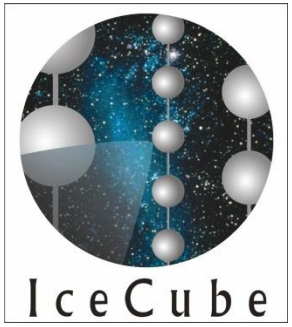
Event Selection with a Boosted Decision Tree



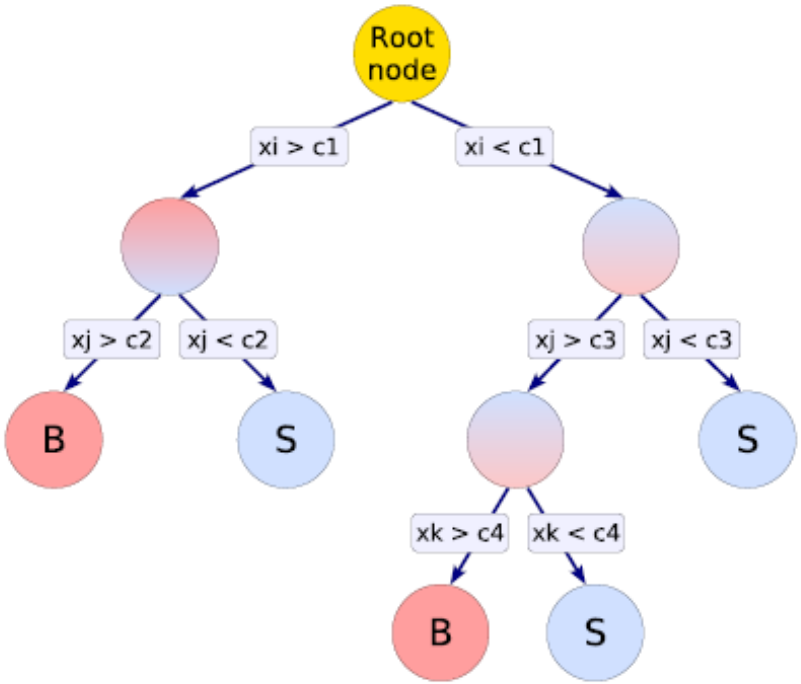
Warren Huelsnitz
University of Maryland

Boosted Decision Tree

- Reference: tmva.sourceforge.net/docu/TMVAUsersGuide/pdf
- Theoretical optimal performance of BDT not as good as some other classifiers. But, easy to get reasonably good results with a BDT.
- Straight cuts carve out a “signal-like” hypercube in multidimensional parameter space; BDT finds multiple hypercubes.
- BDT does not find functional dependencies between parameters, such as a properly set-up neural net would, (but you can specify functional dependencies among variables when you provide the input variables and expressions).

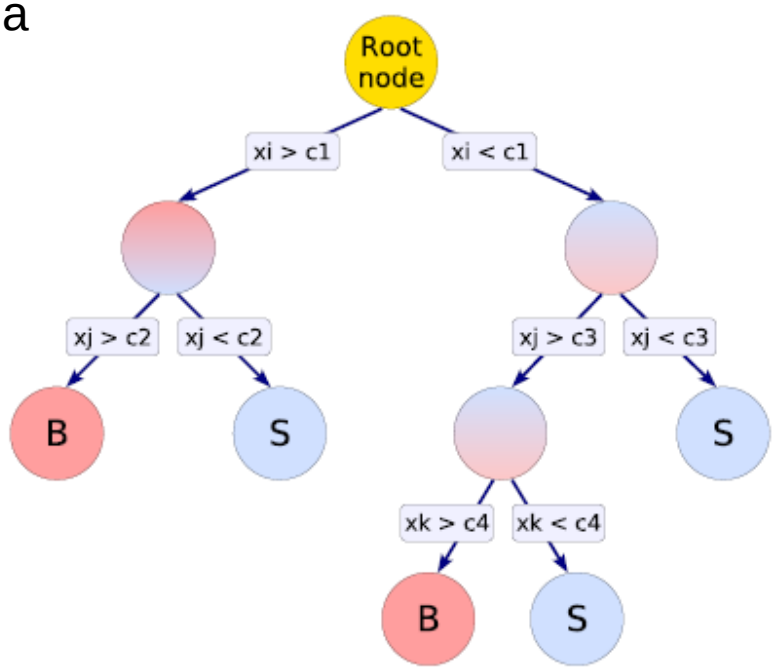
Boosted Decision Tree

- User specifies:
  - Signal and background event input files
  - Signal and background event weights
    - Overall for signal and for background
    - Can also use expressions for variable event weights
  - Input variables (can also be expressions of variables)
  - Number of training signal and background events



Boosted Decision Tree

- Can also specify allowed ranges for variables
- Several options for growing of forest, such as
  - Number of trees in forest
  - Various node splitting criteria
  - Pruning
  - Min events per node
  - Max # split levels
  - Decorelation of variables
  - Boosting method



