PROCEEDINGS OF SCIENCE

# A data mining approach to recognizing source classes for unassociated gamma-ray sources

**Kenji Yoshida***
*Shibaura Institute of Technology, 307 Fukasaku, Minuma-ku, Saitama 337-8570, Japan*
*E-mail:* yoshida@shibaura-it.ac.jp

The Fermi-LAT 3rd source catalog (3FGL) provides the gamma-ray properties for 3034 gamma-ray sources. While 2024 sources in the 3FGL are associated with AGNs (58 % of the total), pulsars (5 %) and the other classes (4 %), 1010 sources (33 %) remain as unassociated sources. In recognizing source classes for unassociated gamma-ray sources of the Fermi-LAT source catalogs, various data mining techniques have been applied, e.g. classification tree and artificial neural network. As a robust alternative to these data mining techniques, we present the Mahalanobis Taguchi (MT) method to recognize source classes. The MT method creates a multidimensional unit space from characteristic variables of a normal class (e.g. AGN) to identify sources of the normal class from those of the other classes using Mahalanobis distances. In this paper, we present the results of the source classification for the unassociated gamma-ray sources in 3FGL by applying the MT method. We also discuss a possibility of dark matter Galactic subhalos for the unclassified sources at $|b| > 20^\circ$.

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

*Speaker.

 

PoS(ICRC2015)857

## 1. Introduction

The Fermi-LAT 3rd source catalog (3FGL) provides spatial, spectral, and temporal properties for 3034 gamma-ray sources. While 2024 sources in the 3FGL are associated with AGNs (1745 sources), pulsars (167 sources) and the other classes (112 sources), 1010 sources remain as unassociated sources [1].

In recognizing source classes for unassociated gamma-ray sources of the Fermi-LAT source catalogs, M.Ackermann *et al.* (2012) [2] employed two data mining techniques to determine likely source classification for the 1FGL unassociated sources: Classification Trees and Logistic Regression. They applied these techniques using the gamma-ray properties that are not related to source significance, as this will bias the results.

N.Mirabal *et al.* [3] investigated the possibility that dark matter annihilation signals coming from Galactic subhalos may account for a small fraction of unassociated point sources in the 2FGL. They applied a Random Forest classifier *Sibyl* that offers predictions on class memberships for unassociated 2FGL sources. In order to construct and train the *Sibyl*, they used the gamma-ray properties dependent on source significance of AGNs and pulsars.

M.Doert and M.Errando (2013) [4] applied a random forest and a neural network method to identify AGN candidates for unassociated sources in 2FGL. Combining the two learning algorithms, they evaluated the false-association rate of 11 % to recognize 80 % of AGNs.

While the classification for unassociated gamma-ray sources are useful for planning multi-wavelength follow-up observations, some sources might be unclassified as known objects such as AGNs and pulsars. Among the unclassified sources, the interesting sources of gamma ray emission are dark matter Galactic subhalos. Numerical simulations of cold dark matter particles indicate that the Galactic halo contains a very large number of dark matter subhalos. As the dark matter annihilations taking place within such dark matter subhalos emit gamma rays, the most nearby and massive subhalos could appear as point-like gamma-ray sources without observable counterparts at other wavelengths. The search for dark matter subhalos in the Fermi catalogs is currently ongoing (e.g. [5]).

In this paper, we investigate the classification of unassociated gamma-ray sources in 3FGL applying a robust alternative data mining technique: the Mahalanobis Taguchi method. We also discuss the possibility of identifying dark matter subhalo candidates.

## 2. Mahalanobis-Taguchi method

The Mahalanobis Taguchi (MT) method is a robust data mining technique developed in quality engineering [6]. The MT method is proposed as a diagnosis and forecasting method using multivariate data. While the MT method has been used in different diagnostic applications to make quantitative decisions by constructing a multivariate measurement scale using data analytic methods, the MT method is applied for particle identification of cosmic ray observations [7]. One of the main objective of the MT method is to construct a Mahalanobis space based on input characteristic variables. The Mahalanobis space, which is also called the unit space, is obtained using the standardized variables of normal data. The Mahalanobis space can be used to discriminate between normal and abnormal data to measure the degree of abnormality, so-called the Mahalanobis

distance. This approach requires a uniformity of the normal data to construct a unit space (Mahalanobis space) from characteristics of samples. By applying sample data to the unit space, we can calculate the Mahalanobis distances from the reference point.

