Decision tree analysis update

Mayly Sanchez
Harvard University





Minos Collaboration Meeting 09/16/04

A bit of context

- Before Ely: old Neugen, reco R1.0, no nutaus. With this we got results comparable with Numi -714.
- At Ely: new Neugen, reco R1.9, nutaus, input from Caldet. Too many changes, hard to track problems, results significantly worse.

Objective so far:
Get comparable results to Numi-714

Before Ely

Results comparison

	Nue	NC	Numu	Bnue	FOM
no cut 20:1	7	22	10	10	1.08
trk cut 25:1	8	25	5	12	1.23
Tingjun	6.1	27.1	3.6	7.8	0.98
Alex	14	139	67	22	0.93
numi 714 (scaled)	7.1	27.2	3.9	5.6	1.17

What's new since Ely?

- MDC R1.9, bugs in 1.7 fixed.
- Vertex finding from Josh Boehm integrated with reconstruction.
- Fixes in framework: normalization and test sample construction.
- New set of cuts, variables and treatment of multiclass trees.

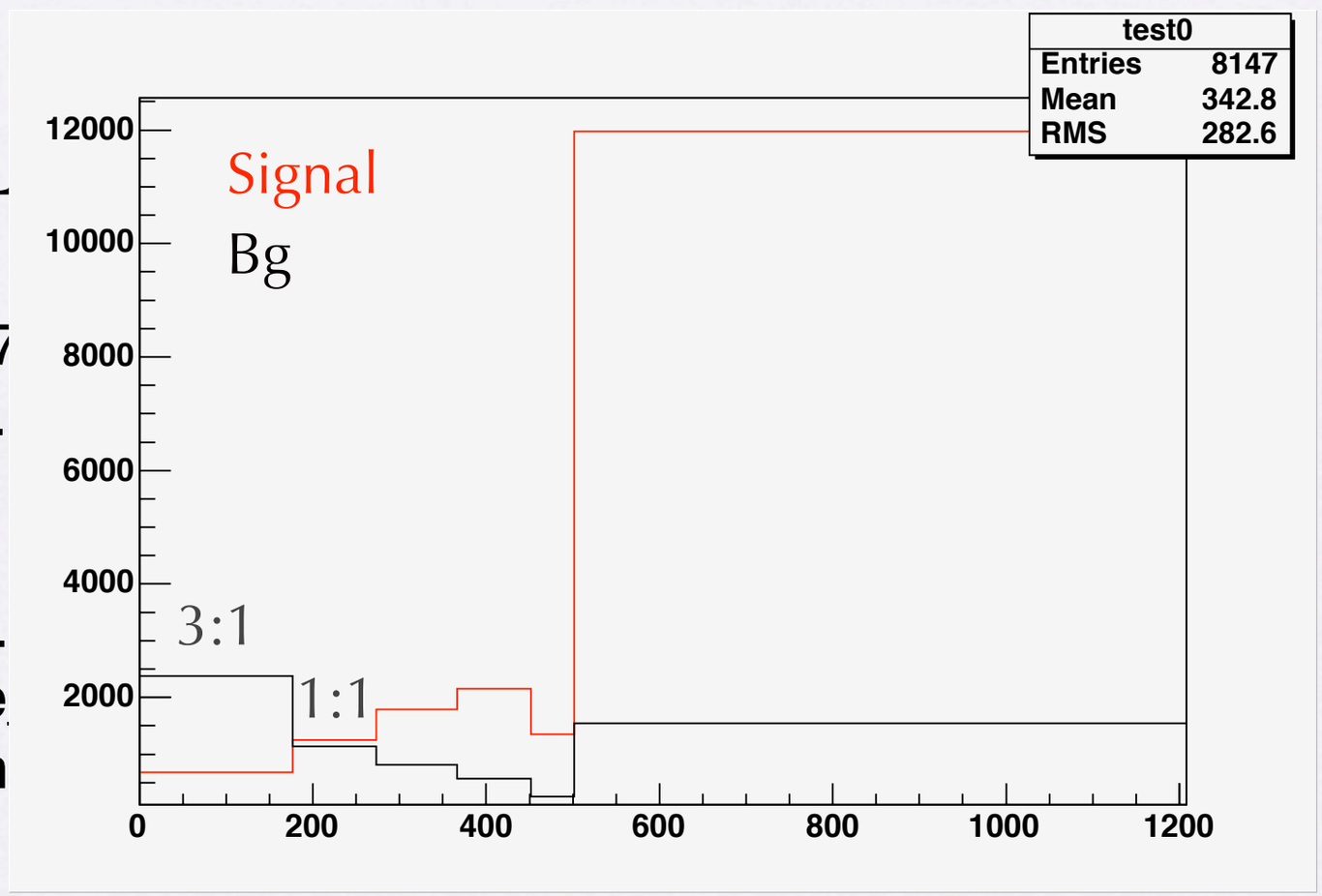
Cuts study

- Idea: cut everything that is obviously background, ie long numus, low energy events, higher energy beam nues.
- Use a set of preliminary decision trees, to optimize these:
  - Low end energy: NC vs CC
  - High end energy cut: Beam nue vs nue, numu
  - Track length cut: numu vs all

Low end energy cut

- NC vs CC: primary purpose start at a reasonable signal/bg ratio.
- Selected cut: ph_pe>273

```
ph_pe <= 366.281
|   ph_pe <= 176.41: nc (22204.0/
|   ph_pe > 176.41
|   |   ph_pe <= 273.472: nc (11867
|   |   ph_pe > 273.472: nc (10035.
ph_pe > 366.281
|   ph_pe <= 501.501
|   |   ph_pe <= 451.762: nc (8704.
|   |   ph_pe > 451.762: numu_nue
|   ph_pe > 501.501: numu_nue_n
```



High end energy cut

- Beam nue vs nue+numu: primary purpose of this cut keep signal range and get good signal/bg for difficult bg.
- Selected cut: ph_pe<1208

ph_pe <= 1208.44

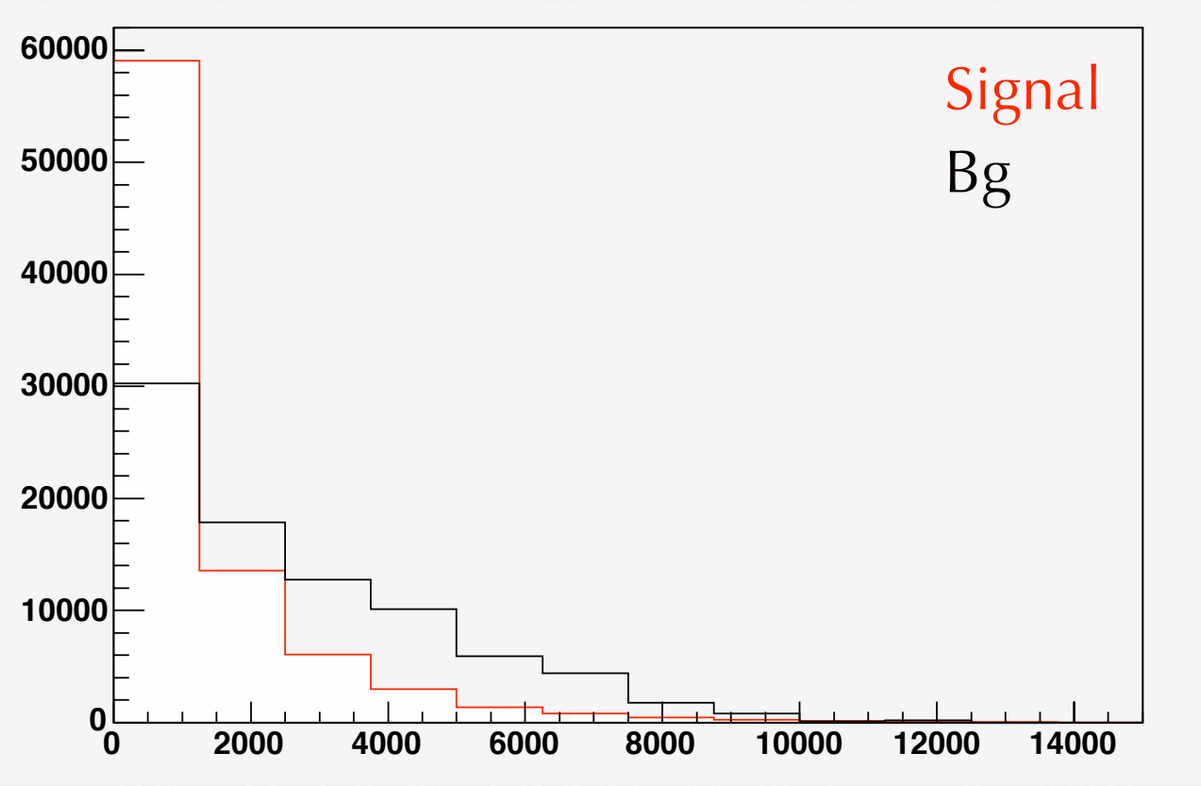
numu_nue	bnue
0.653	0.346

ph_pe > 1208.44

numu_nue	bnue
0.351	0.648

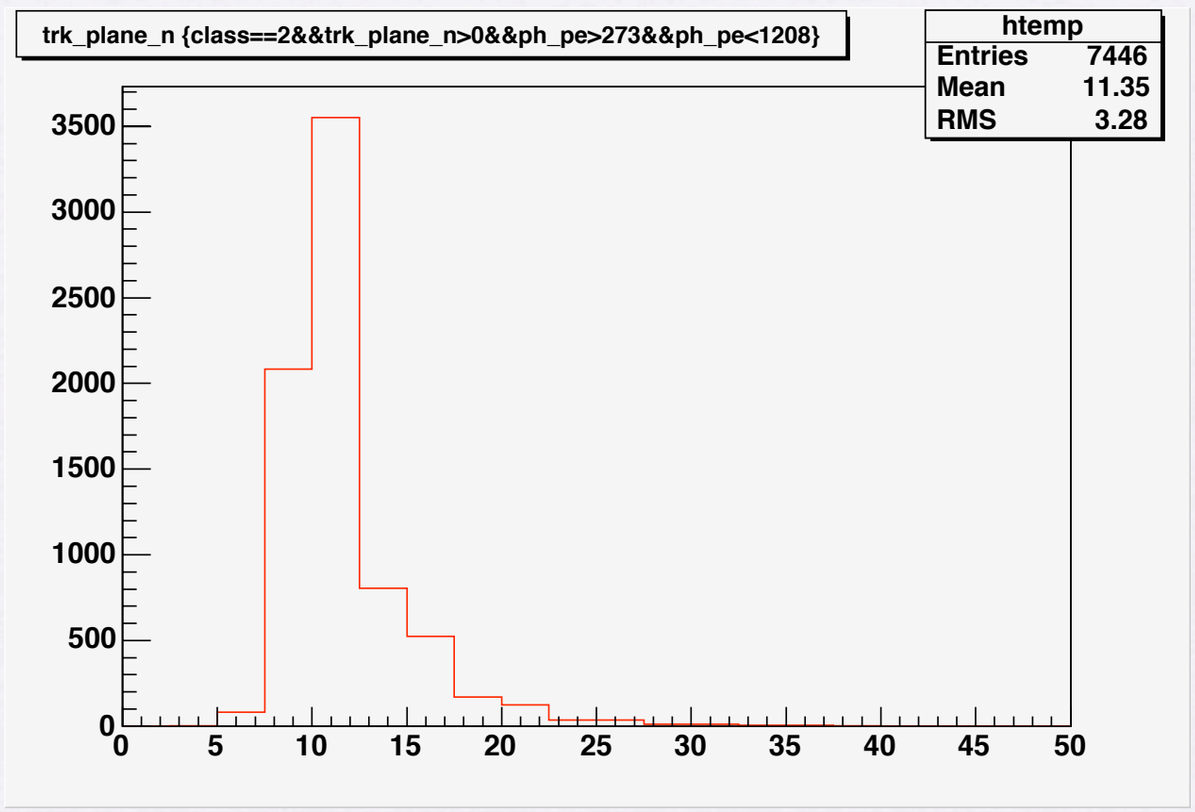
ph_pe is missing

numu_nue	bnue
0.502	0.497



Track length cut

- Numu vs rest: primary purpose eliminate as many numus.
- Selected cut: !=(ntrack==1 && trk_plane_n>23)



With energy cut

trk_plane_n <= 22.5

neg_numu	pos_numu
0.895	0.104

trk_plane_n > 22.5

neg_numu	pos_numu
0.259	0.740

trk_plane_n is missing

neg_numu	pos_numu
0.955	0.044

The sample:

Preliminary cuts on visible energy and track length are applied.

	Nue	NC	Numu	Nutau	Bnue
training	15929	49833	52171	18129	1141
all cuts	8693	18013	8398	3933	330
testing	14507	11739	14436	2224	264
ph_pe>273	13607	6712	13408	2111	255
ph_pe<1208	8277	4635	8257	637	82
trk n < 23	8002	4271	2270	486	78

The sample:

$\Delta m^2 = 2.5\ 10^{-3}$

$(Ue3)^2 = 0.01$

Oscillation weighted

Preliminary cuts on visible energy and track length are applied.

FOM		Nue	NC	Numu	Nutau	Bnue
0.40	testing	28.3	1350.8	3447.3	33.8	87.5
0.37	ph_pe>273	23.9	786.2	3312.1	27.6	84.6
0.48	ph_pe<1208	22.7	538.5	1656.9	20.1	26.9
0.72	trk n < 23	22.1	496.1	403.0	16.9	25.6

Variable selection

- Optimize variable selection for the different problems. Previously used algorithm does not perform well in 3+ class situations.
  - NC vs CC (18)
  - Numu vs Nue+NC (19)
  - Nue vs NC (21)
  - Nue vs Numu (21)
  - Nue+BNue vs Numu/Nutau/NC (29)

  Note: of course they overlap, true total 44

The pool of variables selected

EM shower fit

par_a
par_b
par_e0
e0_pe_ratio
chisq
chisq_ndof
ndof
shwmax
shwmaxplane
conv

Track reco

ntrack
trk_ph_pe
trk_ph_gev
trk_plane_n
trk_plane_ntrklike
ntrklike_ratio
trk_pe_ratio

Shower reco

nshower
shw_ph_pe
shw_nstrip
shw_pe_ratio
shw_nstrip_ratio

Event Tot

ph_pe
nstrip
max_pe_plane
plane_n

Momentum flow

E_Hit_trans
E_Hit_long
E_hits
S_Hit_trans
Ratio_Hit_Far
Ratio_Hit_Energ
ThetaL_Hit_Max
ThetaE_Hit_Max
E_HitT_ratio
E_HitL_ratio
E_Hit_LT_ratio

Transverse

u_mean
u_rms
u_skew
u_asym
u_kurt
u_molrad_peak
u_molrad_vert
v_mean
v_rms
v_skew
v_asym
v_kurt
v_molrad_peak
v_molrad_vert
uv_mean
uv_rms
uv_skew
uv_asym
uv_kurt
uv_molrad_peak
uv_molrad_vert

Multiclass Decision Trees

- Trees are binary in nature, a multiclass problem can be improved by creating trees that separate each of the classes from the rest (1 vs all)
- The pool of variables is used but only the ones useful for each separation will be selected
- In some cases this provides an improvement similar to costly boosting methods