Boosted Decision Tree

- Boosting stabilizes the response of the classifier to fluctuations in the training sample:
  - AdaBoost: increase weights of events misclassified in current tree before next tree is created
  - Bagging: resampling with replacement
- Final decision is based on a (weighted) majority vote of the individual trees:

$$Y_{BDT}(\tilde{x}) = \sum_{i \in forest} \ln(\alpha_i) \cdot h_i(\tilde{x})$$

Y_{BDT} = BDT score for the event

$\tilde{x}$ = values of the input variables for the event

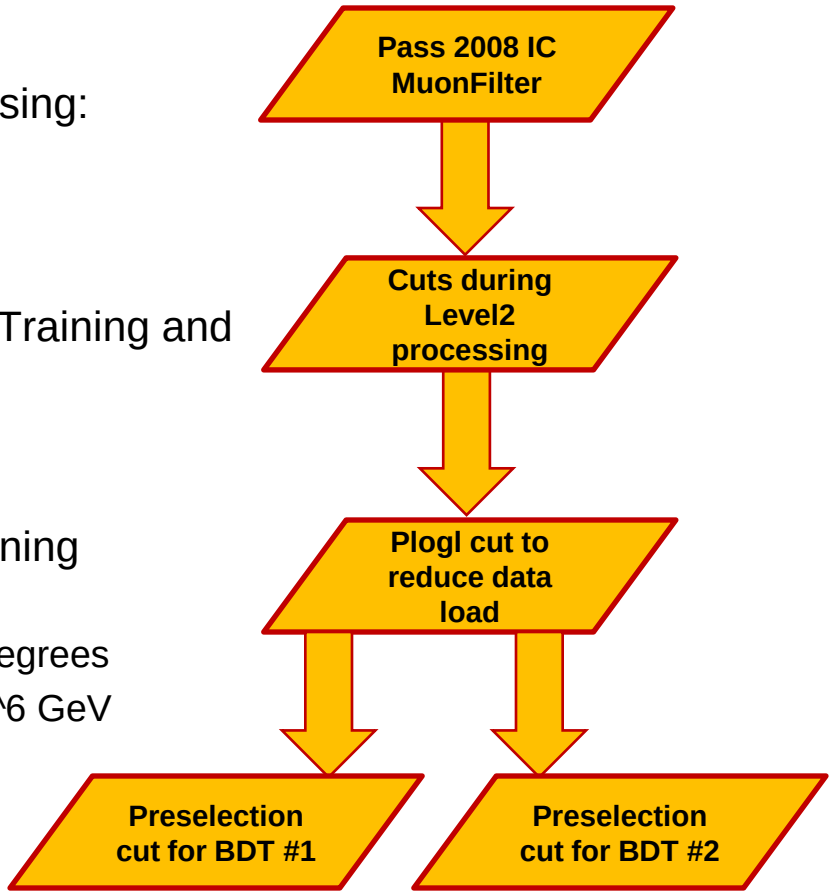
α_i = fraction correctly classified in tree i

h_i = decision result of tree i (1 if in "signal" node, 0 if in "background" node)

- Output score is normalized to values between 0 and 1; you decide what value of the BDT score to cut on

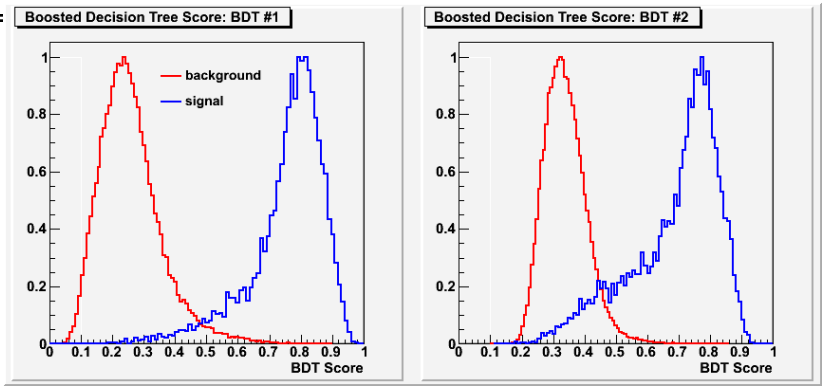
Application to IC40 Atmospheric Neutrinos

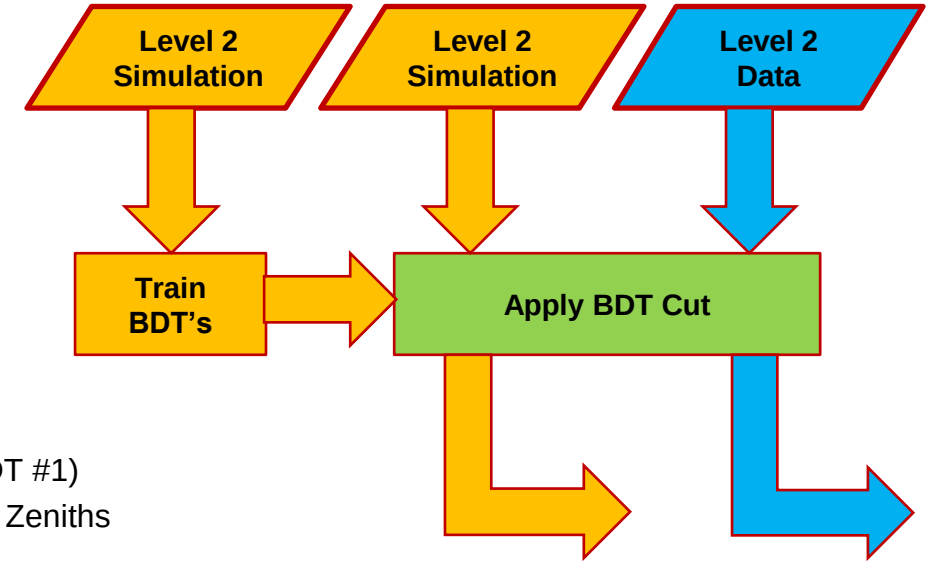
- Cuts Prior to Boosted Decision Trees:
  - 2008 (40-str) IceCube Muon Filter
  - Level 2 Processing and post-processing:
    - SPEFit32_Zenith >= 80
    - SPEFit32_Rlogl <= 12
    - SPEFit32_Logl/(NCh-2.5) < 8
  - Pre-Selection Cuts to improve BDT Training and Performance (also applied to data)
    - BDT #1: LineFit_LFVel >= 0.2
    - BDT #2: Split Track Zeniths >= 80
  - Cuts on Signal Events Used for Training (not applied to data)
    - |MC_Zenith – Reco_Zenith| <= 10 degrees
    - For BDT #1 only, Nu_Energy <= 10^6 GeV