The Mahalanobis distance is a squared distance (also denoted as $D^2$) given by the following formula:

$$D^2 = \frac{1}{k} Z_i^T C^{-1} Z_i, \tag{2.1}$$

where $k$ is the number of characteristics, $T$ is transpose of a vector, $C^{-1}$ is inverse of the correlation matrix, and $Z_i$ is a standardized vector obtained by $i$-th characteristic $X_i$ as follows:

$$Z_i = (X_i - m_i)/s_i (i = 1, 2, \cdots, k), \tag{2.2}$$

where $m_i$ is a mean of $i$-th characteristic and $s_i$ is a standard deviation of $i$-th characteristic. We used $D$ as the Mahalanobis distance in this paper. As the Mahalanobis distance of a source is closer to 1, the source is more similar to the reference sources. By using a fiducial threshold of the Mahalanobis distance, we can discriminate signal sources from the background sources.

## 3. Classification of unassociated sources

In the 3FGL catalog, the gamma-ray spatial, spectral, and temporal properties measured by the Fermi-LAT are summarized for individual sources. In this study, we selected the following properties given by Ackermann *et al.* (2012) [2]. As source significance will bias results, these properties are not related to the source significance. The hardness ratios $HR_{ij}$ are constructed as

$$HR_{ij} = \frac{\nu F_{\nu_i} - \nu F_{\nu_j}}{\nu F_{\nu_i} + \nu F_{\nu_j}}. \tag{3.1}$$

where $\nu F_{\nu_i}$ is the spectral energy distribution for energy band $i$. The energy bands $i$ of 1, 2, 3, 4, and 5 correspond to $0.1-0.3$ GeV, $0.3-1.0$ GeV, $1.0-3.0$ GeV, $3.0-10.0$ GeV, and $10.0-300$ GeV, respectively. It is also possible to define a quantity that discriminates curvature by the difference between two hardness ratios such as $HR_{23} - HR_{34}$. To remove the source significance dependency for variability, we use the fractional variability described in Ackermann *et al.* (2010) [8]. The fractional variability is given by

$$\frac{\delta F}{F} = \sqrt{\frac{\Sigma_i (F_i - F_{av})^2}{(N_{int} - 1) F_{av}^2} - \frac{\Sigma_i \sigma_i^2}{N_{int} F_{av}^2} - f_{rel}^2}, \tag{3.2}$$

where $N_{int}$ is the number of time intervals (48 in 3FGL), $F_{av}$ is the average flux, $\sigma_i$ is the statistical uncertainty in the flux $F_i$, and the $f_{rel}$ is an estimate of the systematic uncertainty on the flux for each interval (2% in 3FGL). In addition, we use the spectral index $\Gamma$ of the best fitted power-law spectrum, the Galactic Longitude ($\ell$), and the Galactic Latitude ($b$).

Among 2024 associated sources in the 3FGL, the most abundant class of sources is AGN (1745 sources), and the second most abundant class is pulsar (167 sources). The other classes contain 112 sources. In this study, we focus on the two most abundant classes of sources in the 3FGL, AGN and pulsar.