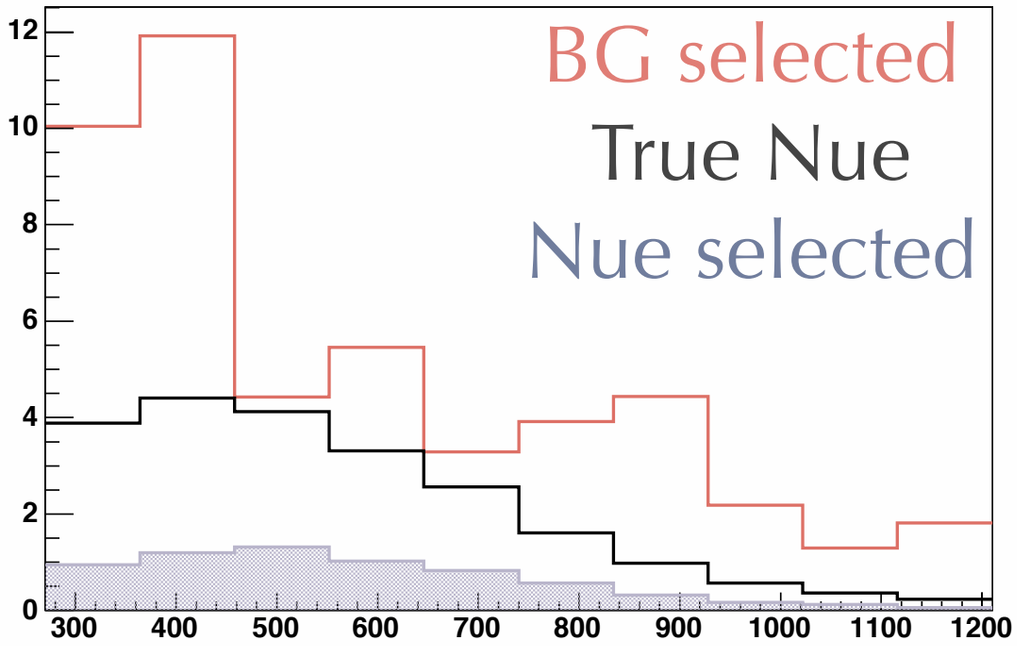
Results comparison

Testing several methods

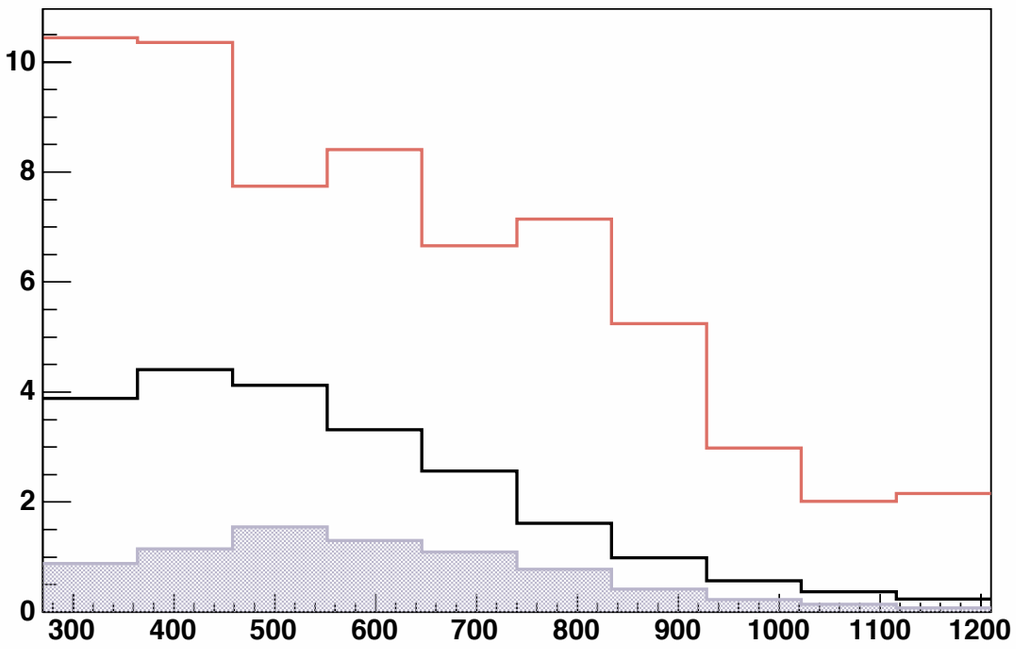
	Nue	NC	Numu	Nutau	BNue	FOM
j48 5:1	8.0	49.5	16.8	3.2	9.6	0.90
multij48 c0.5 20	5.6	23.9	8.5	2.0	6.6	0.87
multij48 c0.5 200	8.5	56.5	16.6	3.2	8.6	0.92
multij48 m1 c0.5 200	6.6	30.2	8.8	2.2	7.6	0.94
Reptree 20	7.6	37.4	12.8	2.8	10.2	0.96
numi 714 (scaled)	7.1	27.2	3.9	5.3	5.6	1.12

Signal and Background

Multij48 M1

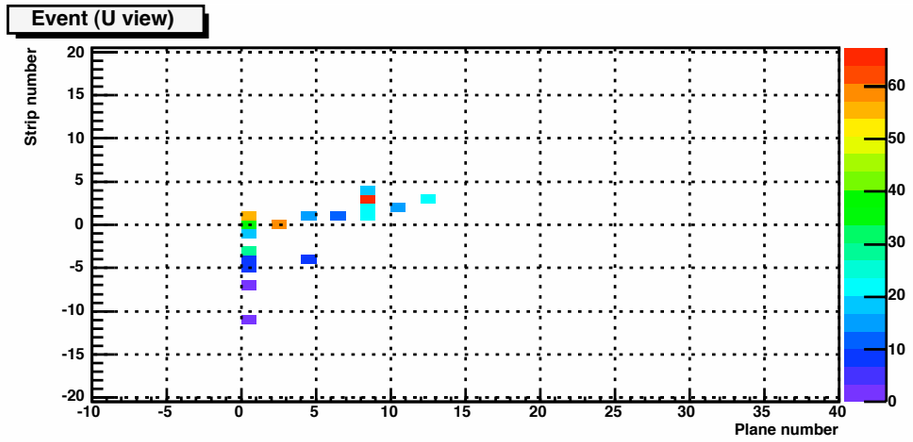


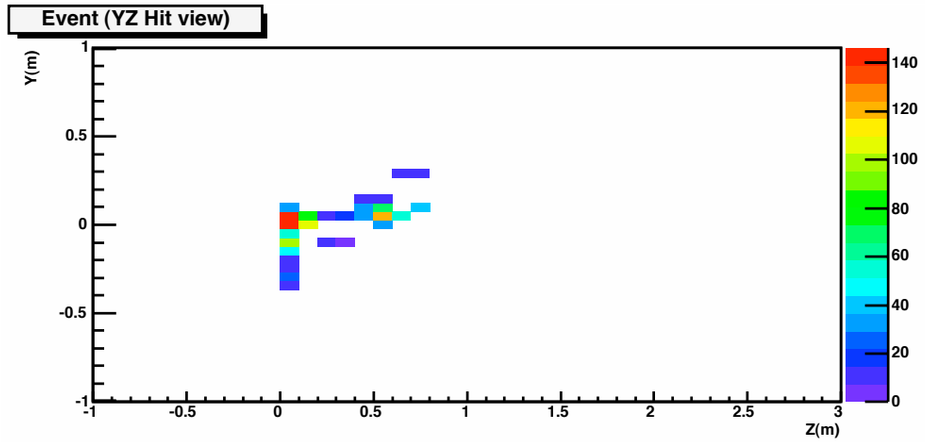
Reptree 20

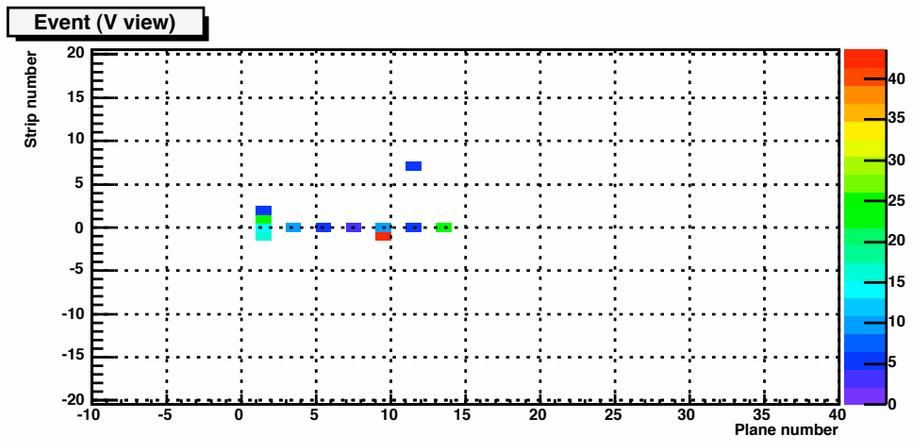


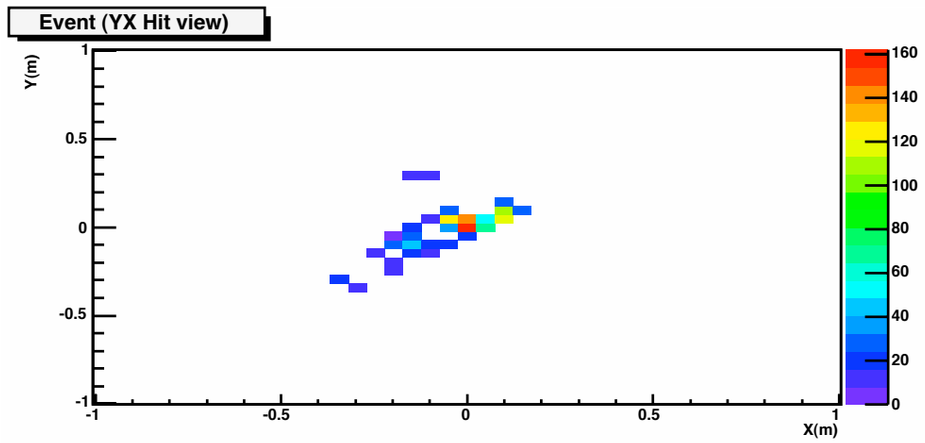
- Comparison of two methods, using 44 variables. Multij48 is using error correction codes to do multiclassing.
- Reptree is using the 1 vs all method for multiclassing.

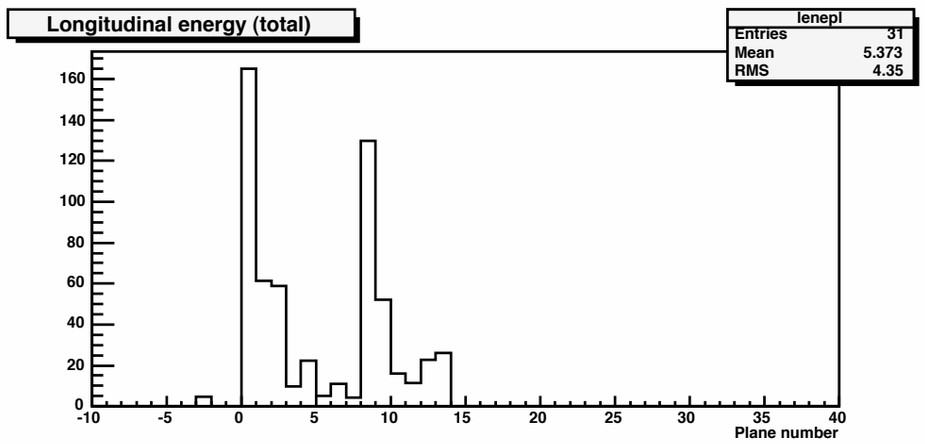
Event pictures

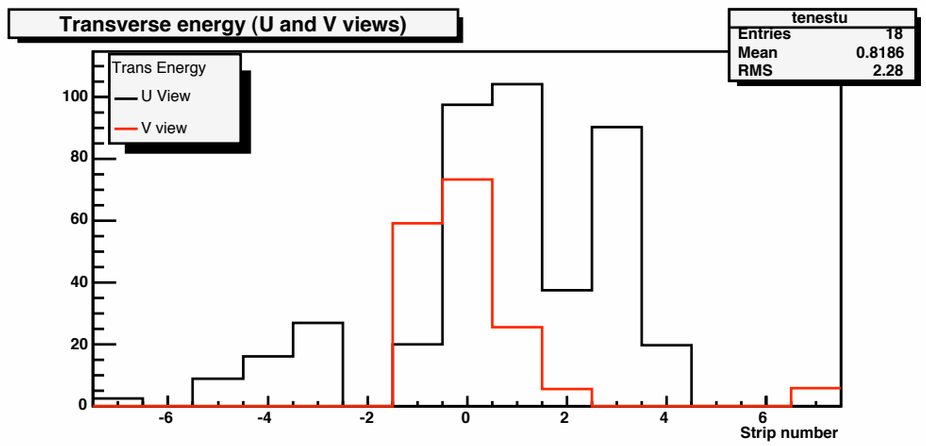

Event (U view)
Strip number
Plane number

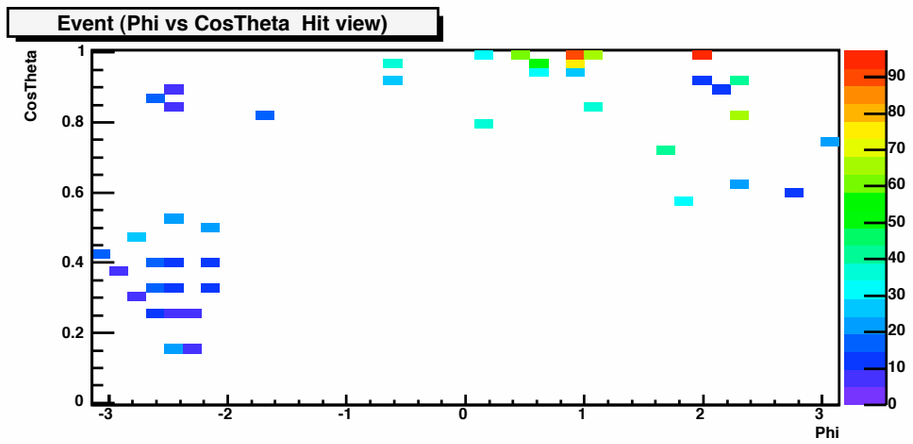

Event (YZ Hit view)
Y(m)
Z(m)

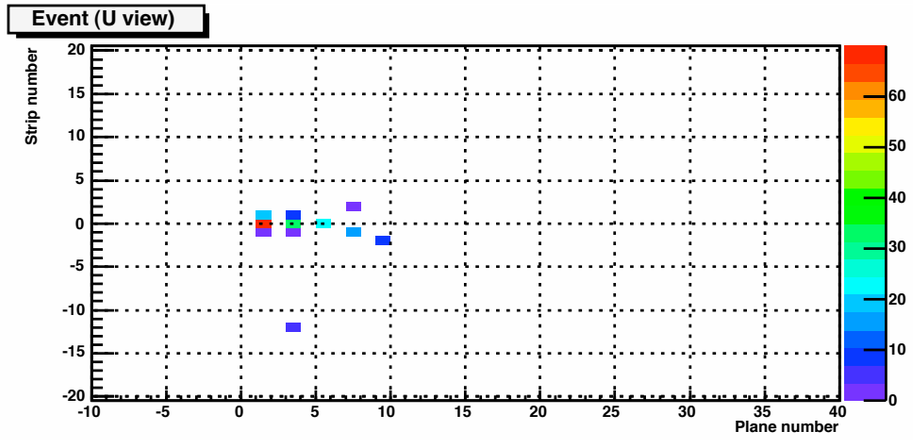

Event (V view)
Strip number
Plane number

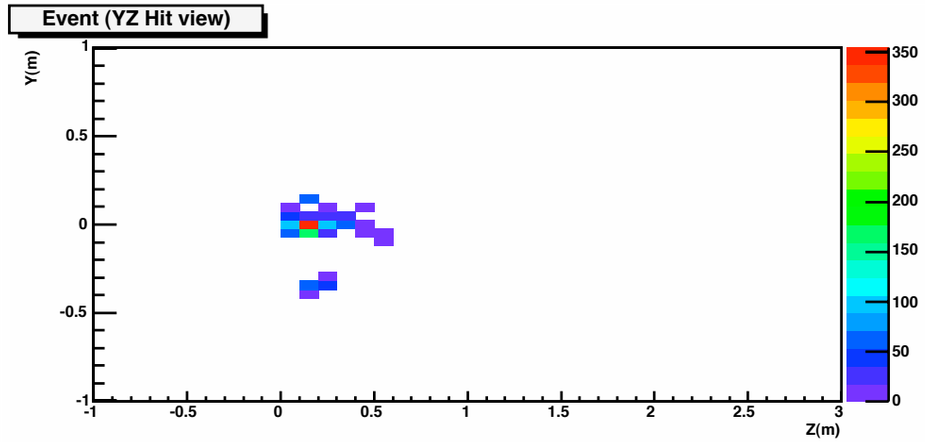

Event (YX Hit view)
Y(m)
X(m)

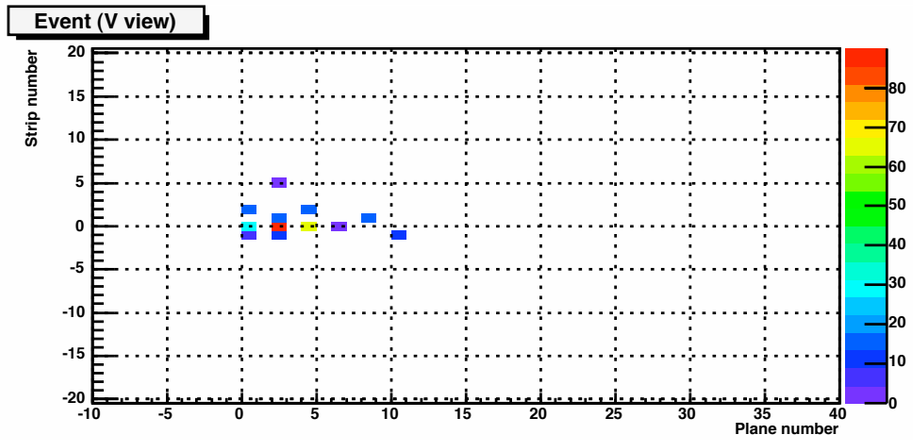

Longitudinal energy (total)
lenepl
Entries 31
Mean 5.373
RMS 4.35
Plane number

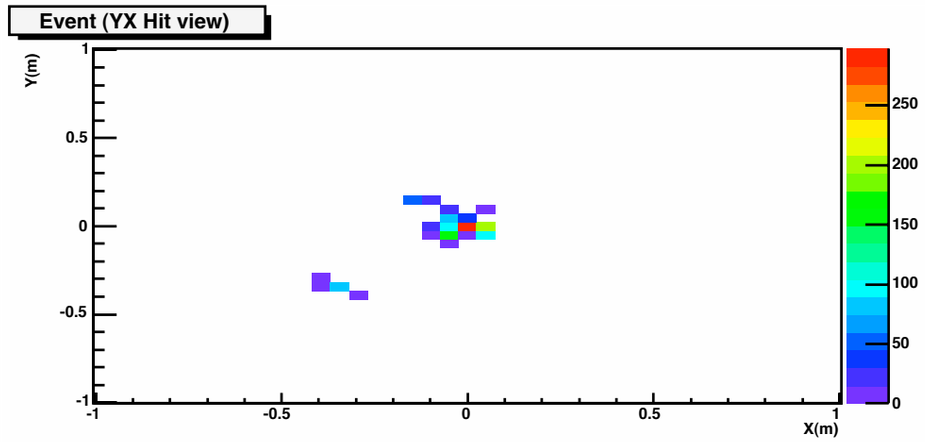

Transverse energy (U and V views)
tenestu
Entries 18
Mean 0.8186
RMS 2.28
Trans Energy
U View
V view
Strip number