Application to IC40 Atmospheric Neutrinos

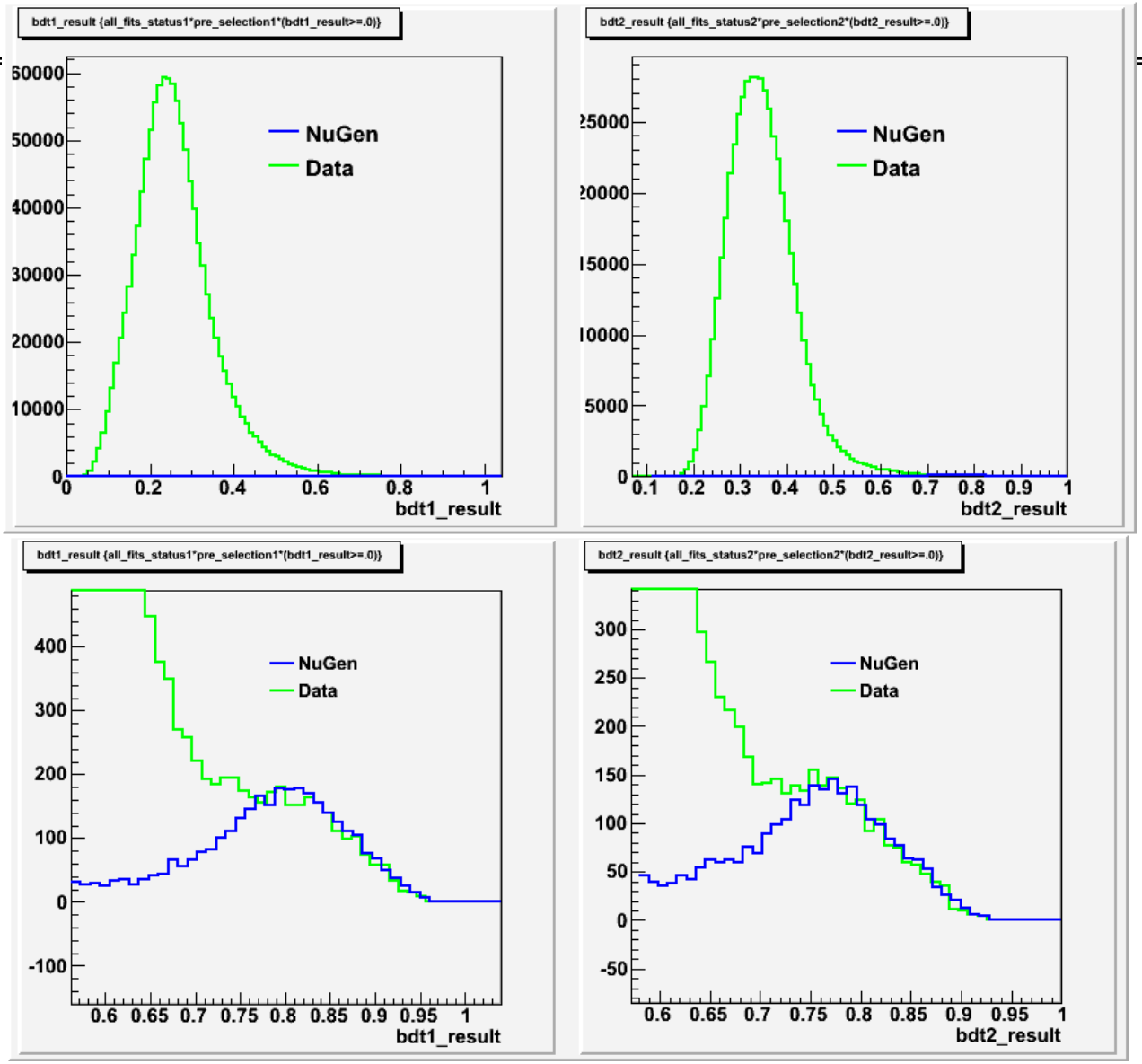
- Two BDT's used:
  - Both BDT's reject all background
  - Event retained if it passes either BDT
- Boosted Decision Tree Cut: (BDT1 || BDT2)
- BDT variables:
  - Parab_sigma
  - Rlogl
  - PLogl
  - SmoothAll
  - NDirC
  - LDirC
  - NDirC/NHits
  - Bayes_Logl – Logl
  - Umbrella_Logl - Logl
  - SingleLLH_Zenith – LF_Zenith
  - NChannel
  - Nstring (BDT #1 only)
  - LF Geo and Time Split Track Zeniths (BDT #1)
  - LF and SPE16 Geo and Time Split Track Zeniths (BDT #2)



