our results for obtaining a preliminary estimate of the pulsar and AGN fractions among unassociated sources, as well as for planning multiwavelength follow-up observations.

Fermi-LAT First Source Cataog

The *Fermi*-LAT [1] First Source Catalog (1FGL) lists 1451 sources detected during the first 11 months of operation by the LAT in the 100 MeV to 100 GeV energy range [2].

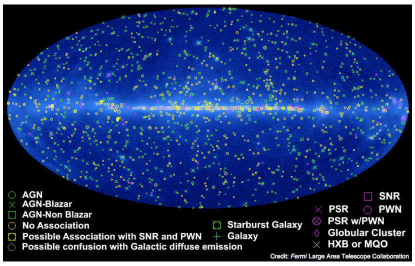

Figure 1: *The 1451 1FGL catalog sources.*

For each LAT source, the proposed associations with sources in other astronomical catalogues is based primarily on positional coincidence. The 95% uncertainty radii for 1FGL sources are typically 10 arcmin. Thus, these position measurements are often inadequate to make identifications based solely on location. As a result,
693 sources were associated with AGNs, 8 with galaxies,
63 with pulsars, 57 with SNRs, PWNe, GC and binaries,
while **630 1FGL sources remain unassociated**.

Distribution of 1FGL sources

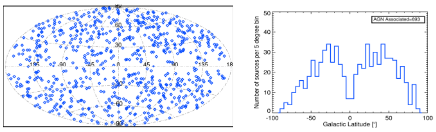

Figure 2: *Spatial distribution of 1FGL AGNs.*

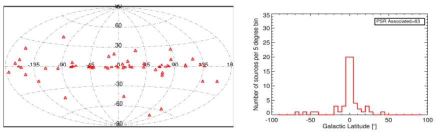

Figure 3: *Spatial distribution of 1FGL pulsars.*

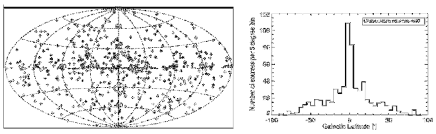

Figure 4: *Spatial distribution of 1FGL unassociated sources.*

Logistic Regression analysis

We decided to use a Logistic Regression (LR) analysis to quantify the probability of an unassociated source to be a specific gamma-ray source based on their gamma-ray properties, in order to plan multiwavelength follow-up observations. LR is part of a class of generalized linear models and it allows us to form a multivariate relation between a dependent variable and several independent variables (e.g. spectral shape or variability). The logistic function is defined as:

$f(z) = (1+e^{-z})^{-1}$

where f(z) is the probability of an event to occur and z can be defined as

$Z = b_0+b_1X_1+...+b_nX_n$

where b_0 is the "intercept" of the model, b_i (i = 1, 2, ..., n) are the "regression coefficients" (they were found using a Maximum-Likelihood method) and the X_i (i = 1, 2, ..., n) are the independent variables.

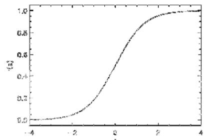

Figure 5: *Logistic Regression function*

Training sample and predictor variables

LR must be trained on known objects in order to predict the membership of a new object to a given class on the basis of its observables. We decided to train the predictor using the pulsars and AGNs identified in the 1FGL catalog [2] because they are abundant and have differing phenomenology. We then set f(z)=1 to the 693 associated AGNs and f(z)=0 to the 63 pulsars. In such a way the output of our training process is a parameter that determines the probability that an unassociated source to belong to the AGN or pulsar source class.
To evaluate the best predictor variables, we used the likelihood ratio test. We started by using the Fractional Variability, the Hardness Ratios for the 5 energy bands in the catalog [3] and the position on the sky.