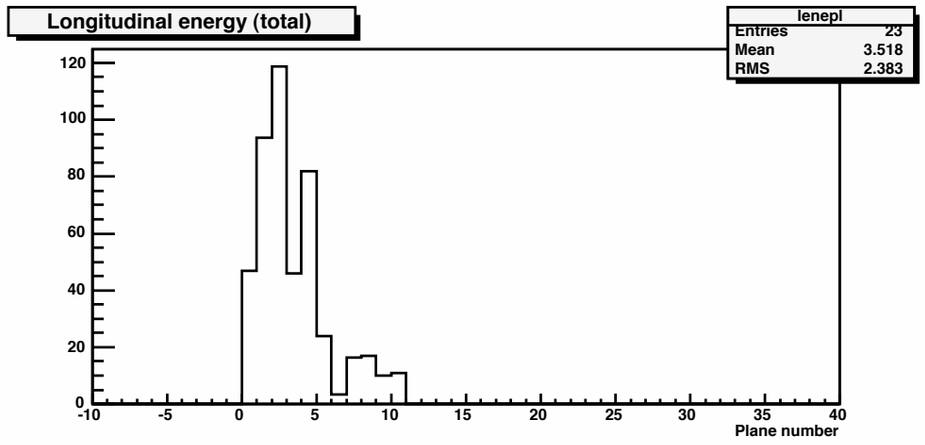

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

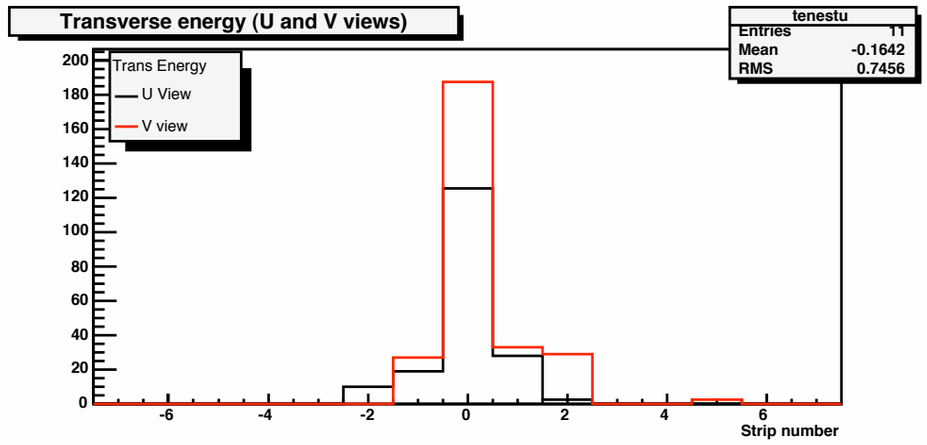

Event (U view)
Strip number
Plane number

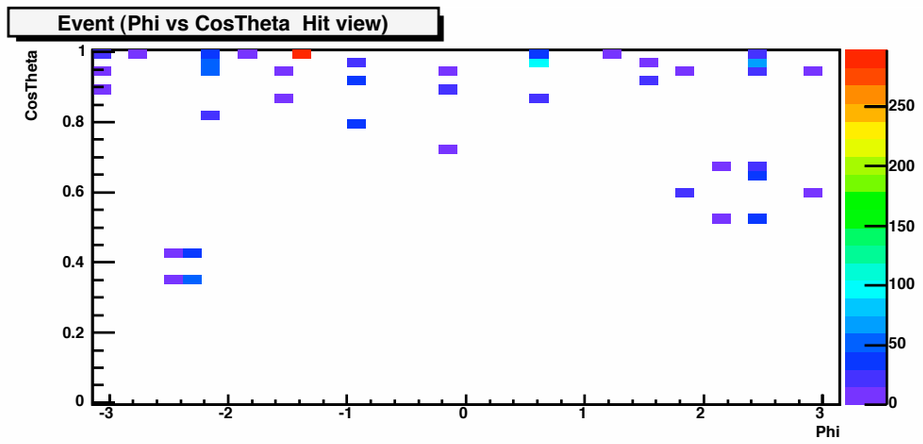

Event (YZ Hit view)
Y(m)
Z(m)


Event (V view)
Strip number
Plane number

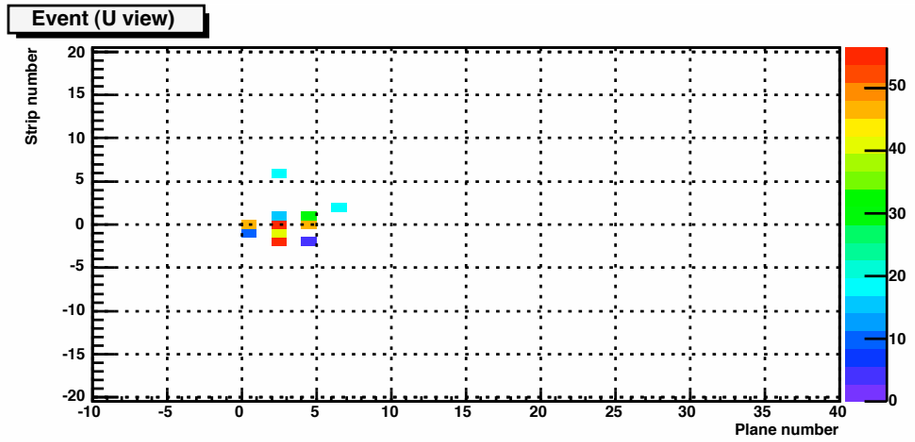

Event (YX Hit view)
Y(m)
X(m)

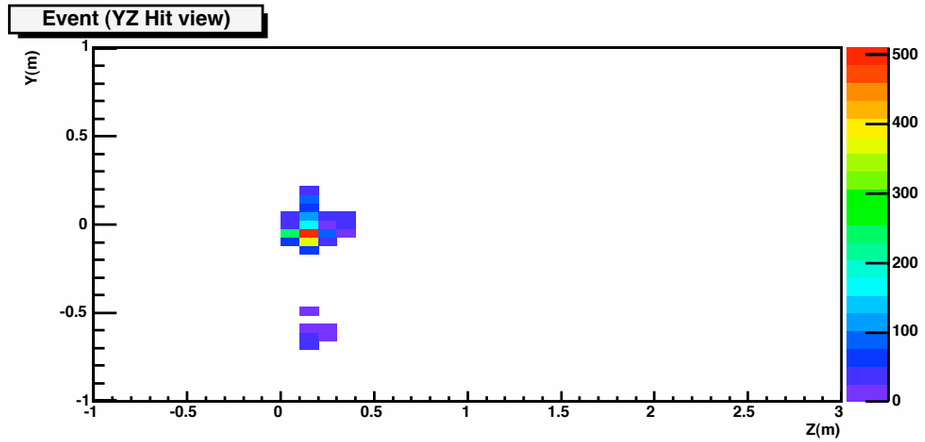

Longitudinal energy (total)
lenepl
Entries 23
Mean 3.518
RMS 2.383
Plane number

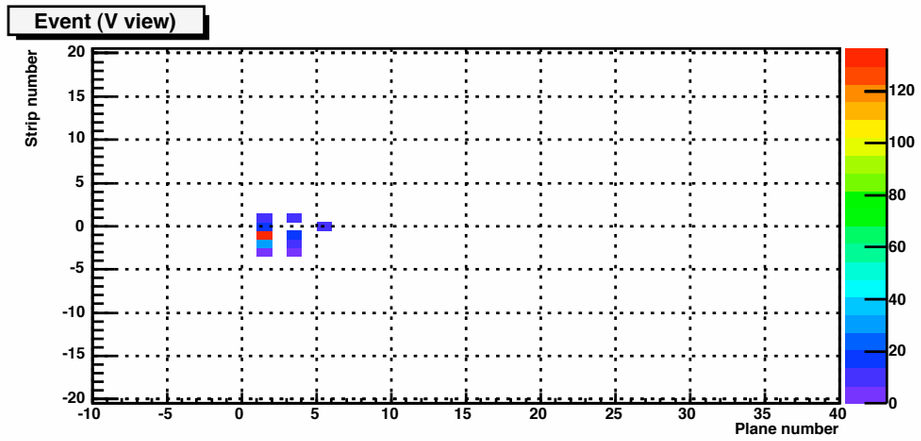

Transverse energy (U and V views)
tenestu
Entries 11
Mean -0.1642
RMS 0.7456
Trans Energy
U View
V view
Strip number

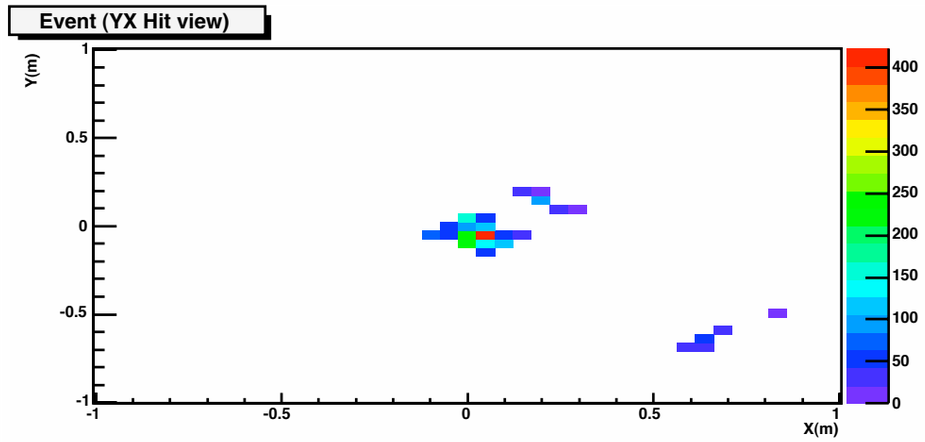

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

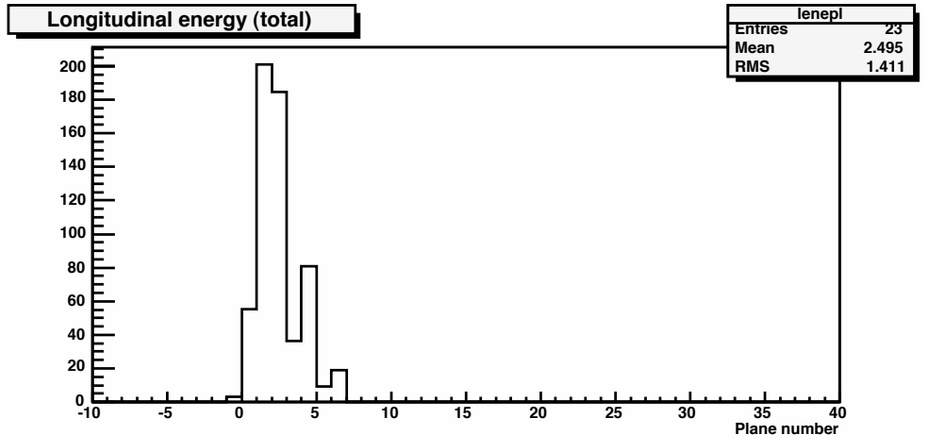


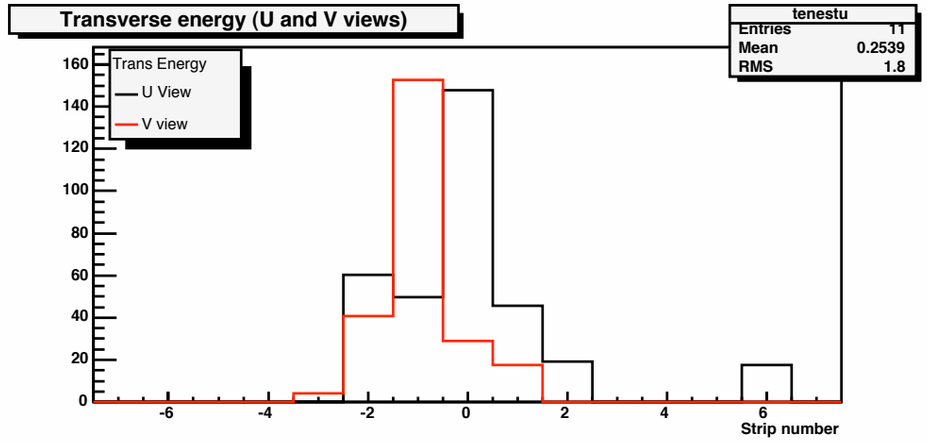

Event (U view)
Strip number
Plane number

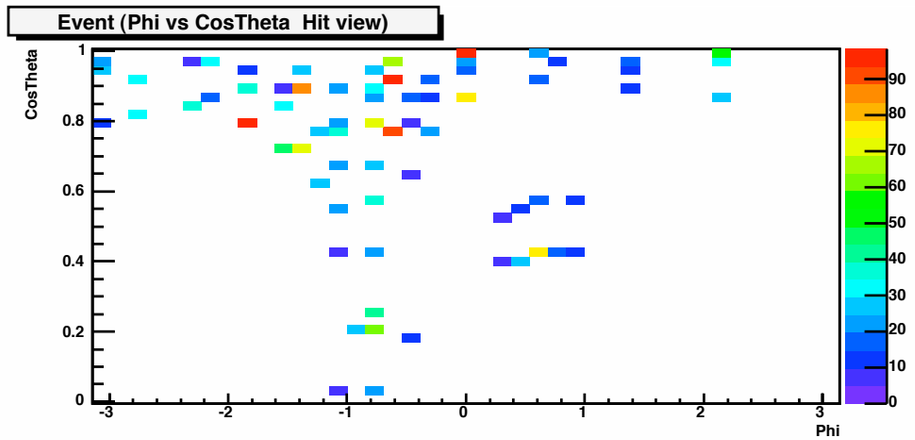

Event (YZ Hit view)
Y(m)
Z(m)

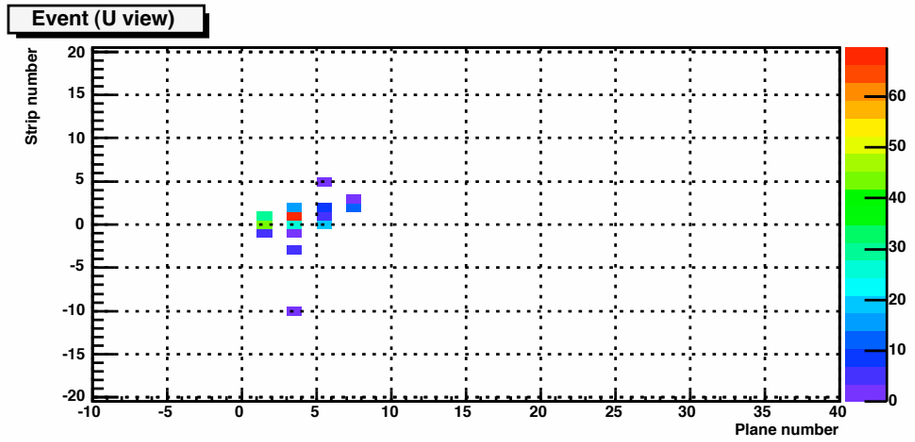

Event (V view)
Strip number
Plane number

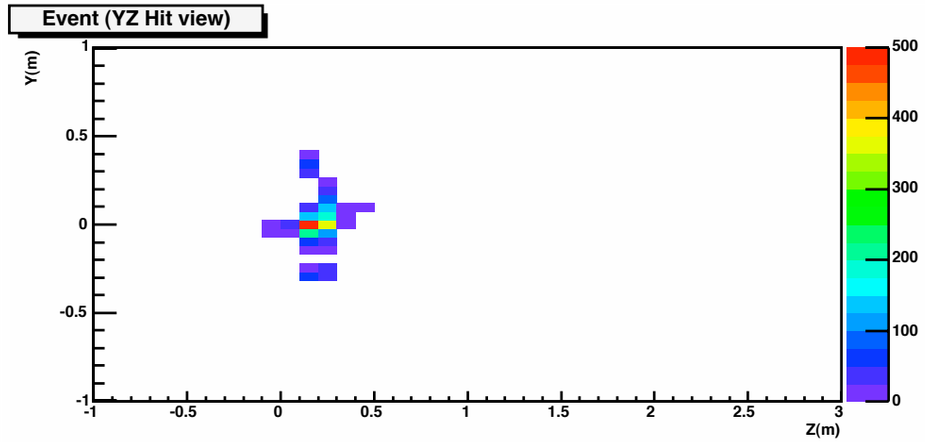

Event (YX Hit view)
Y(m)
X(m)