### 3.1 AGN classification

For the AGN classification, we constructed a unit space of the attributes of AGNs as a normal data set, and derived the Mahalanobis distances of AGNs and non-AGNs. To construct the AGN unit space, we used the properties of the fractional variability $\delta F/F$, the hardness ratio $HR_{12}$, $HR_{34}$, $HR_{45}$, the spectral index $\Gamma$, the Galactic longitude $\ell$, and the Galactic latitude $b$ to transform these properties into the following characteristics:

$$HR_{12}, \quad HR_{34}, \quad HR_{45}, \quad \log(\frac{\delta F}{F}), \quad \Gamma, \quad \frac{\ell - 180.0}{b}, \quad \frac{1.0}{|b|}.$$

To evaluate the classification performance of our method, we cross-validated using the 1745 AGNs and the other 279 sources in the 3FGL. We held out 1/5 of the sample at random to be the testing data set, and we used the remaining 4/5 of the sample for constructing the AGN unit space. In this method, the construction of the unit space corresponds to training in machine learning algorithms. We repeated this procedure 5 times, using a different set of 1/5 of the sample in each data set. At the end, by this 5-fold cross-validation we can evaluate the testing efficiency rates for 1745 AGNs and 279 non-AGNs.

### 3.2 Pulsar classification

For the pulsar classification, we constructed a unit space of the attributes of pulsars as a normal data set, and derived the Mahalanobis distances of pulsars and non-pulsars. For the pulsar unit space, we used the properties of the fractional variability $\delta F/F$, the hardness ratio $HR_{23}$, $HR_{34}$, and $HR_{45}$ to transform these properties into the following characteristics,

$$HR_{23} - HR_{34}, \quad HR_{45}, \quad \frac{\delta F}{F}.$$

In the similar way with AGN classification, we also cross-validated using the 167 pulsars and the 1857 non-pulsars of AGNs, supernova remnants, pulsar wind nebulae, and so on. For the pulsar classification, we held out nearly $1/5$ (33 sources) of the pulsar sample at random for the testing data set, and we used remaining 134 sources to construct the pulsar unit space. By the 5-fold cross-validation, we can evaluate the testing efficiency rates for 165 pulsars and 1857 non-pulsars.

## 4. Results

Figure1 shows the distributions of the Mahalanobis distance $D$ in the AGN unit space for the 3FGL associated sources (left panel) and for the unassociated sources (right panel). The distribution of associated sources clearly shows that we can select a set of AGN and non-AGN candidates, when setting the appropriate fiducial threshold. For 80 % efficiency rate of AGN sources with the fiducial threshold of $D = 1.08$ in the AGN unit space, 9.3 % non-AGN sources (26 of 279 sources) remains as AGN classification. For 95 % efficiency rate of AGN sources with the fiducial threshold of $D = 1.44$ in the AGN unit space, 22.6 % non-AGN sources (63 of 279 sources) remains as AGN classification.

Figure2 shows the distributions of the Mahalanobis distance $D$ in the pulsar unit space for the 3FGL associated sources (left panel) and for the unassociated sources (right panel). For 80 %

efficiency rate of pulsars with the fiducial threshold of $D = 1.10$ in the pulsar unit space, 5.2 % non-pulsar sources (97 of 1857 sources) remains as pulsar classification. For 95 % efficiency rate of pulsar with the fiducial threshold of $D = 2.43$ in the pulsar unit space, 33.0 % non-pulsar sources (612 of 1857 sources) remains as pulsar classification.