Variable	Coefficient	Standard Error	p-value
Fractional Variability	12.14	1.47	<0.001
$Hardness_{23}$	-7.63	1.15	<0.001
$Hardness_{34}$	2.91	0.86	<0.001
$Hardness_{45}$	2.04	0.56	<0.001
Intercept (b_0)	0.34	0.54	-

Table 1: *List of the predictor variables included in the LR model with the relative statistical significance. All the other variables have a p-value > 5.*

Defining classification thresholds

We set two classification thresholds, one to single out AGN candidates (C_A) and one for pulsar candidates (C_P). We chose these two thresholds so that 80% of the AGNs in 1FGL have a predictor value greater than C_A and 80% of the pulsars have a predictor value smaller than C_P. Thus, we set C_P to 0.63 and C_A to 0.97. By using these thresholds, only 1% of the AGNs are misclassied as a pulsar and 5% of pulsars as an AGN.

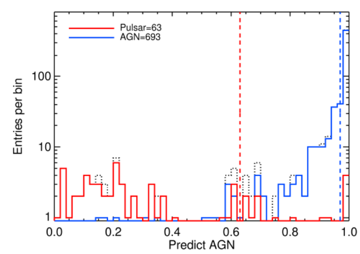

Figure 6: *Distribution of the LR predictor for the training sample. Vertical lines indicate the value of the thresholds we set to identify pulsar and AGN candidates.*

LR classification of 1FGL unassociated sources

Applying the LR model to the 1FGL unassociated sources we found that 122 sources are classified as pulsar candidates (P < 0.63), 368 as AGN candidates (P > 0.97) while 140 remain unclassified.

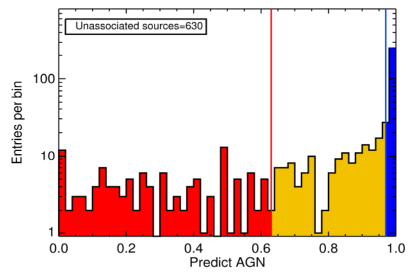

Figure 7: *Distribution of the LR predictor for 1FGL unassociated sources. Vertical lines indicate the two classification thresholds.*

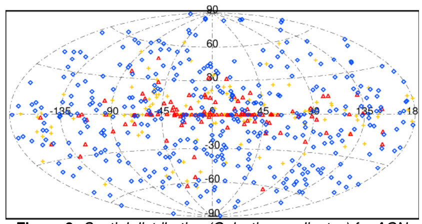

Figure 8: *Spatial distribution (Galactic coordinates) for AGN candidates (blue diamond) and pulsar candidates (red triangles). Yellow crosses represent unclassified sources*

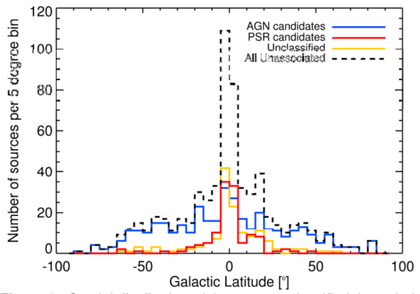

Figure 9: *Spatial distribution of the sources classified through the LR method, as a function of their Galactic latitude.*

Testing our LR classification with fresh source associations

Classification	New AGN associations	New *gamma-ray* PSR associations	New *radio* PSR associations
Correct	85%	63%	28%
Negative	2%	0%	22%

Table 2: *Success rate of the LR algorithm for identifying the newly associated sources.*

After the publication of the 1FGL catalog, multiwavelengths studies yielded 128 new AGN associations and 34 (16 gamma-ray pulsars and 18 radio ones discovered inside 1FGL error boxes) new pulsar ones. Using these fresh findings, we tested the efficiency of the LR algorithm for classifying both new AGNs and pulsars. Table 2 summarizes our results showing high success rate coupled with few false negatives. Only the success rate computed for the new radio pulsars is unsatisfactory, probably owing to the faintness of their gamma-ray yield.
LR classification results can, indeed, serve as a guide for finding associations for unassociated gamma-ray sources.

References

[1] *Fermi*-LAT Collaboration, *ApJ*, 697, 1071 (2009)
[2] *Fermi*-LAT Collaboration, *ApJS*, 188, 405 (2010)
[3] *Fermi*-LAT Collaboration, in preparation