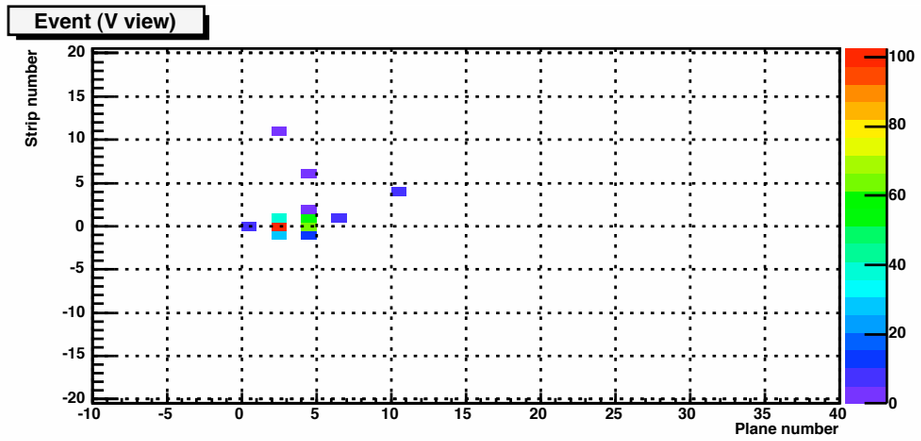

Longitudinal energy (total)
lenepl
Entries 23
Mean 2.495
RMS 1.411
Plane number

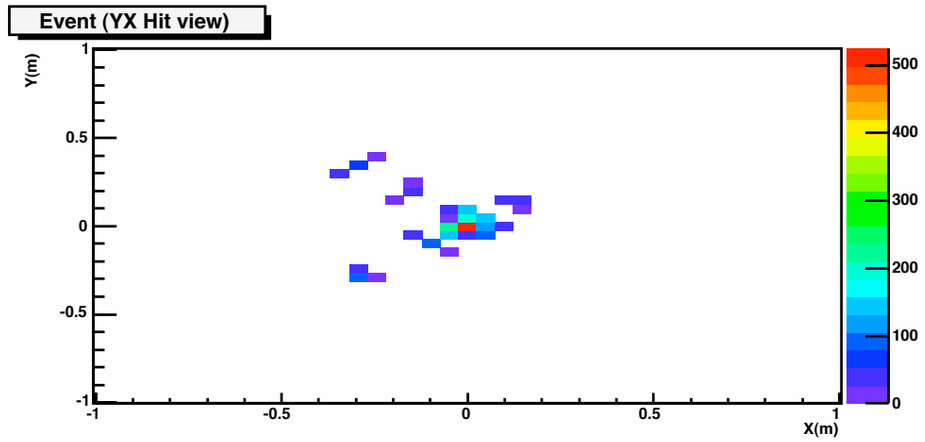

Transverse energy (U and V views)
tenestu
Entries 11
Mean 0.2539
RMS 1.8
Trans Energy
U View
V view
Strip number

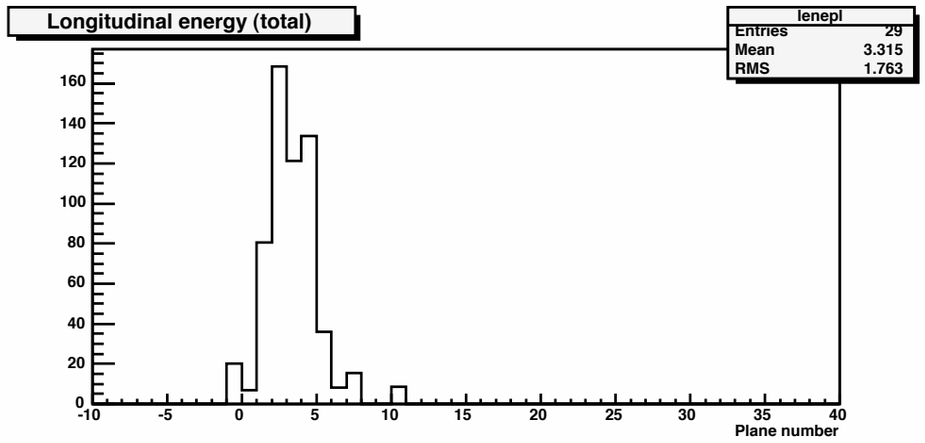

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

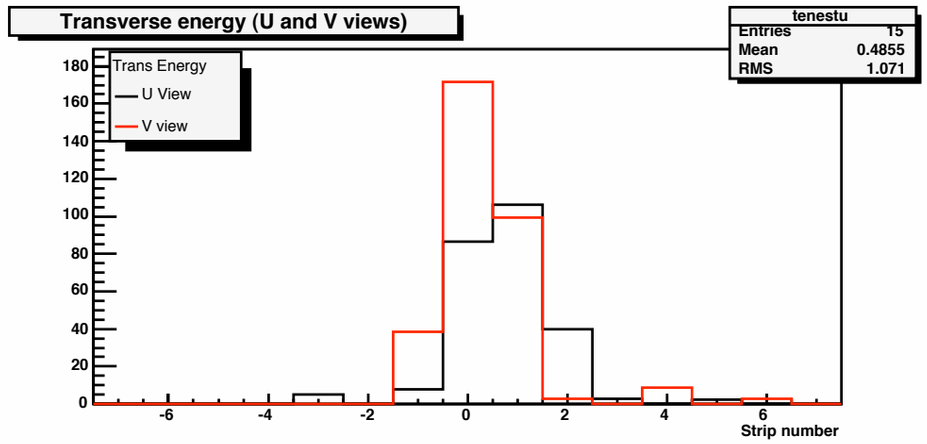

Event (U view)
Strip number
Plane number

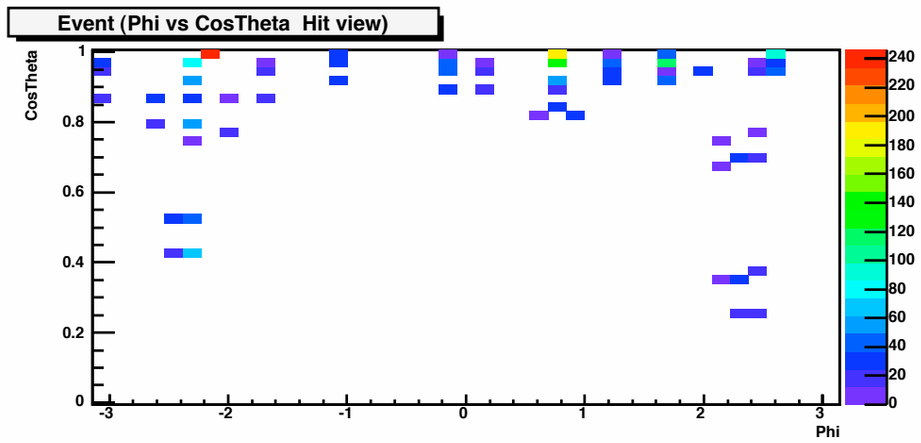

Event (YZ Hit view)
Y(m)
Z(m)


Event (V view)
Strip number
Plane number


Event (YX Hit view)
Y(m)
X(m)


Longitudinal energy (total)
lenepl
Entries 29
Mean 3.315
RMS 1.763
Plane number


Transverse energy (U and V views)
tenestu
Entries 15
Mean 0.4855
RMS 1.071
Trans Energy
U View
V view
Strip number


Event (Phi vs CosTheta Hit view)
CosTheta
Phi



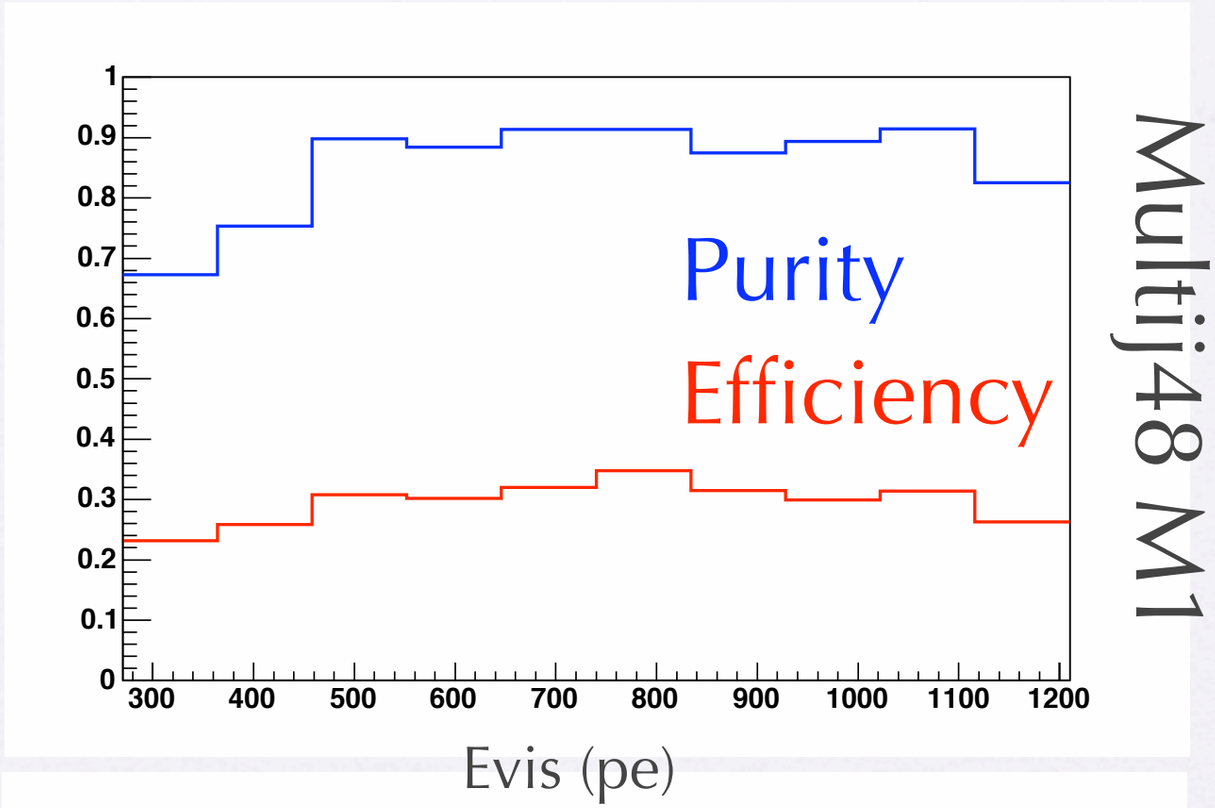
Ongoing work

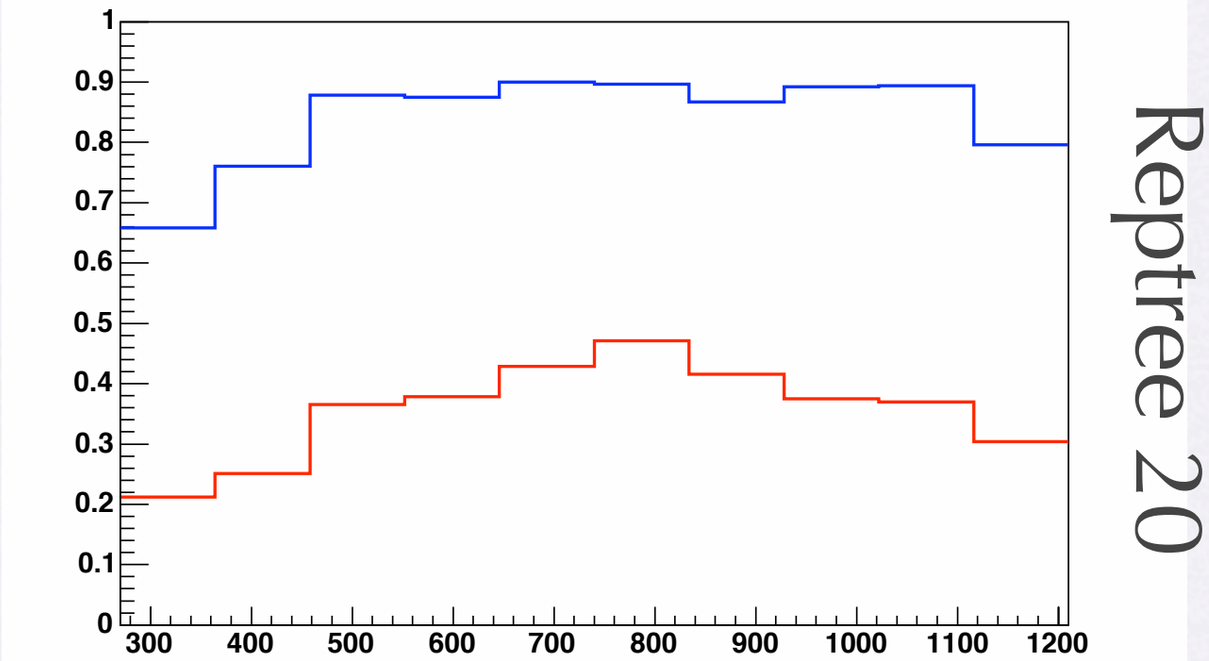
- Evaluation of CandShowerEM with respect to SR. So far the variables look promising but we adding them without detailed variable selection worsen results.
- Optimization of tree method. Boosting might yield further improvement.
- Do parameter estimation.

Efficiency and Purity

Before oscillations

- Low efficiency but high purity.
- Reptree manages a slight gain in the efficiency.

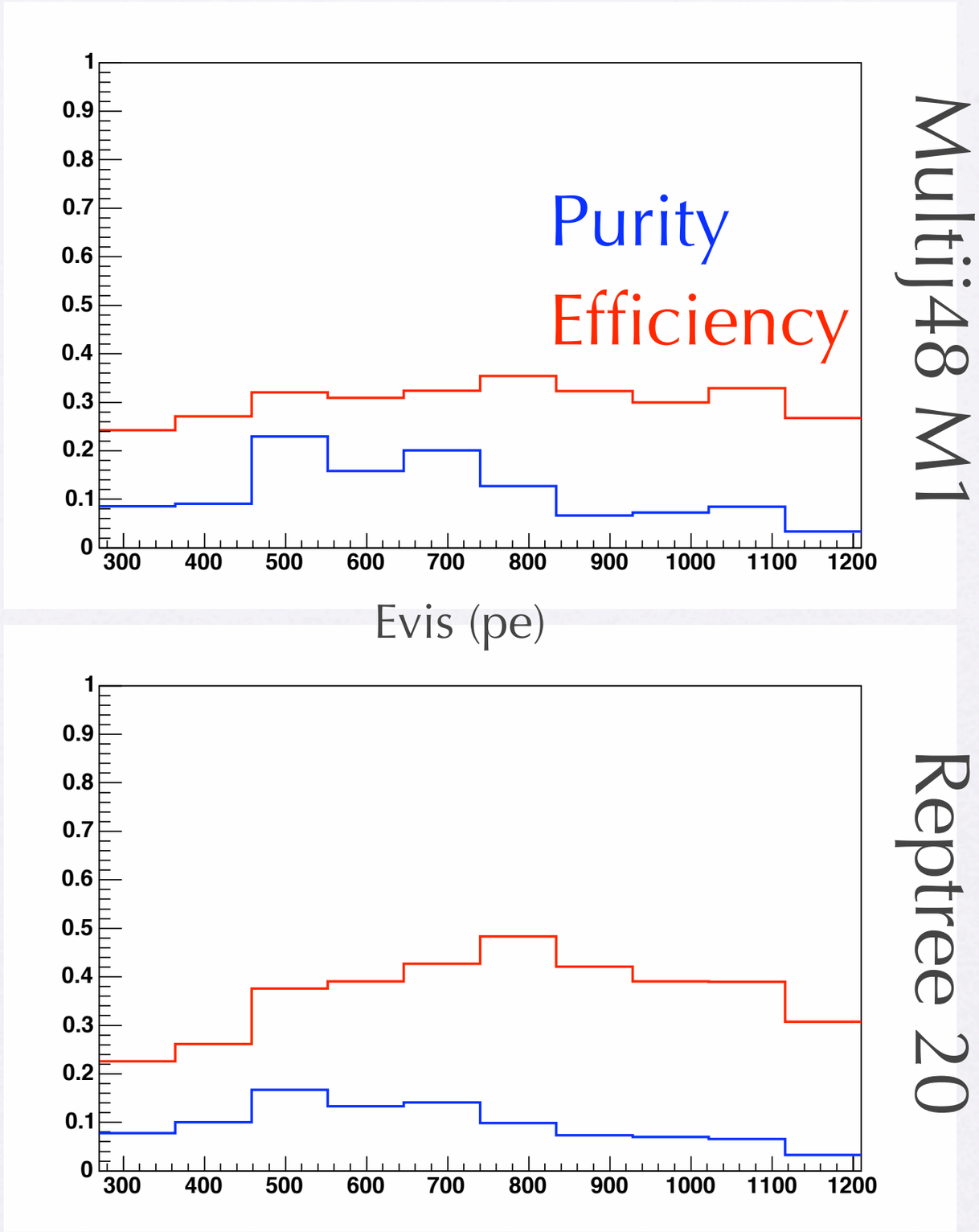




Efficiency and Purity

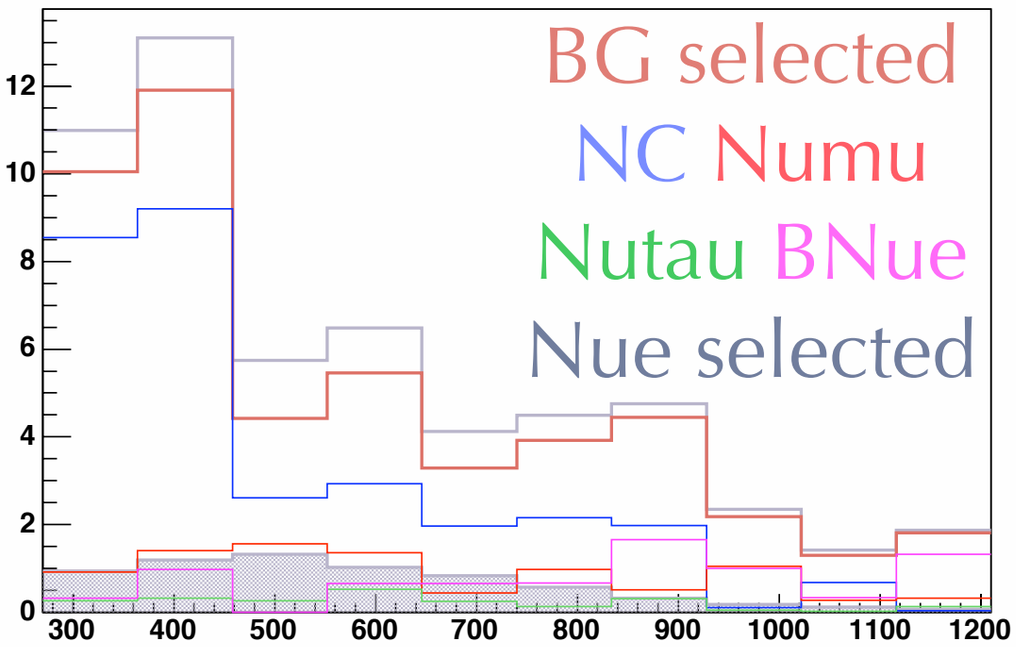
After oscillations

- Low efficiency but high purity.
- Reptree manages a slight gain in the efficiency.