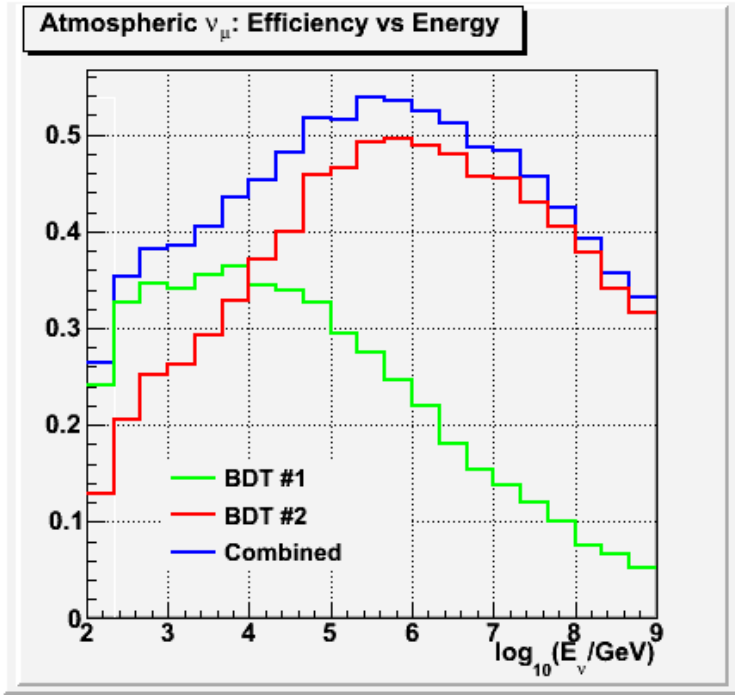


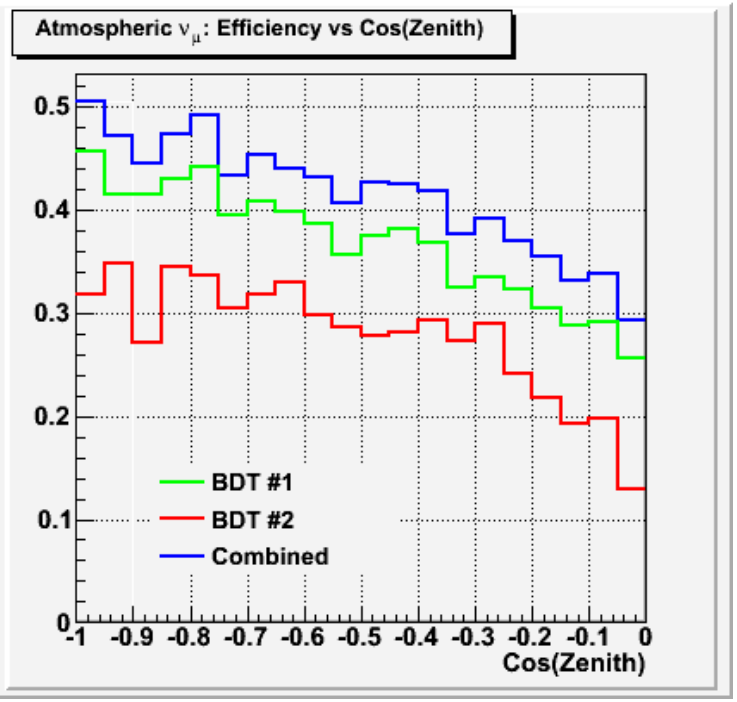
BDT Scores for Data

BDT Scores for Burn Sample (top)
Zoomed in to region of interest (bottom)

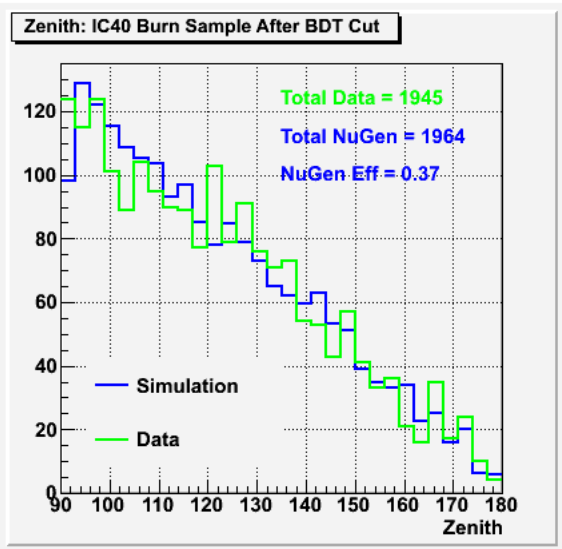


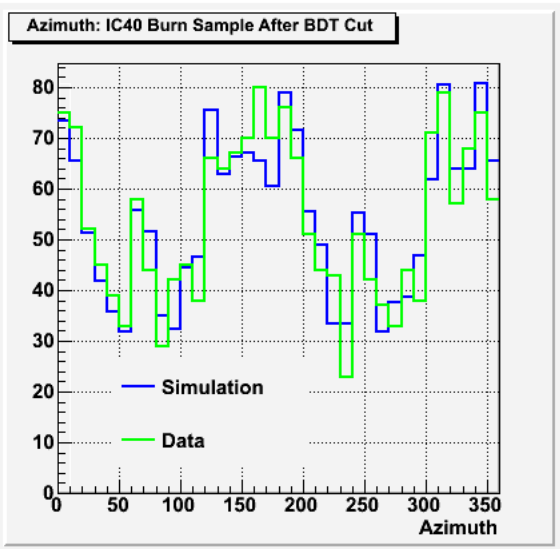
Net Efficiency of the Boosted Decision Trees (for ~100% purity)