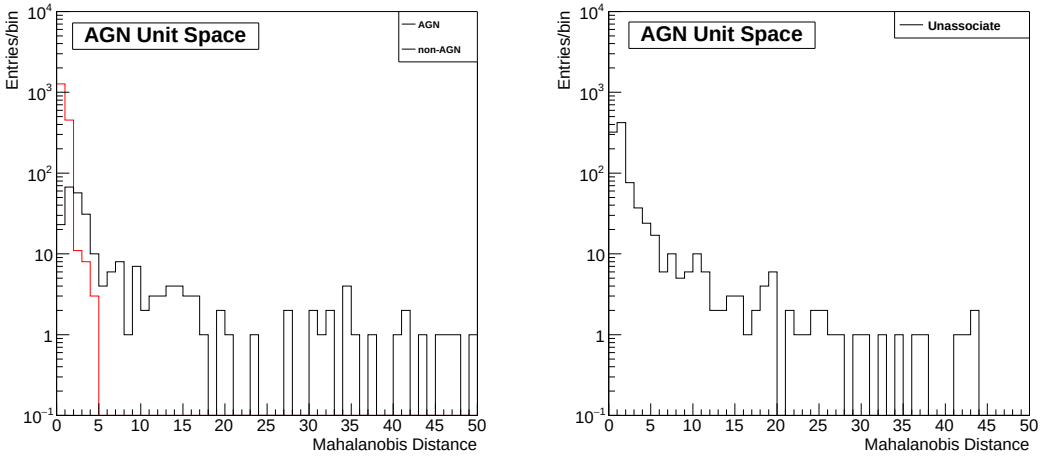


**Figure 1:** Distributions of the Mahalanobis distances for AGN classification. Left: For sources of the 3FGL catalog associated as AGNs (red histogram) and non-AGNs (black histogram). Right: For 3FGL unassociated sources.

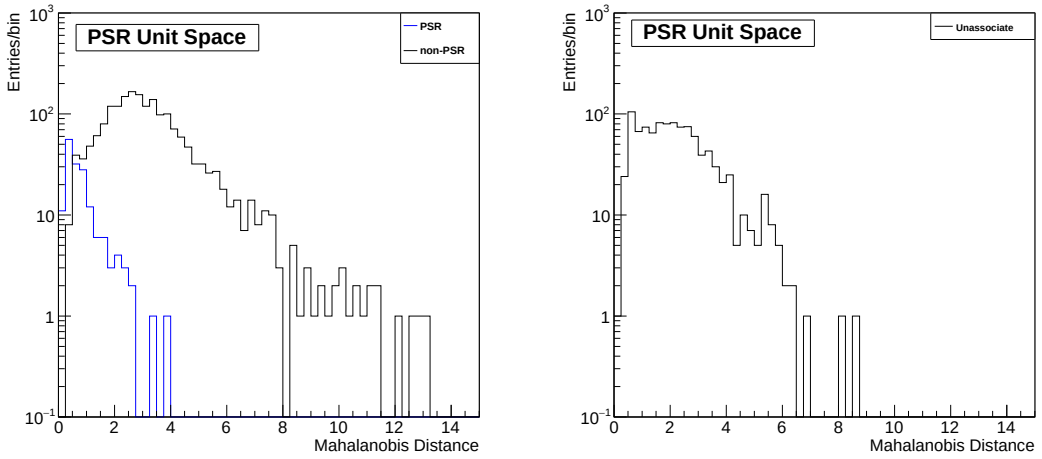


**Figure 2:** Distributions of the Mahalanobis distances for pulsar classification. Left: For sources of the 3FGL catalog associated as pulsars (blue histogram) and non-pulsars (black histogram). Right: For 3FGL unassociated sources.

Combining the two classifications, we classified 1010 unassociated sources into AGN candidates, pulsar candidates, and unclassified candidates. By using a 95 % fiducial threshold, the

sources are classified as 511 AGN candidates, 360 pulsar candidates, and 139 unclassified candidates. In conflict case between AGN and pulsar classification, we classified the sources to be a class with the smaller Mahalanobis distance. Figure 3 presents a spatial distribution of the combined classification for the unassociated 3FGL sources in Galactic coordinates.