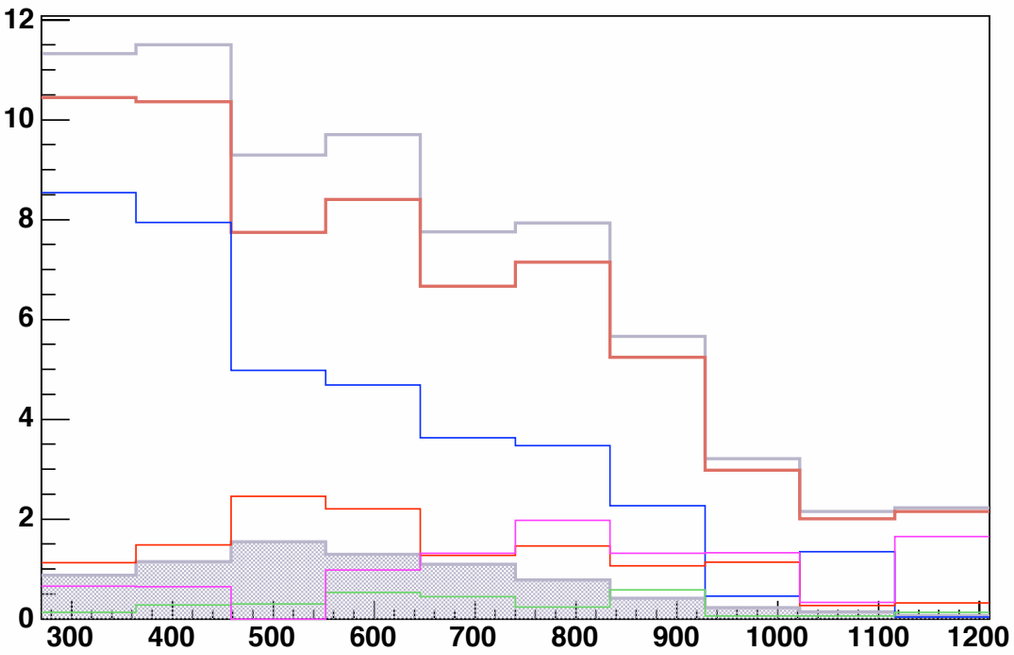


Separate BG behavior

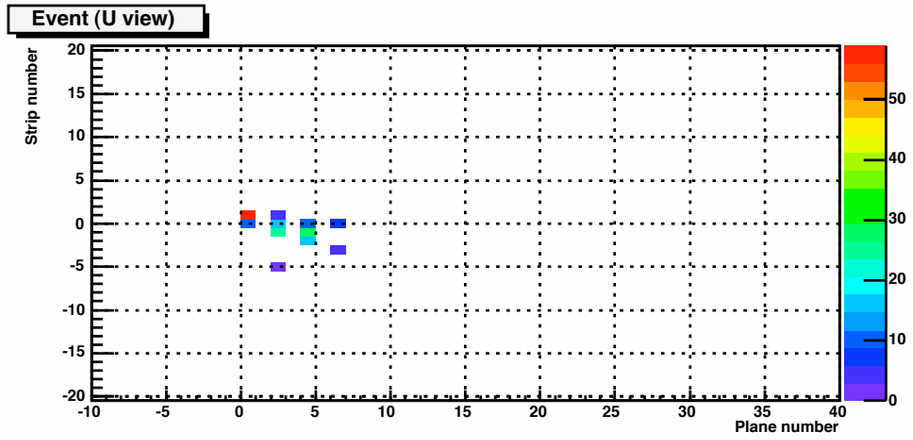
Multij48 M1

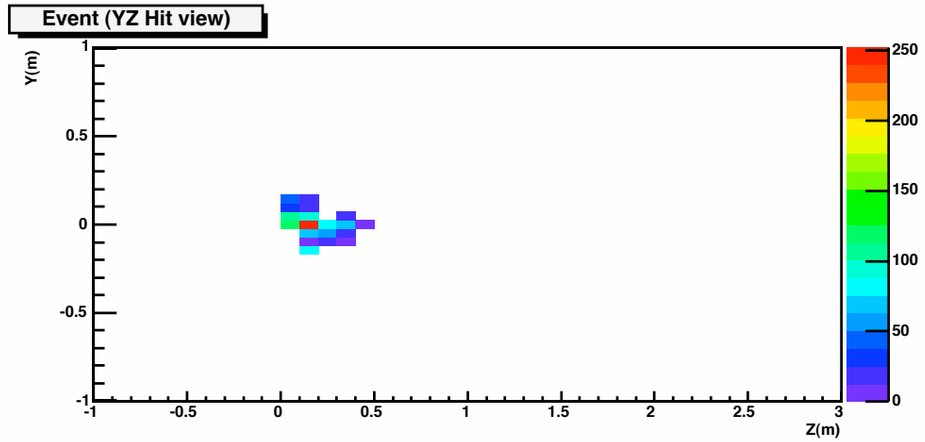


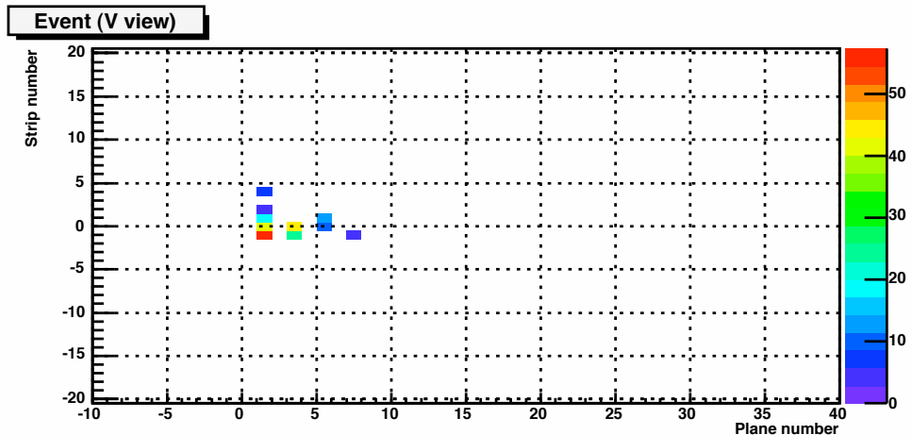
Reptree 20

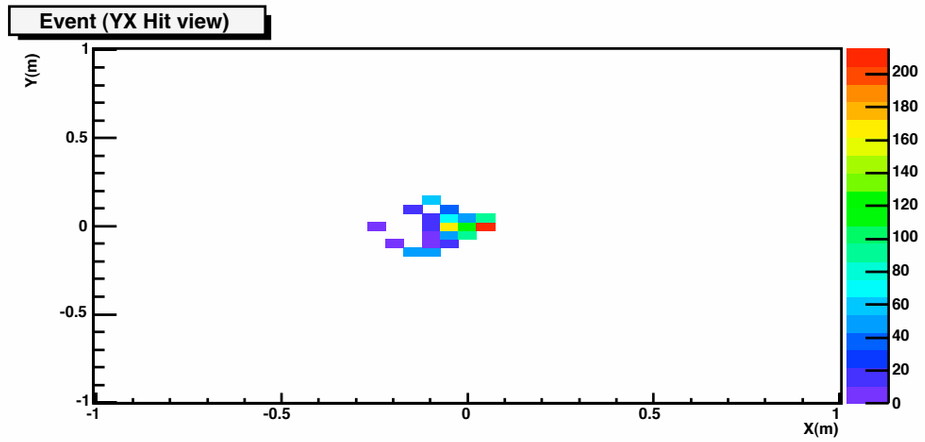


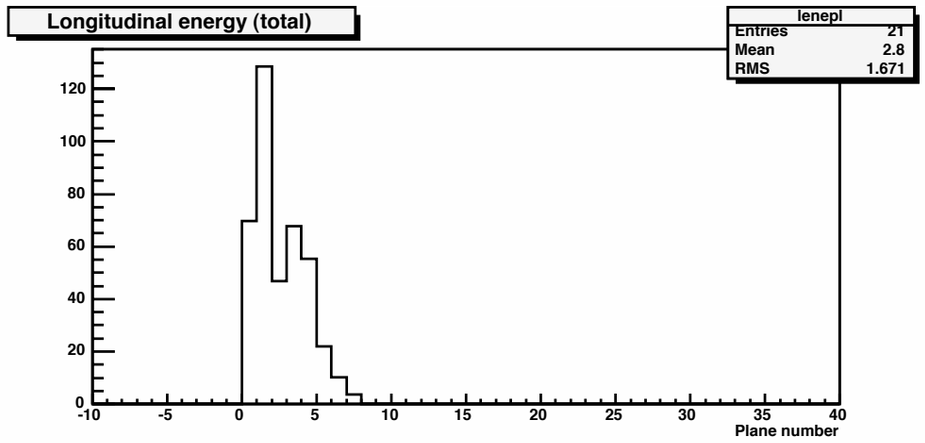
- Multij48 does a stronger cut in BG above 450 pe.
- Reptree does have a slight gain in signal, at the expense of more background.

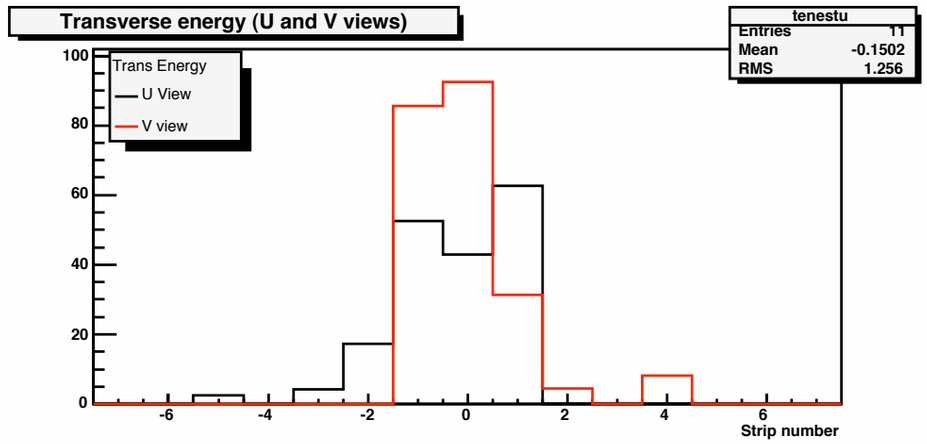

Event (U view)
Strip number
Plane number

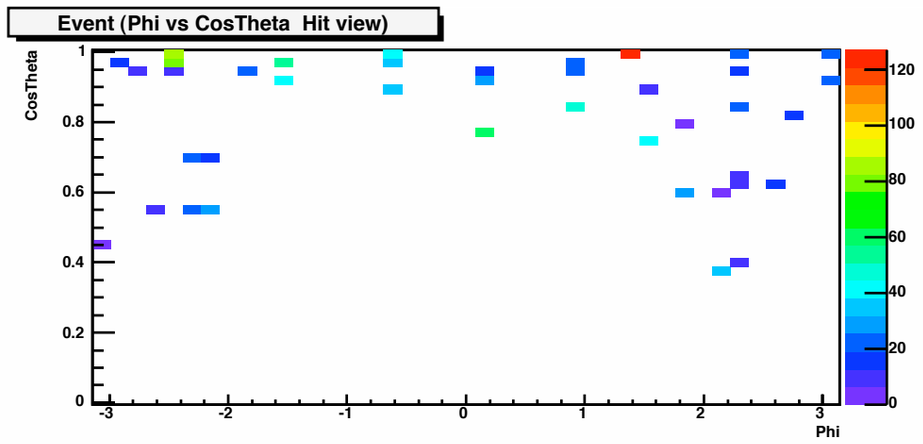

Event (YZ Hit view)
Y(m)
Z(m)


Event (V view)
Strip number
Plane number

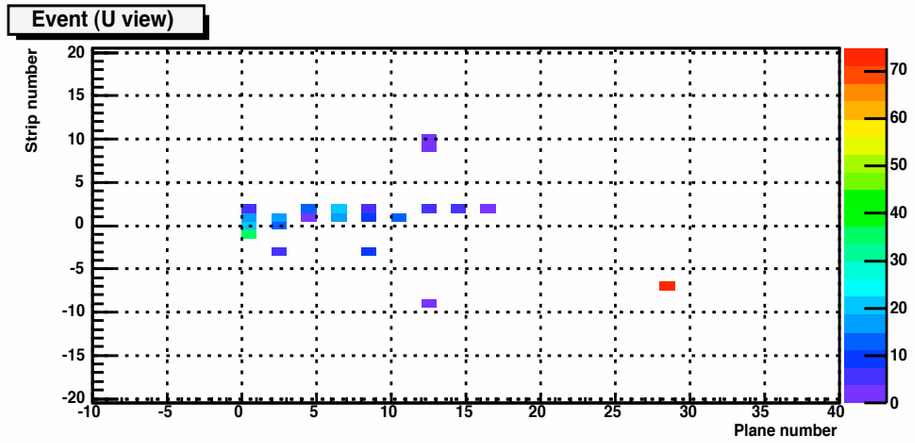

Event (YX Hit view)
Y(m)
X(m)

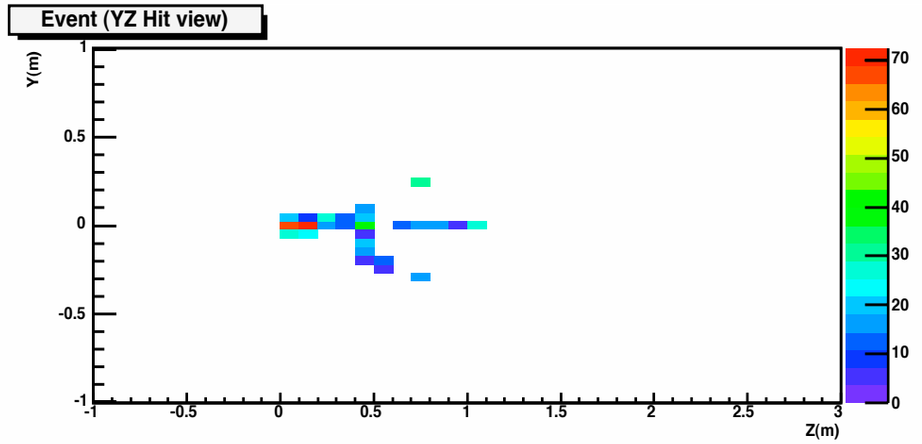

Longitudinal energy (total)
lenepl
Entries 21
Mean 2.8
RMS 1.671
Plane number

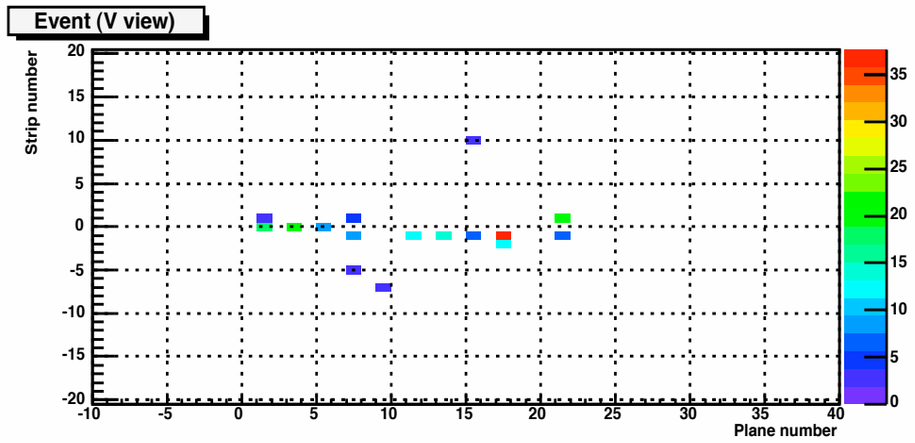

Transverse energy (U and V views)
tenestu
Entries 11
Mean -0.1502
RMS 1.256
Trans Energy
U View
V view
Strip number