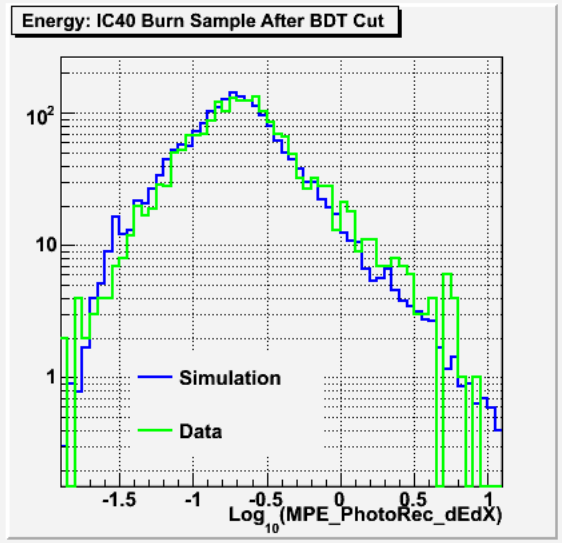


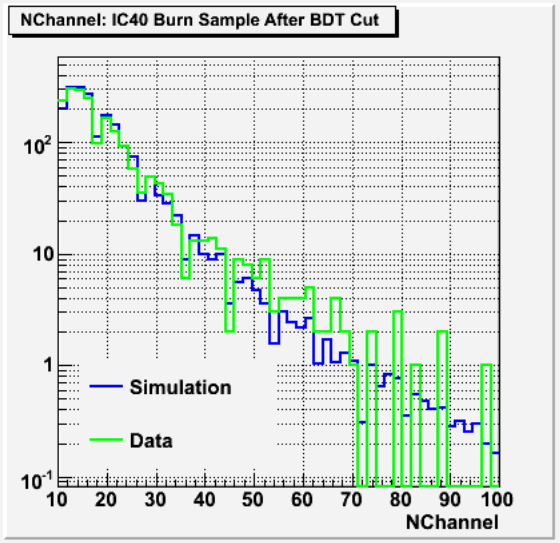


Data and NuGen, after BDT

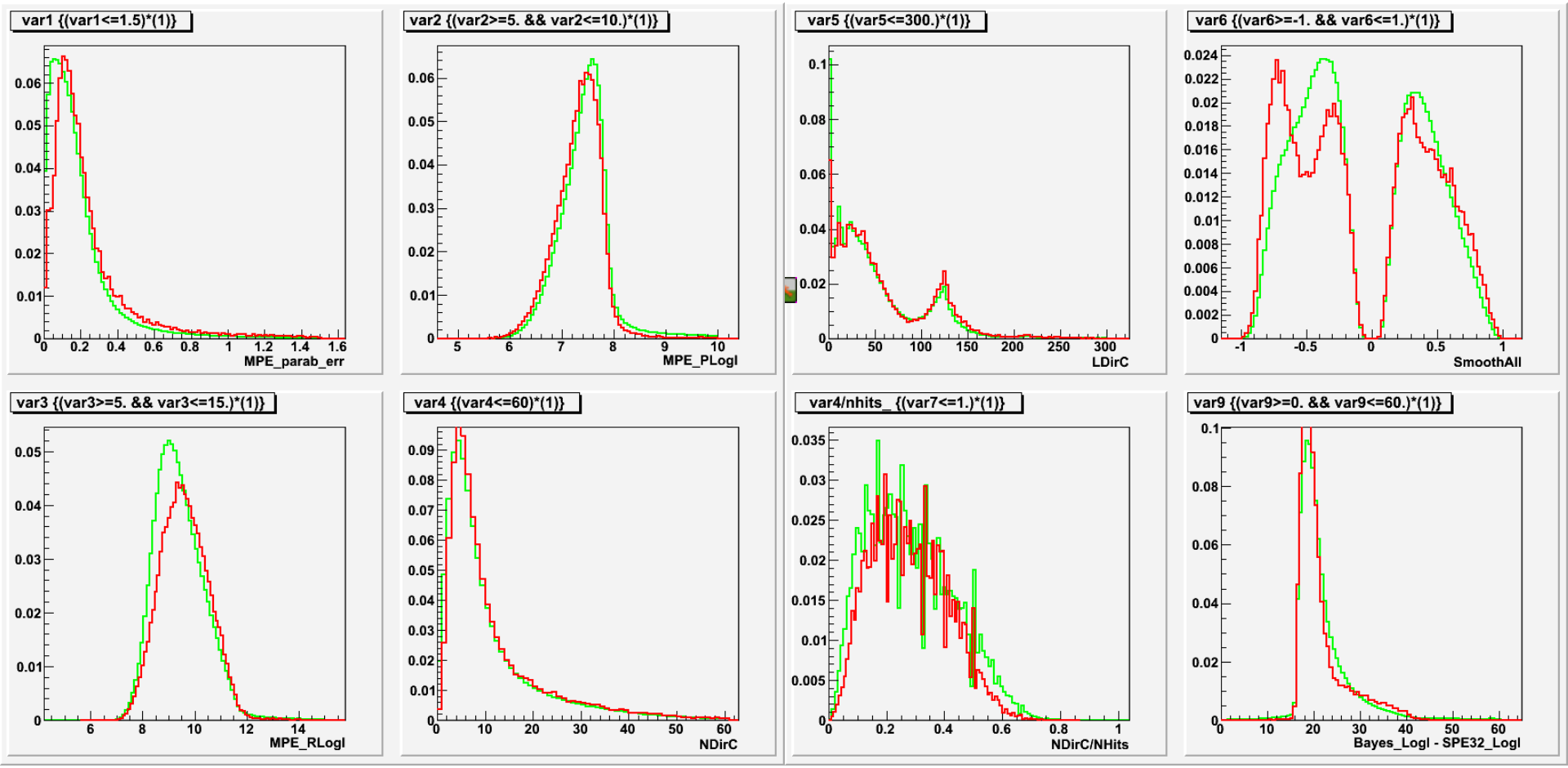