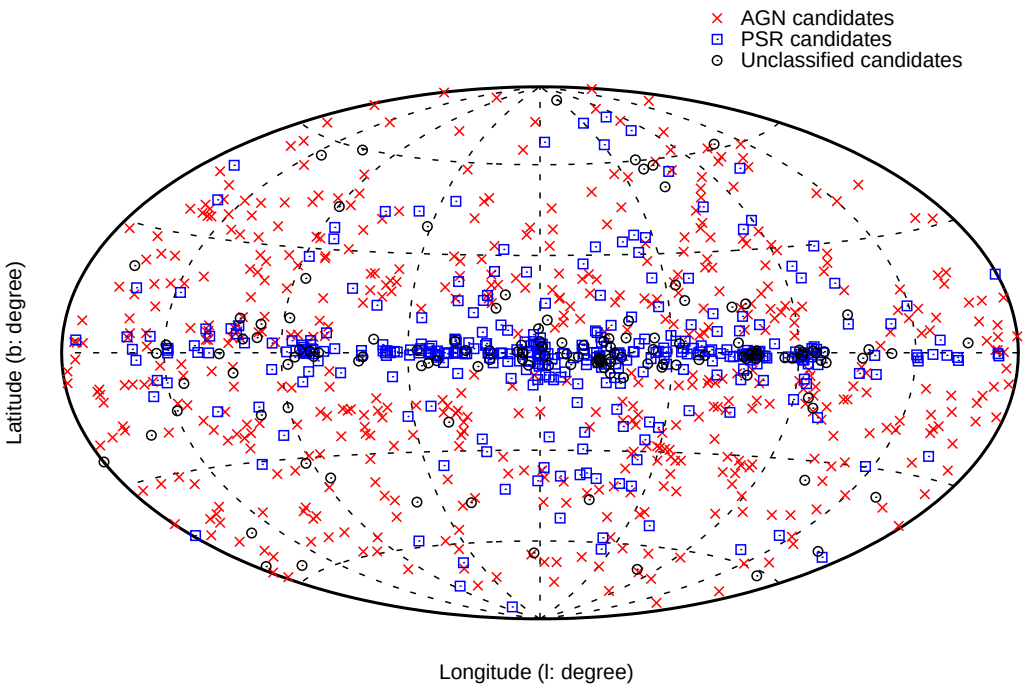


**Figure 3:** Spatial distribution of the combined classification for unassociated 3FGL sources in Galactic coordinates. Sources are classified as AGN candidates (red cross), pulsar candidates (blue square), or unclassified (black circle).

## 5. Discussion

As shown in Fig. 1 and Fig. 2, the shapes of the unassociated source distributions are different from the associated source distributions. For the discrimination between AGNs and non-AGNs in Fig. 1, there is an apparent absence of AGN-like sources in the unassociated source distribution, compared with the associated source distribution. For the discrimination between pulsars and non-pulsars in Fig. 2, there is also an apparent absence of sources larger than $D = 6$ in the unassociated source distribution. This may be due to the presumably different fractions of AGNs and pulsars in the associated and unassociated samples, or there may be an additional contributing component.

While this work might be useful for planning future multi-wavelength follow-up observations, there may be gamma-ray sources without detectable counterparts at other wavelengths. Dark matter

annihilations taking place in nearby dark matter Galactic subhalos could appear as such gamma-ray sources. Bertoni *et al.* (2015) [9] indicates that the 3FGL might contain on the order of $\sim$10 dark matter subhalos. In discussion, we consider the collection of unclassified gamma-ray candidates in 3FGL.

In order to isolate outliers that might constitute dark matter subhalo candidates, we accept the MT prediction at the 95% confidence level, in which at least 95% of the AGN and pulsar sources agree on the MT decision. Otherwise, the sources remain without a prediction. Such threshold value is set based on the results explained in the previous section. In total, predictions for the 380 unassociated 3FGL sources at $|b| > 20^\circ$ suggest that 281 sources are AGN candidates and 69 sources are pulsar candidates with the 95% efficiency rate. The remaining 30 sources at $|b| > 20^\circ$ are left without a firm prediction. While most of the Galactic sources, except for pulsars, are concentrated at $|b| < 20^\circ$ among the associated sources, there are 3 associated Galactic sources at $|b| > 20^\circ$: 2 globular clusters and 1 pulsar wind nebula. In order to better understand the nature of the remaining 30 objects it is desired to compute their outlyingness, which is a measure of how far away a source is from its closed class. The Mahalanobis distances in the unit space directly present the outlyingness from the normal class. Table 1 presents the top 12 outliers among high-latitude ($|b| > 20^\circ$) unassociated sources in 3FGL with the Mahalanobis distances in the AGN unit space and the pulsar unit space. These sources are relatively faint sources with the source significances of $4-9\sigma$. Dark matter subhalo candidates proposed by Bertoni *et al.* (2015) [9] using a theoretical approach were not included in Table 1. As the outliers in Table 1 have the relatively small Mahalanobis distances in the AGN unit space, there may be a possibility that some of the outliers are AGNs.