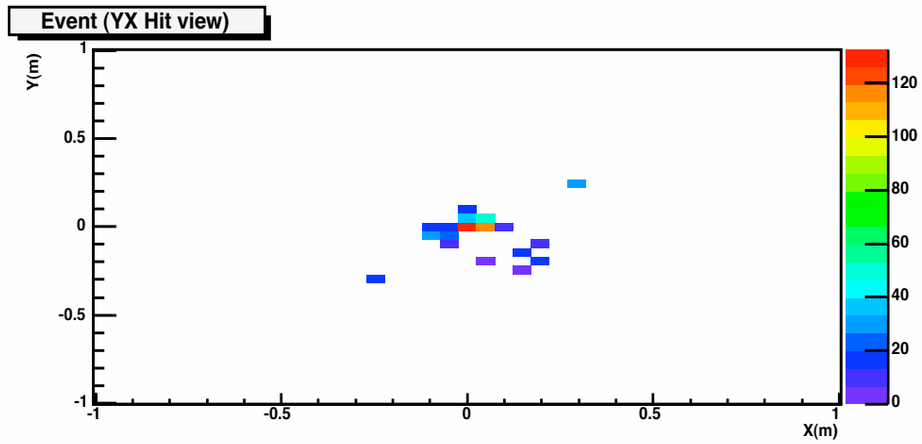

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

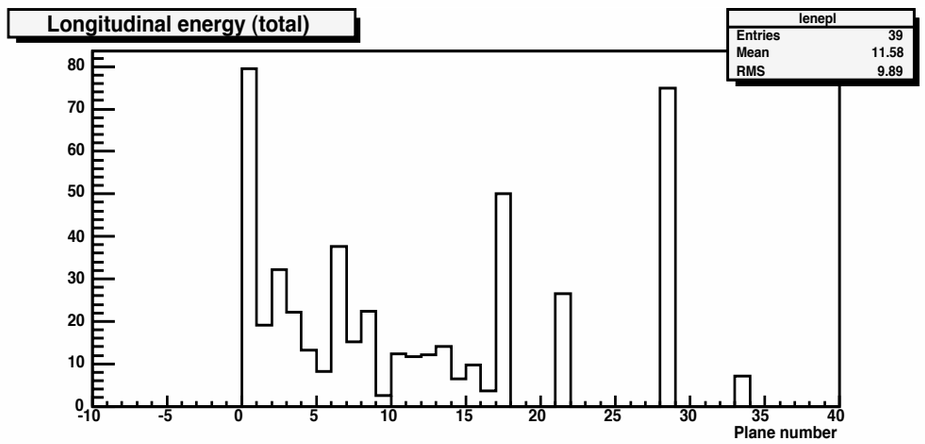


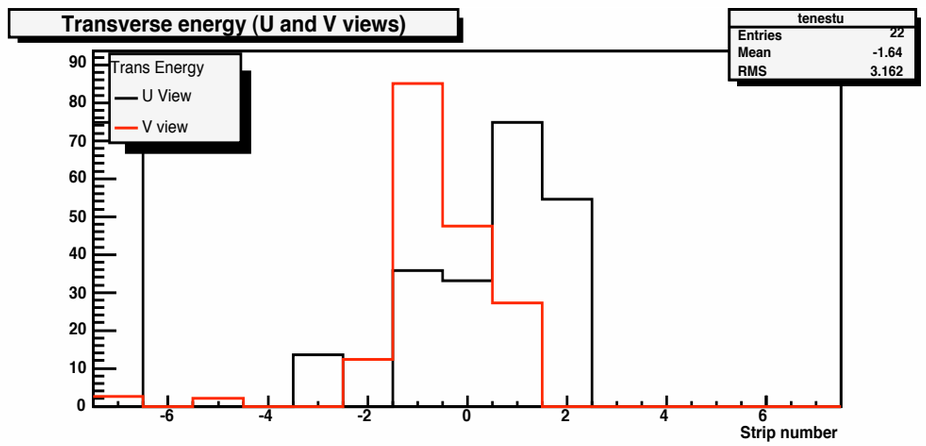

Event (U view)
Strip number
Plane number

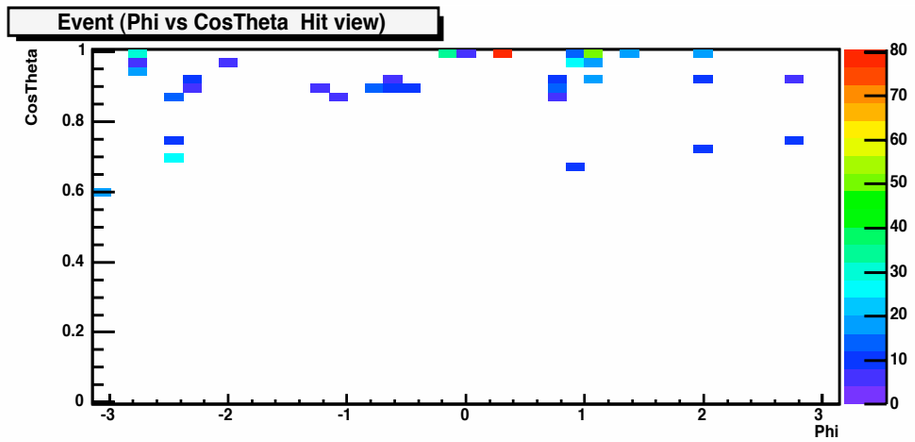

Event (YZ Hit view)
Y(m)
Z(m)


Event (V view)
Strip number
Plane number

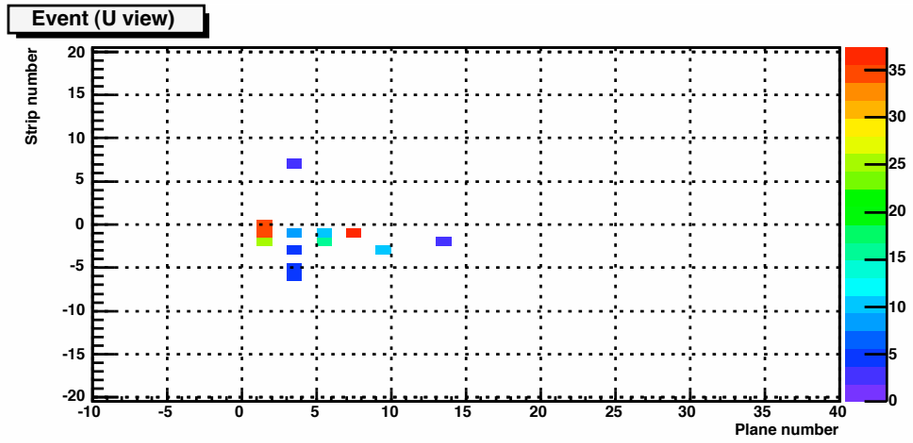

Event (YX Hit view)
Y(m)
X(m)

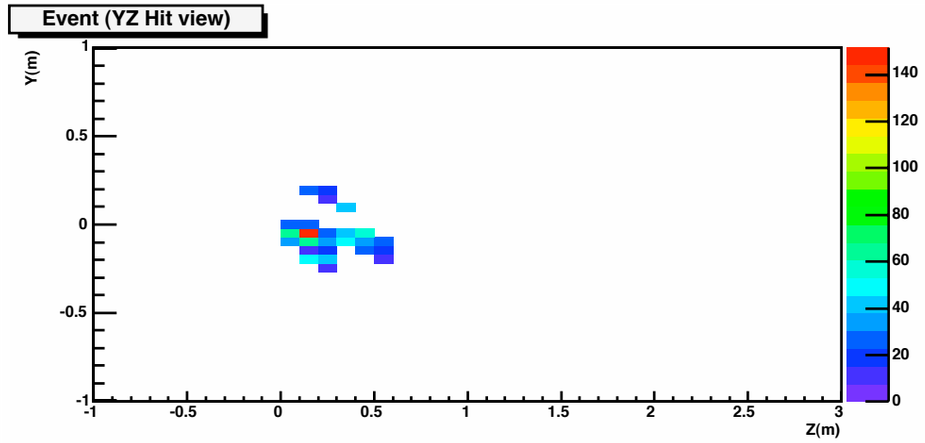

Longitudinal energy (total)
lenepl
Entries 39
Mean 11.58
RMS 9.89
Plane number

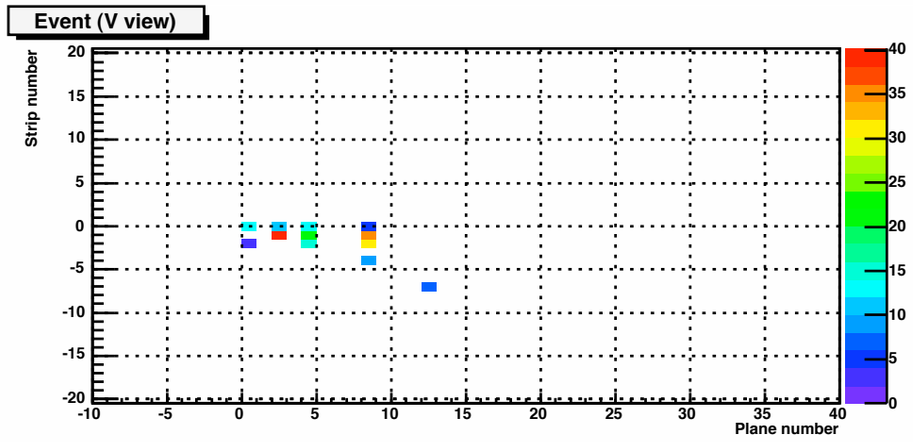

Transverse energy (U and V views)
tenestu
Entries 22
Mean -1.64
RMS 3.162
Trans Energy
U View
V view
Strip number

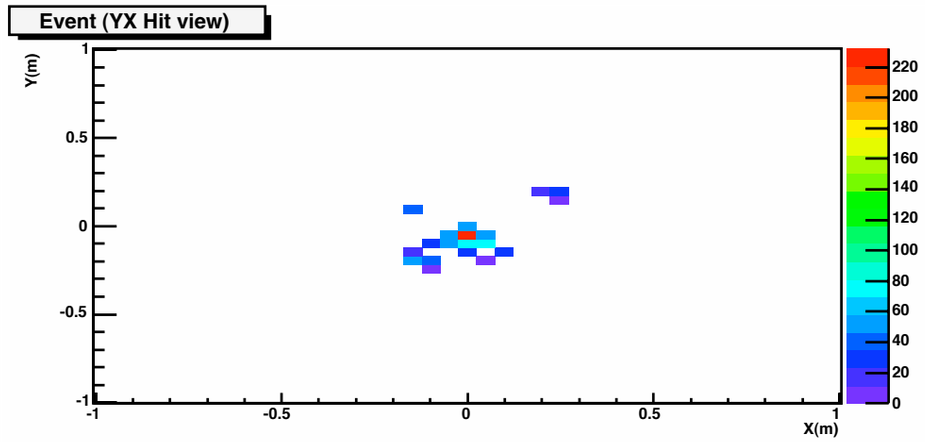

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

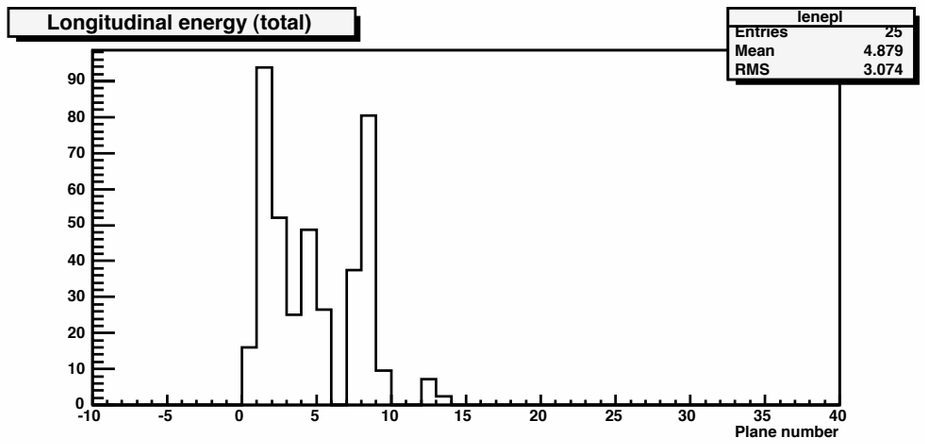


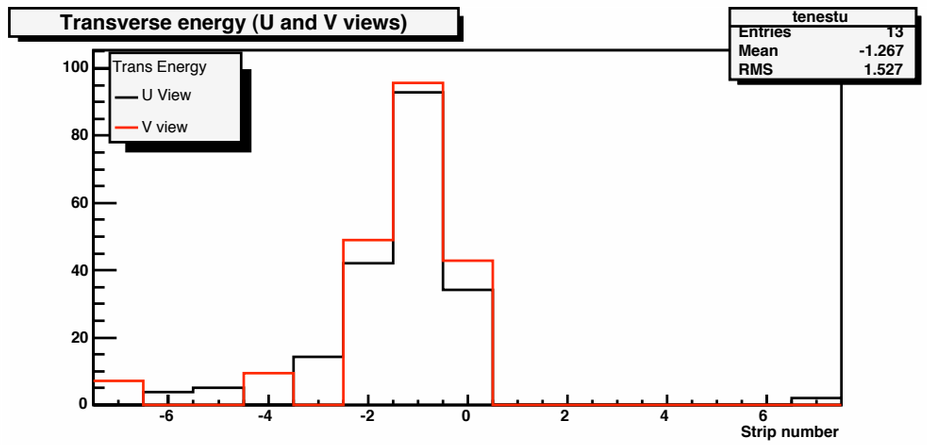

Event (U view)
Strip number
Plane number

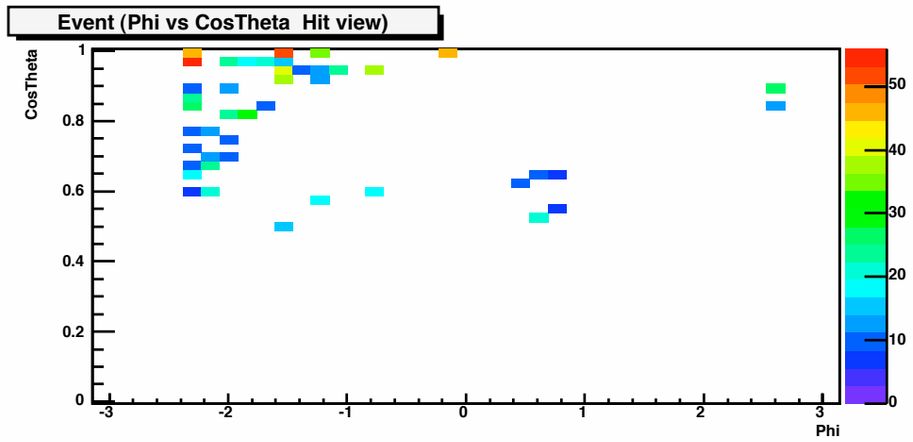

Event (YZ Hit view)
Y(m)
Z(m)

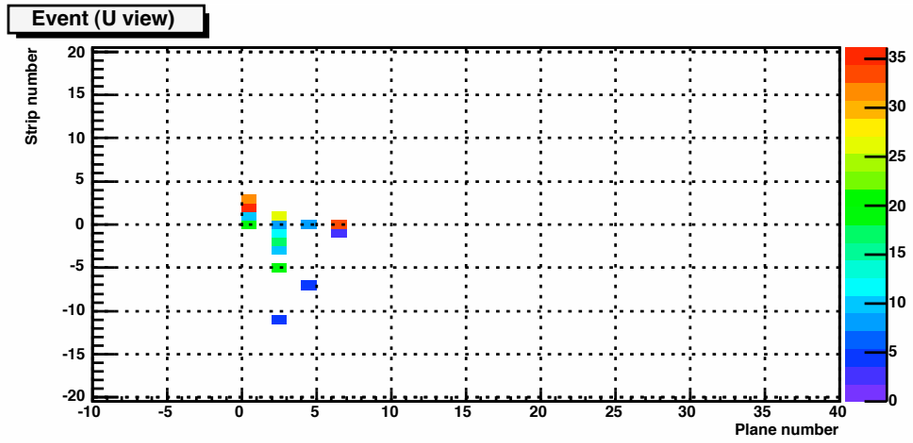

Event (V view)
Strip number
Plane number

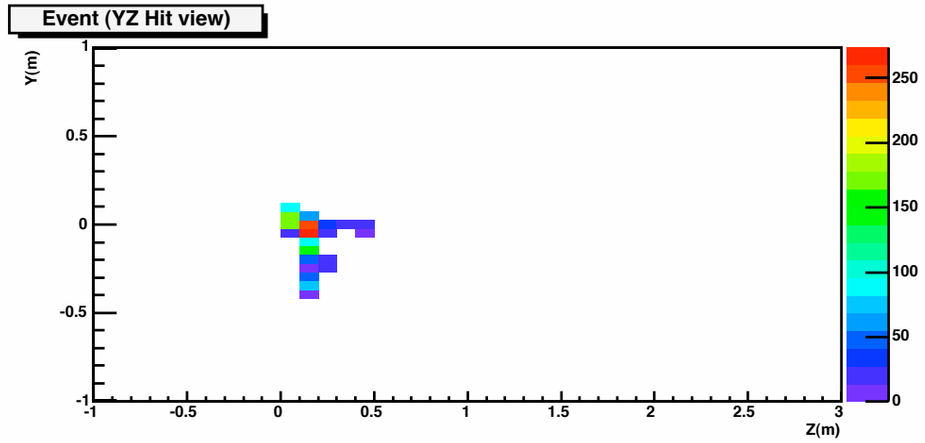

Event (YX Hit view)
Y(m)
X(m)