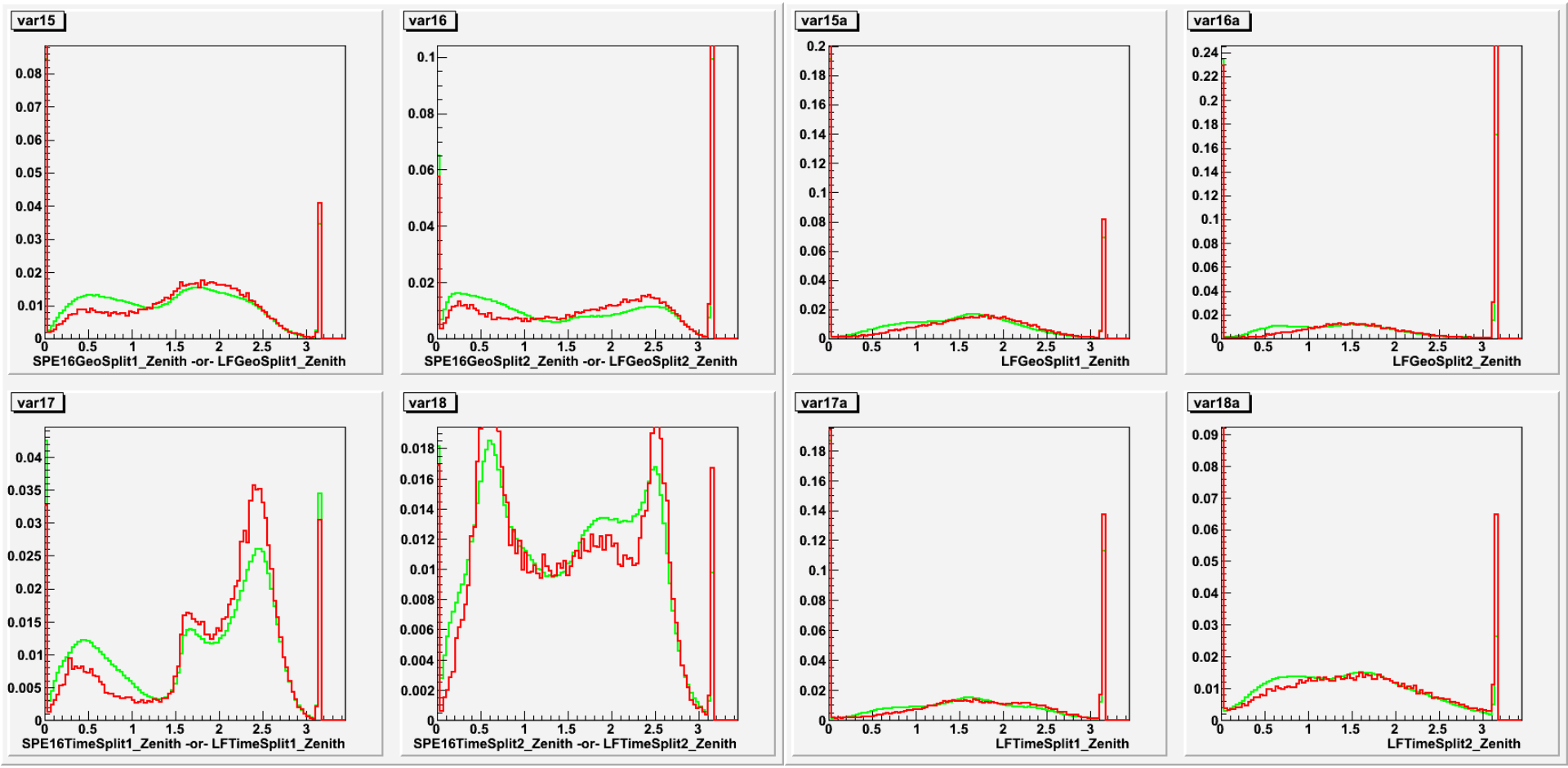




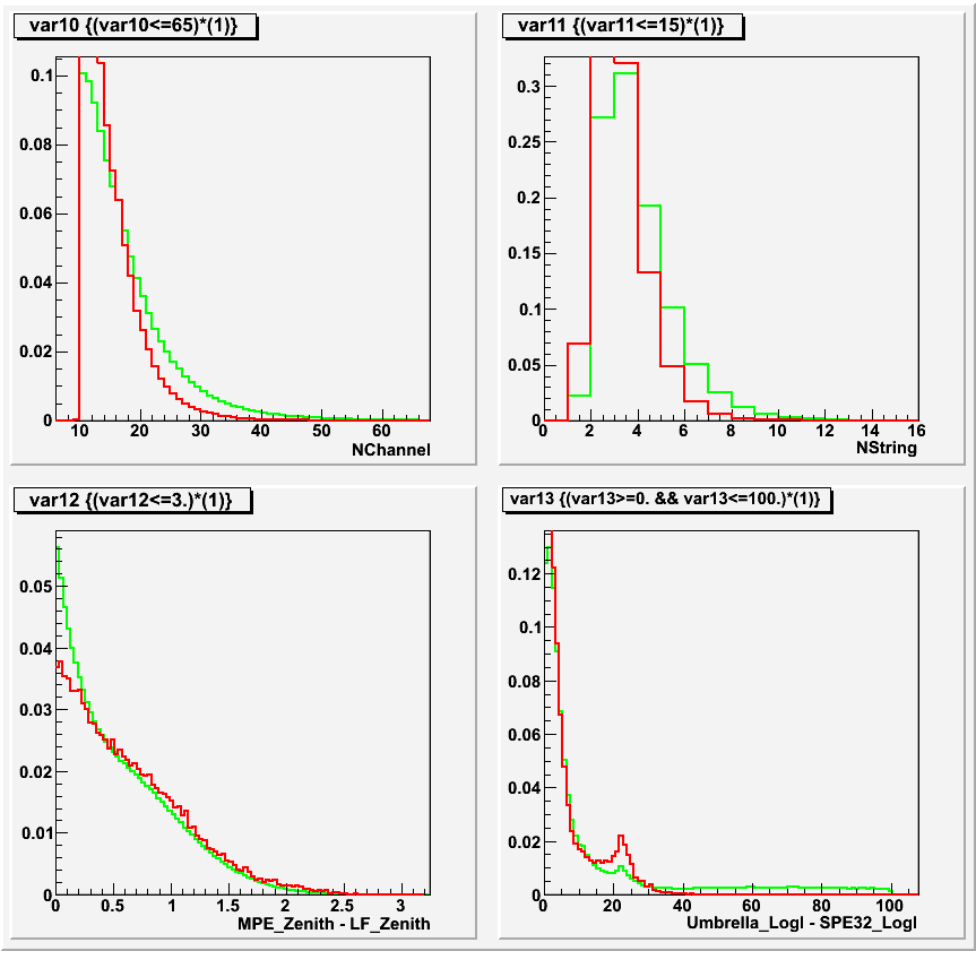
Variables before BDT Cut
Corsika(Red) and Data(Green)