**Table 1:** Top 12 outliers among high-latitude ($|b| > 20^\circ$) unassociated sources in 3FGL

| Source | $\ell$ | $b$ | $D$ (AGN) | $D$ (pulsar) |
|---|---|---|---|---|
| 3FGLJ1234.7-0437 | -64.913 | 57.996 | 1.975 | 6.214 |
| 3FGLJ0240.0-0253 | 174.600 | -54.492 | 1.895 | 5.988 |
| 3FGLJ2258.2-3645 | 3.903 | -64.252 | 1.659 | 6.052 |
| 3FGLJ1616.8+5846 | 89.516 | 42.688 | 2.536 | 5.679 |
| 3FGLJ2142.6-2029 | 31.142 | -46.557 | 2.161 | 5.764 |
| 3FGLJ2250.3+1747 | 86.354 | -36.331 | 1.831 | 5.843 |
| 3FGLJ0258.2+3555 | 149.895 | -20.218 | 1.963 | 5.615 |
| 3FGLJ1258.4+2123 | -41.094 | 84.038 | 2.281 | 5.316 |
| 3FGLJ1250.2-0233 | -57.656 | 60.307 | 1.867 | 5.433 |
| 3FGLJ2006.5-0939 | 32.637 | -21.030 | 1.880 | 5.400 |
| 3FGLJ0251.1-1829 | -158.133 | -61.166 | 1.557 | 5.457 |
| 3FGLJ1334.3-4152 | -48.569 | 20.294 | 1.599 | 5.345 |
| 3FGLJ0514.6-4406 | -110.526 | -35.393 | 1.699 | 5.291 |
| 3FGLJ0434.3-1411c | -149.264 | -36.714 | 1.580 | 5.202 |
| 3FGLJ0420.4+1448 | 179.885 | -24.215 | 2.073 | 4.692 |
| 3FGLJ1330.4+5641 | 112.329 | 59.630 | 2.161 | 4.600 |

## 6. Conclusion

In recognizing source classes for unassociated gamma-ray sources of the 3FGL, we applied the Mahalanobis-Taguchi method that is a robust data mining technique. This method has a capability to recognize 80 % of the AGNs in the sample of the associated sources, while having a contamination of sources incorrectly labeled as AGNs of 9.3 %. This fraction is significantly better than the previous report of 11 % [4]. To recognize 80 % of the pulsars, a contamination of sources incorrectly labeled as pulsars is 5.2 %. In this paper, we suggest the source classification for the unassociated gamma-ray sources in 3FGL using the MT method. Among 380 unassociated 3FGL sources at $|b| > 20^{\circ}$, we listed unclassified sources left without a firm prediction. While theoretical approaches start with ad hoc theoretical dark matter spectra and non-variable, high-significance unassociated sources, this approach could give us an another useful method to search for dark matter Galactic subhalos.

## References

[1] The Fermi-LAT Collaboration, axXiv:1501.02003v2, 2015.

[2] M.Ackermann *et al.*, Astrophys. J. 753, 83, 2012.

[3] N.Mirabal *et al.*, Mon. Not. R. Astron. Soc. 424, L64-L68, 2012.

[4] M.Doert & M.Errando, Proc. of 33rd ICRC (Rio de Janeiro), 764, 2013.

[5] A.Berlin & D.Hooper, Phys. Rev. D 89, 016014, 2014.

[6] G.Taguchi & J.Rajesh, The Indian Journal of Statistics 62, 233-248, 2000.

[7] T.Aoki, K.Yanagisawa, K.Yoshida, JPS Conf. Proc. 1, 013105, 2014.

[8] M.Ackermann *et al.*, Astrophys. J. Suppl. Ser. 188, 405, 2010.

[9] B.Bertoni, D.Hooper, & T.Linden, axXiv:1504.02087v1, 2015.