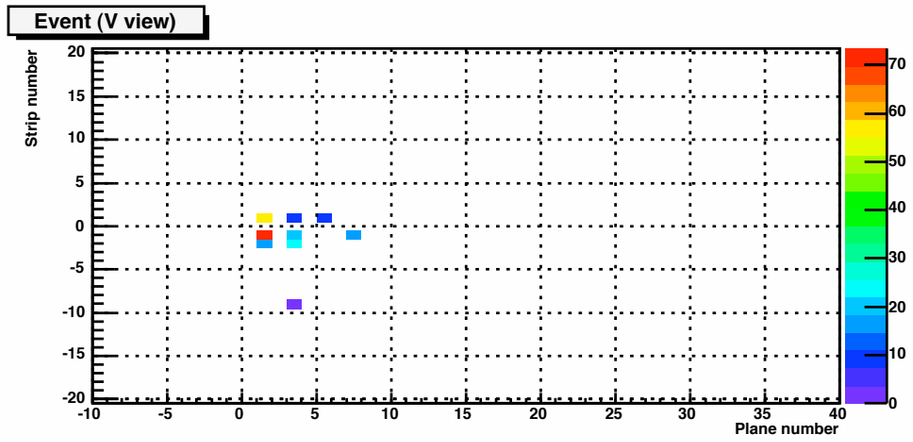

Longitudinal energy (total)
lenepl
Entries 25
Mean 4.879
RMS 3.074
Plane number

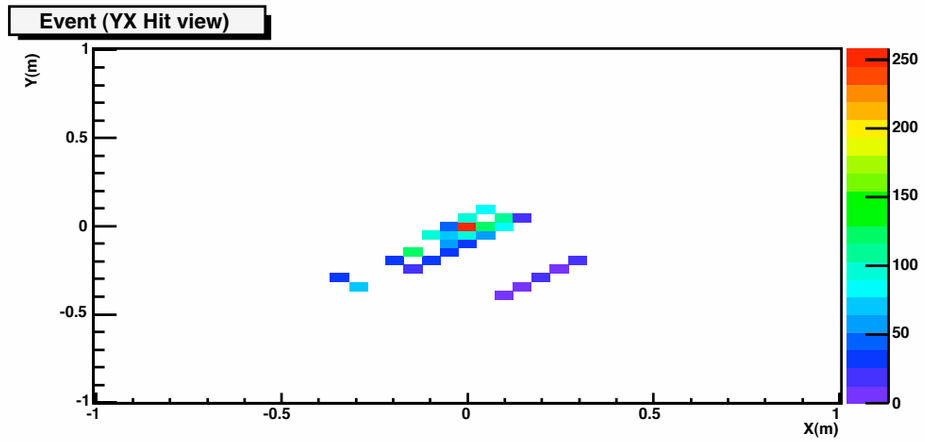

Transverse energy (U and V views)
tenestu
Entries 13
Mean -1.267
RMS 1.527
Trans Energy
U View
V view
Strip number

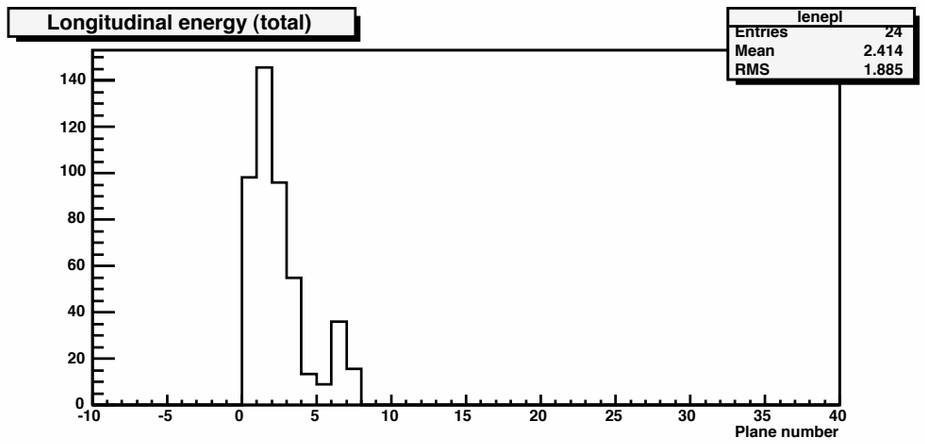

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

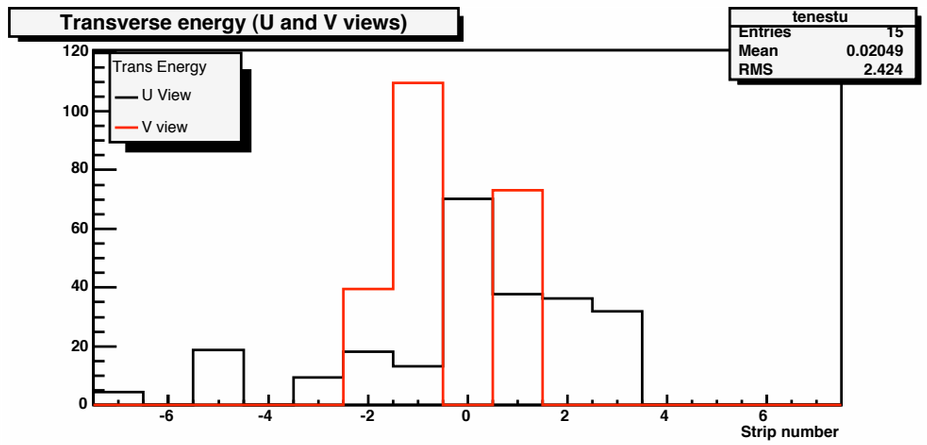

Event (U view)
Strip number
Plane number

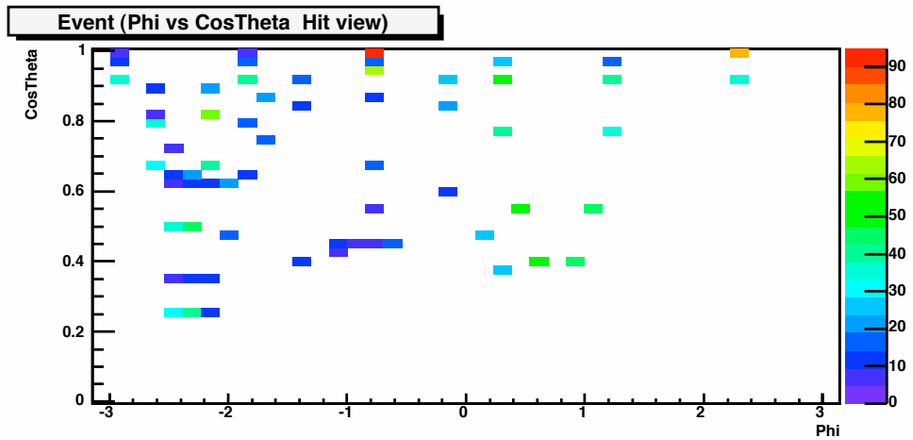

Event (YZ Hit view)
Y(m)
Z(m)

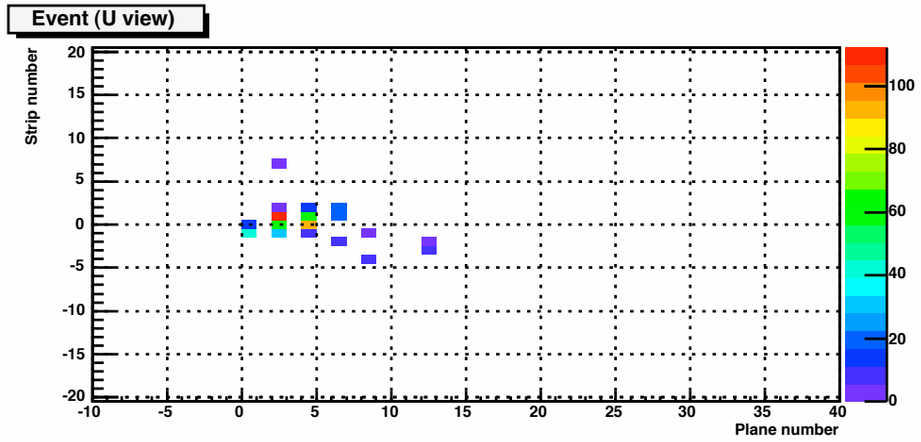

Event (V view)
Strip number
Plane number

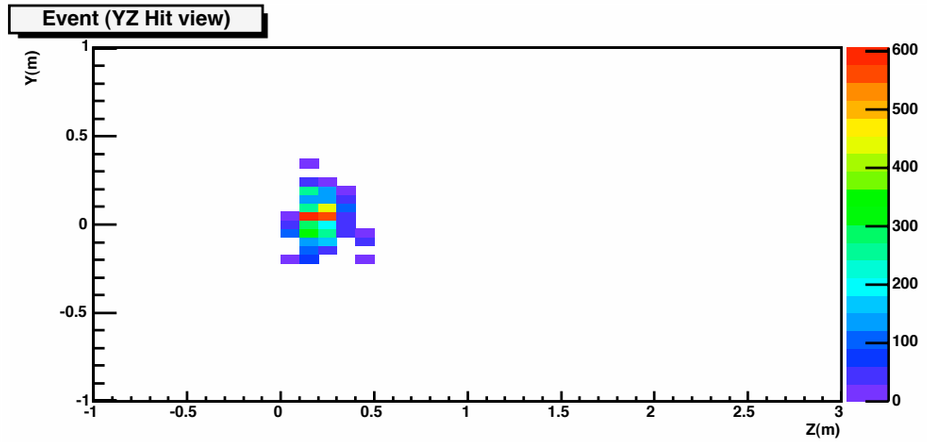

Event (YX Hit view)
Y(m)
X(m)

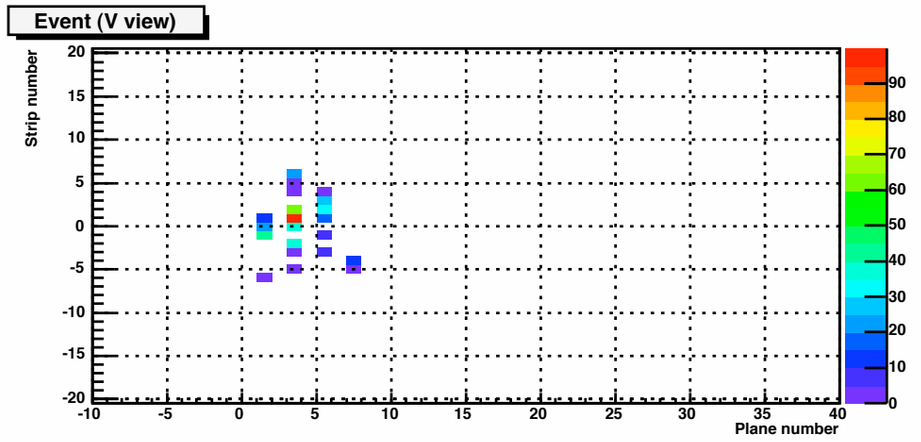

Longitudinal energy (total)
lenepl
Entries 24
Mean 2.414
RMS 1.885
Plane number

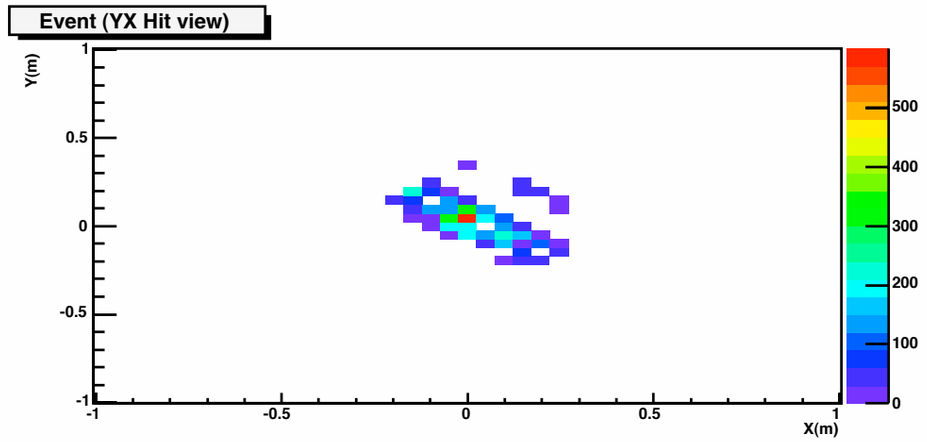

Transverse energy (U and V views)
tenestu
Entries 15
Mean 0.02049
RMS 2.424
Trans Energy
U View
V view
Strip number

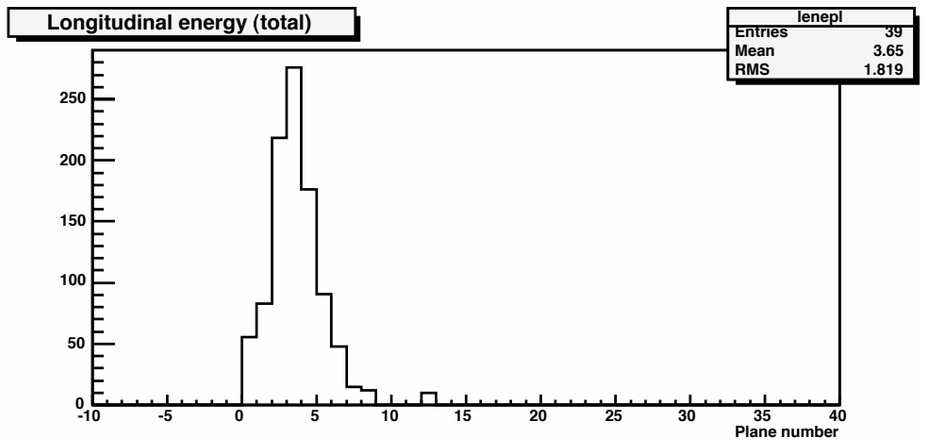

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

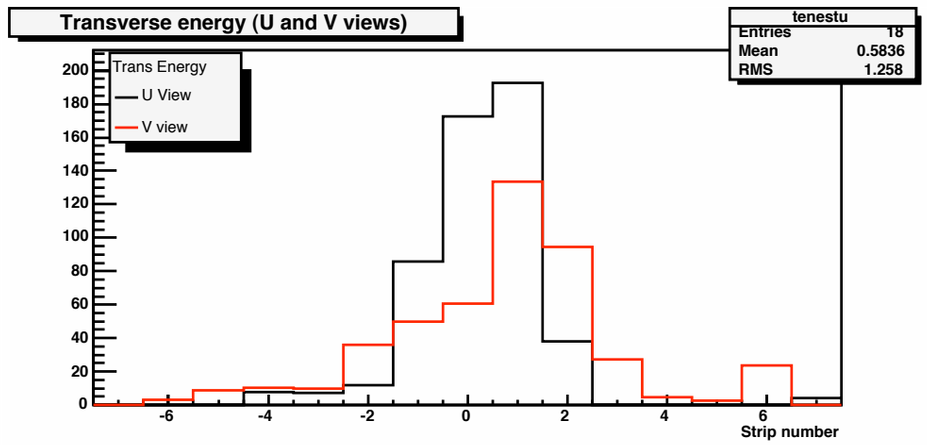

Event (U view)
Strip number
Plane number

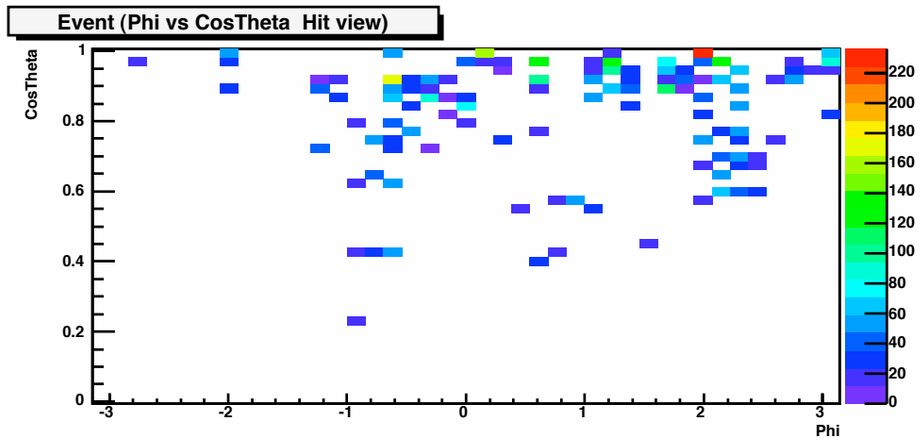

Event (YZ Hit view)
Y(m)
Z(m)

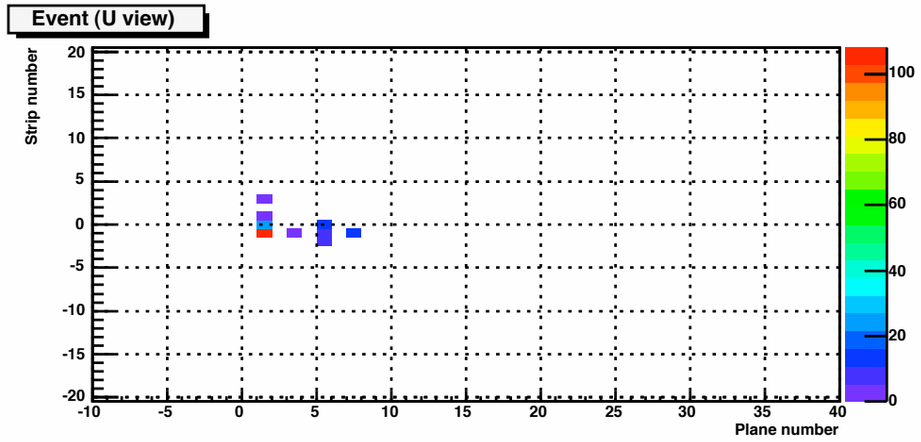

Event (V view)
Strip number
Plane number

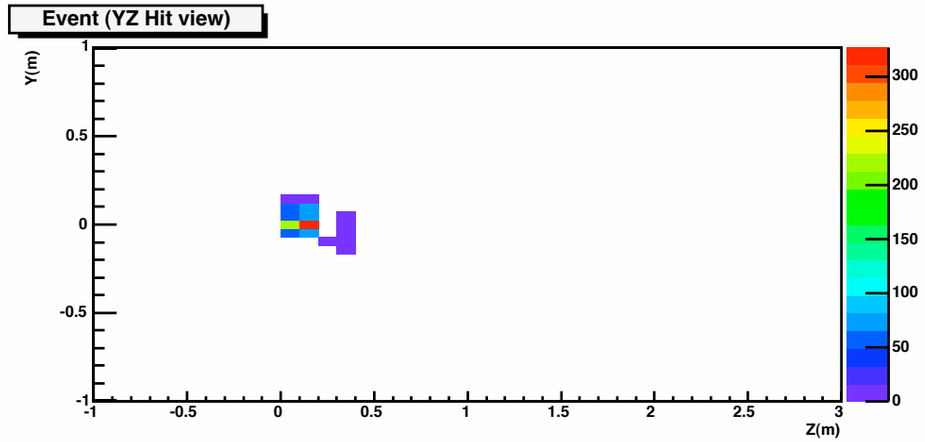

Event (YX Hit view)
Y(m)
X(m)