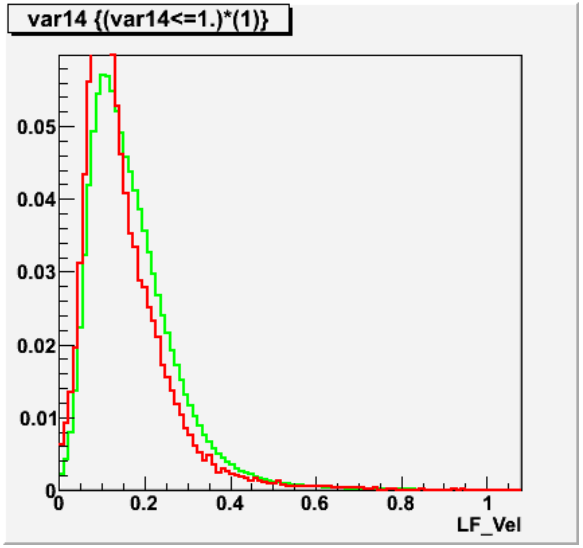


Variables before BDT Cut
Corsika(Red) and Data(Green)

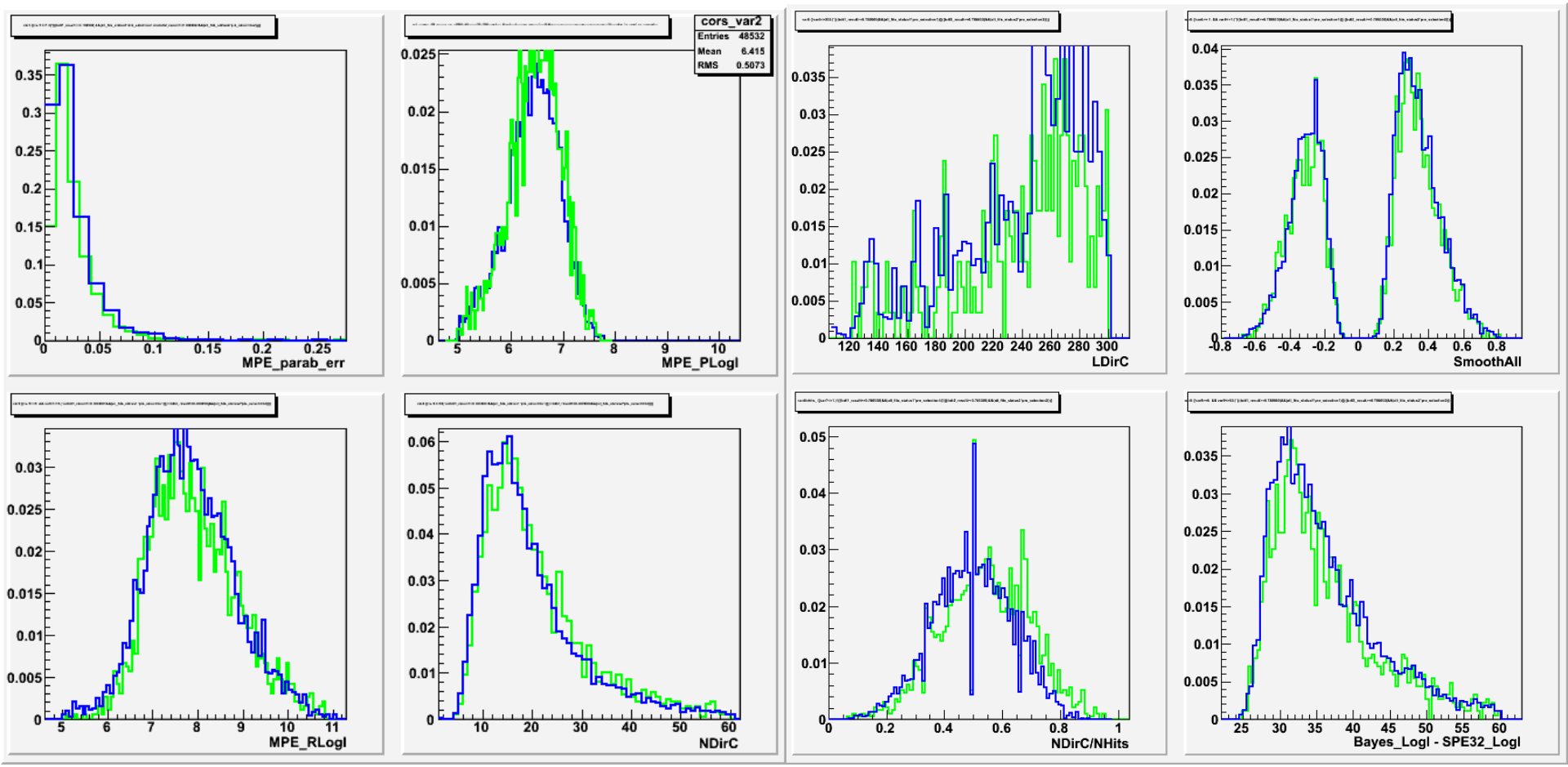


Variables before BDT Cut
Corsika(Red) and Data(Green)