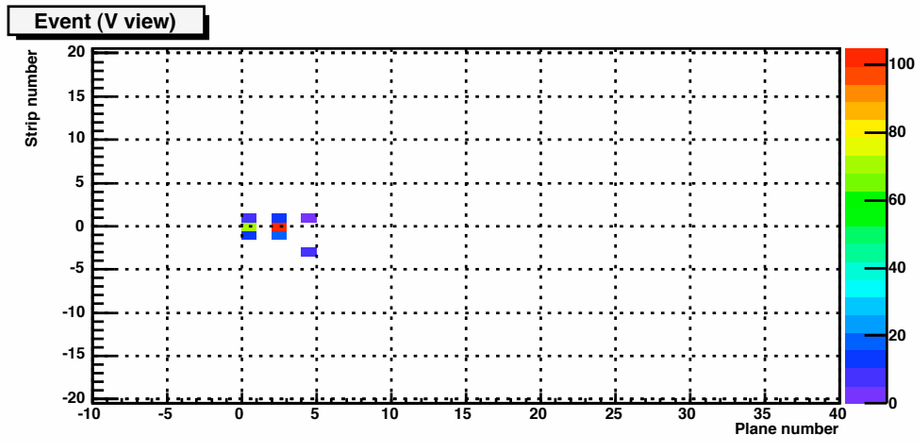

Longitudinal energy (total)
lenepl
Entries 39
Mean 3.65
RMS 1.819
Plane number

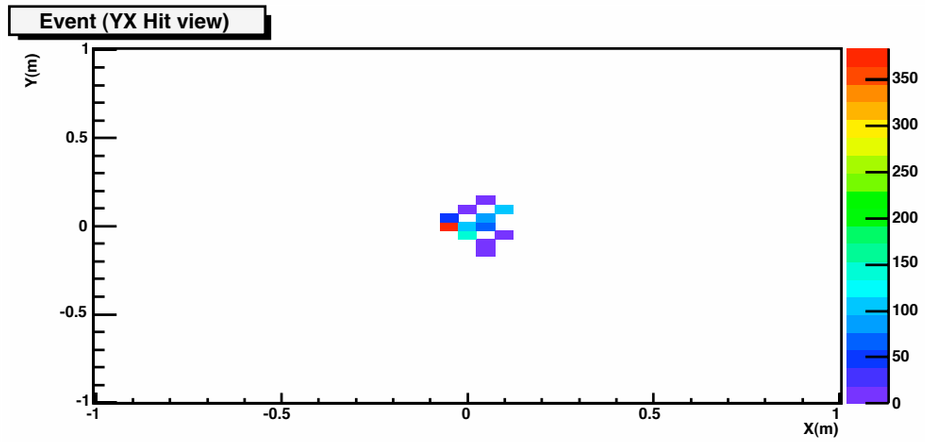

Transverse energy (U and V views)
tenestu
Entries 18
Mean 0.5836
RMS 1.258
Trans Energy
U View
V view
Strip number

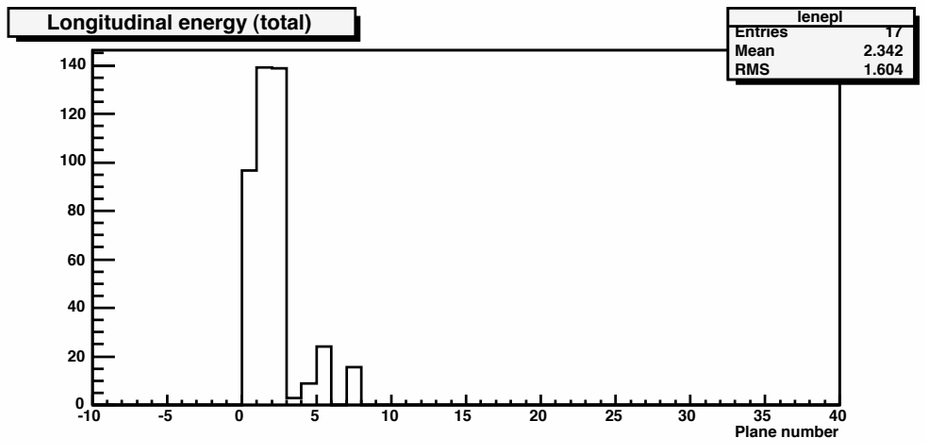

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

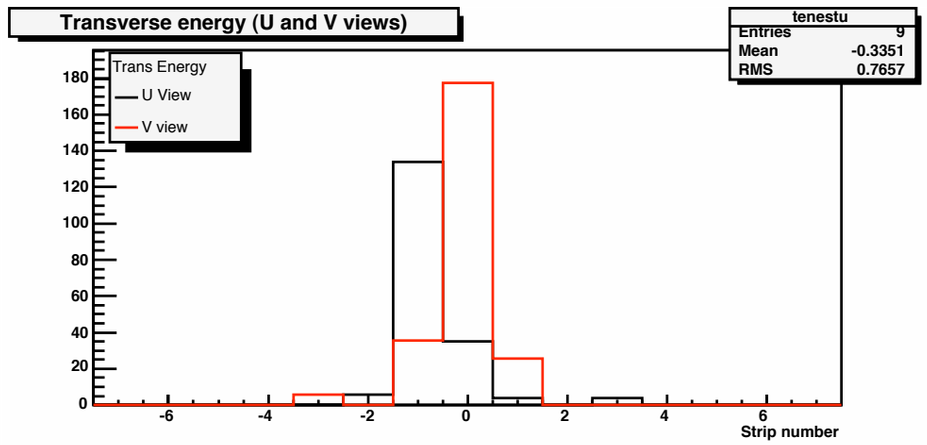

Event (U view)
Strip number
Plane number

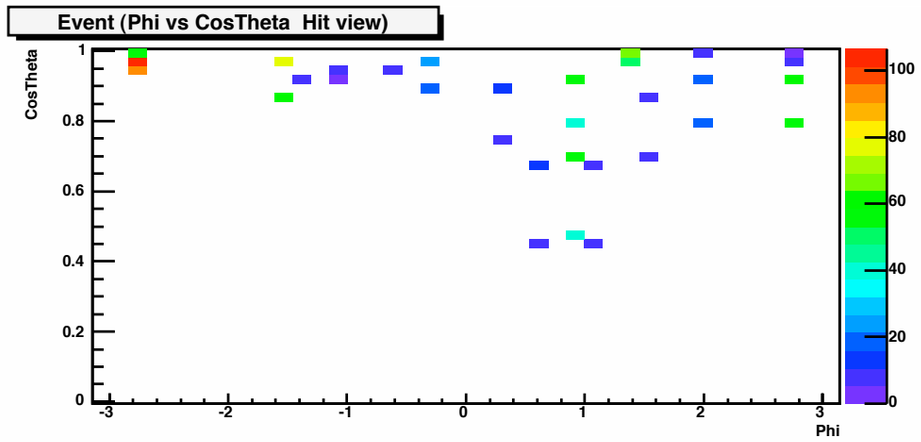

Event (YZ Hit view)
Y(m)
Z(m)

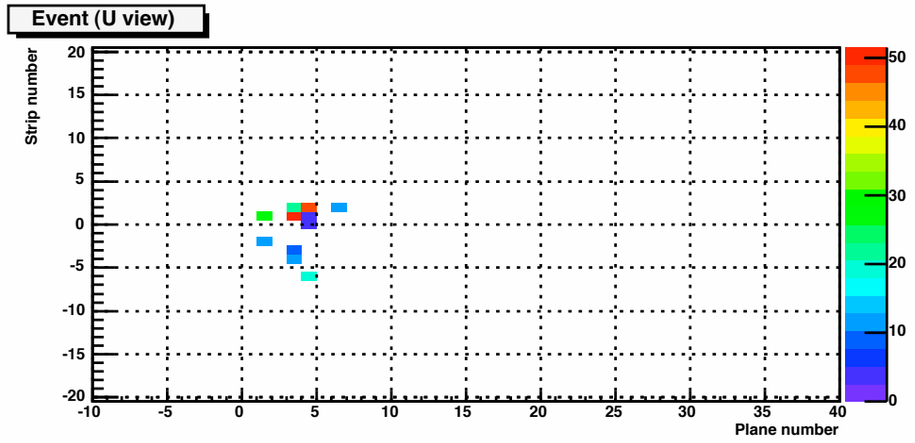

Event (V view)
Strip number
Plane number

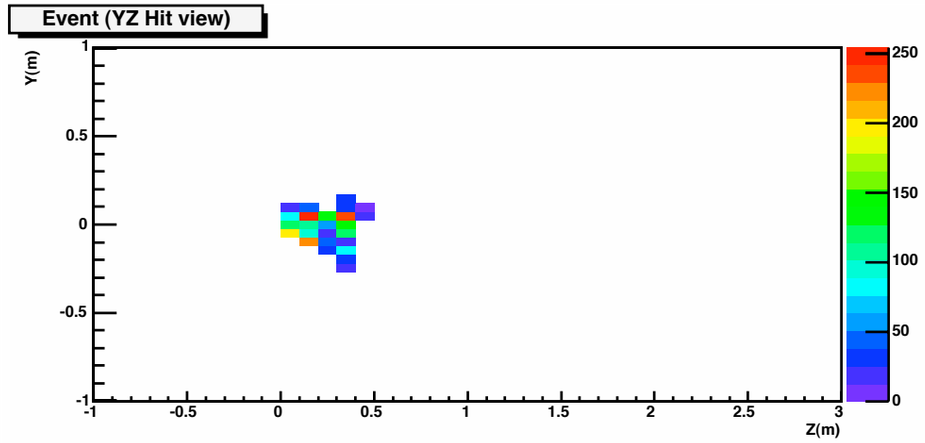

Event (YX Hit view)
Y(m)
X(m)

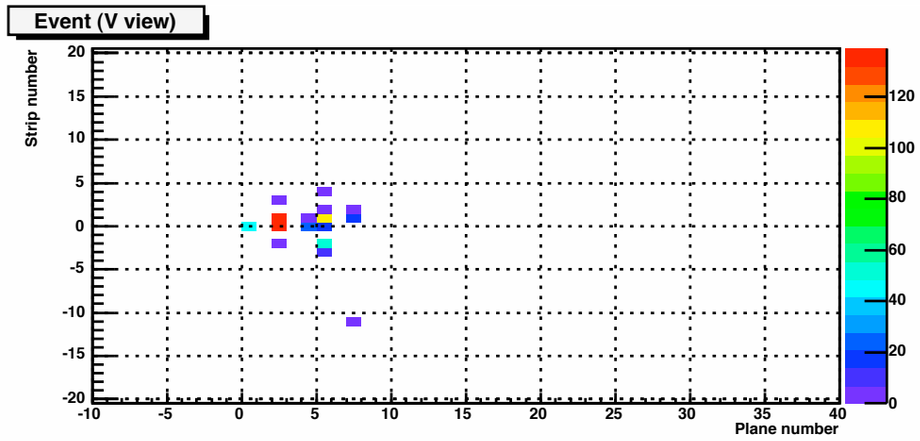

Longitudinal energy (total)
lenepl
Entries 17
Mean 2.342
RMS 1.604
Plane number

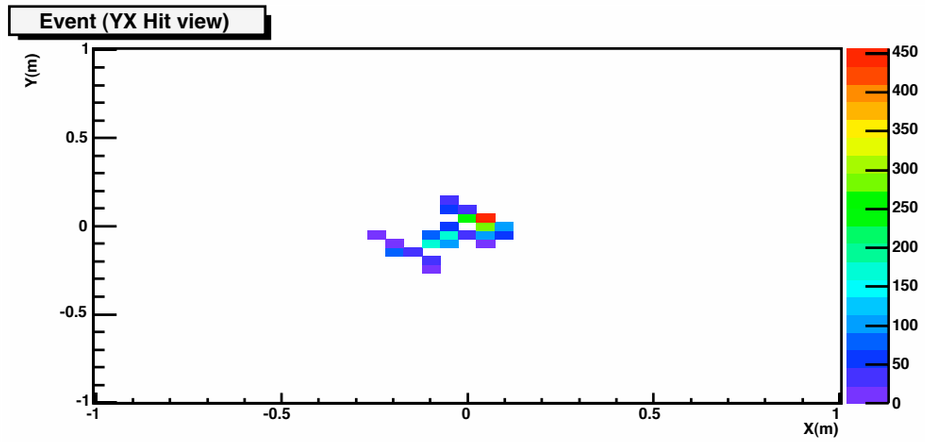

Transverse energy (U and V views)
tenestu
Entries 9
Mean -0.3351
RMS 0.7657
Trans Energy
U View
V view
Strip number

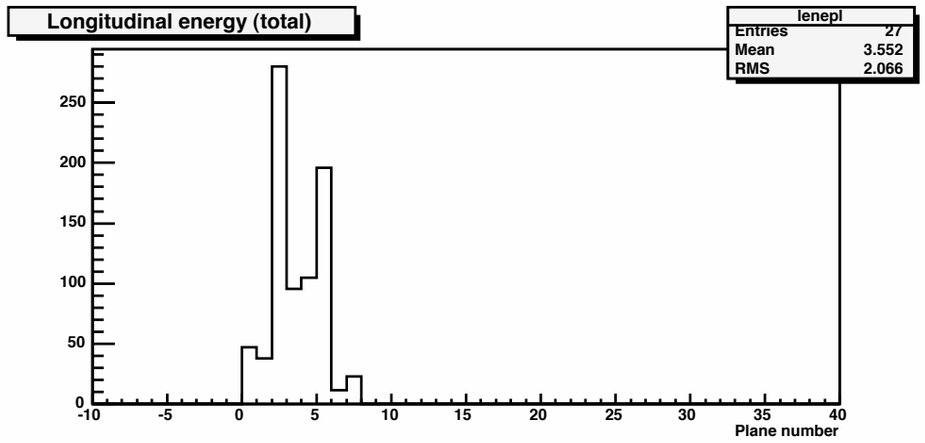

Event (Phi vs CosTheta Hit view)
CosTheta
Phi

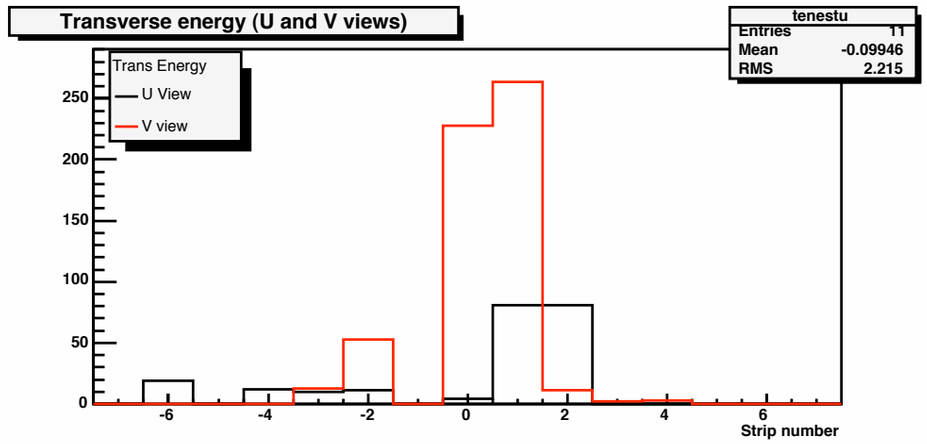

Event (U view)
Strip number
Plane number

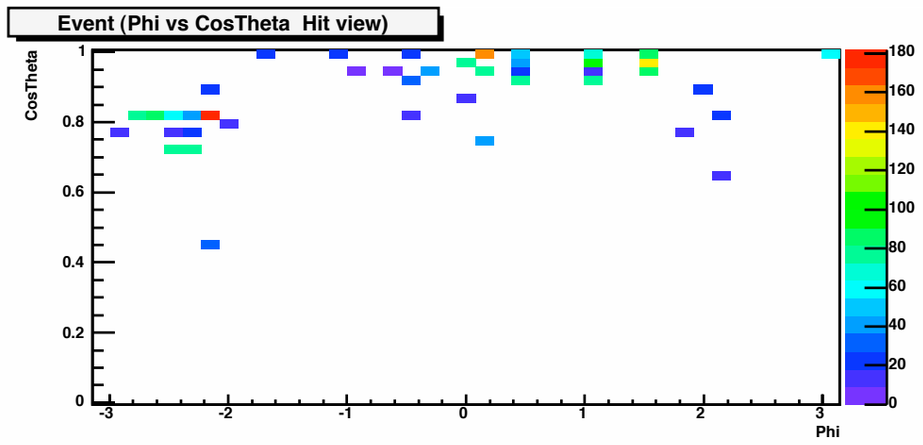

Event (YZ Hit view)
Y(m)
Z(m)


Event (V view)
Strip number
Plane number


Event (YX Hit view)
Y(m)
X(m)


Longitudinal energy (total)
lenepl
Entries 27
Mean 3.552
RMS 2.066
Plane number


Transverse energy (U and V views)
tenestu
Entries 11
Mean -0.09946
RMS 2.215
Trans Energy
U View
V view
Strip number


Event (Phi vs CosTheta Hit view)
CosTheta
Phi