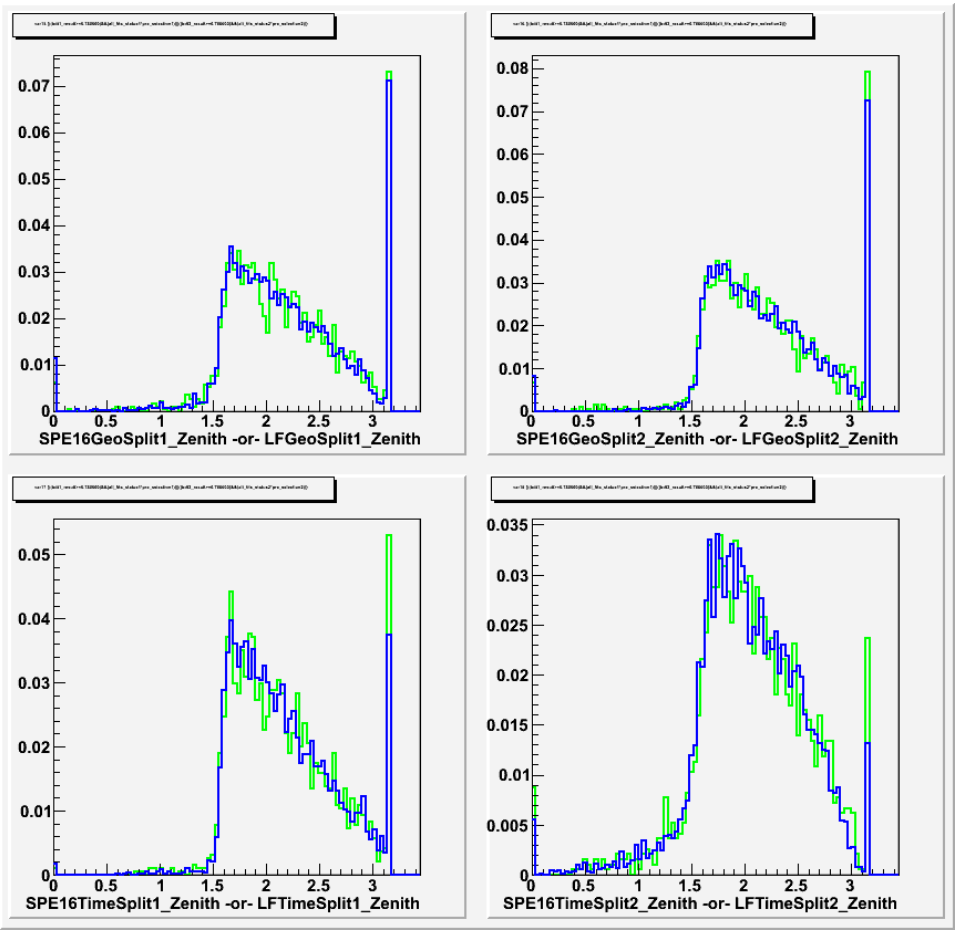


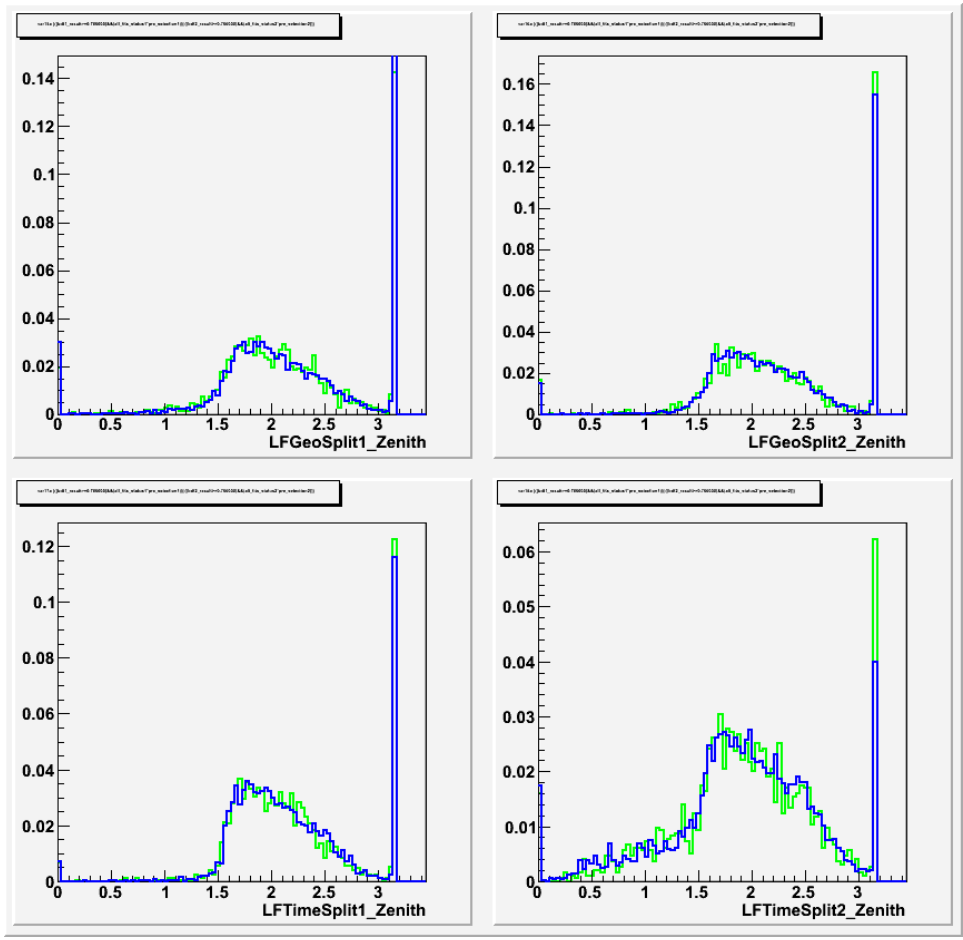


Variables after BDT Cut
NuGen(Blue) and Data(Green)



Variables after BDT Cut
NuGen(Blue) and Data(Green)





Variables after BDT Cut
NuGen(Blue) and Data(Green)