Validation of nanoAOD variables in the Higgs to four leptons case

Paula Martínez
DESY Summer Student Programme 2019

Overview

Introduction

A brief introduction to data formats

Description of the project

Validation of variables
- Variable correspondence
- Selection cuts
- Data 2011
- Data 2012

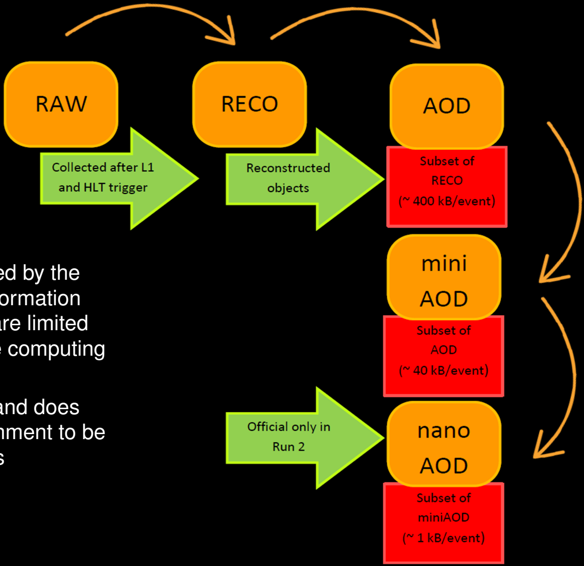
Higgs to four leptons plot

Conclusions

Introduction

- CERN Open Data Portal (CODP) is a platform which allows public access to the data produced by the different research activities at CERN. Most of the information contained in the CODP includes simplified data formats, reconstructed data and simulations, and the necessary analysis software.
- The CODP contains a research level example for the Higgs to four leptons process using 2011 and 2012 data. A histogram of the invariant mass of the Higgs boson reconstructed from the four leptons plus the contributions of the different backgrounds is produced from AOD data.
- The aim of this project is to produce the same results from nanoAOD-like (nanoAODplus) format data files.

A brief introduction to data formats



- The amount of data collected by the detectors, as well as the information contained in the datasets, are limited by (among other things) the computing power.
- nanoAOD format is lighter and does not need a CMSSW environment to be run, but it also contains less information.

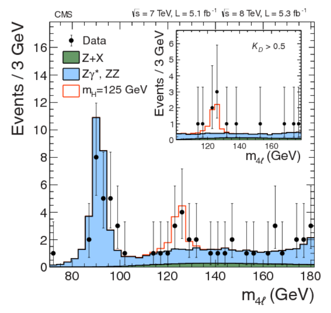
Description of the project

- This particular project consists on validating the variables from a nanoAODplus ntuple filled with Run 1 data. It is called nanoAODplus because it contains extra information with respect to the official nanoAOD (from Run 2).

- To do so, we compare the histograms produced by the Higgs example from the CODP and the histograms produced by the nanoAODplus (under the same conditions).

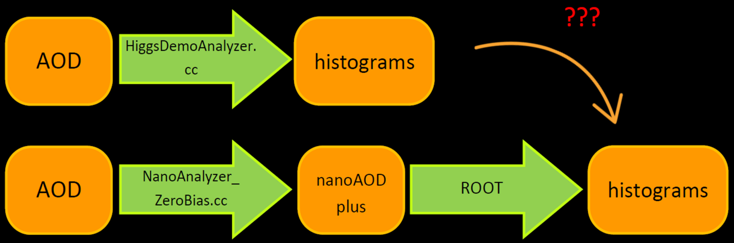
- If a variable is a match, we can say that it has been validated.

- In the end, we will use these new datasets to produce the histogram of the invariant mass of $H \to 4\ell$, and compare it with the official one (https://arxiv.org/pdf/1207.7235.pdf)



Validation of variables

Variable correspondence



- The variables in the Higgs example (HiggsDemoAnalyzer.cc) are not necessarily the same as the ones in the nanoAODplus (NanoAnalyzer_ZeroBias.cc).
- To reproduce the cuts, we first need to identify all the variables.
- As we will see later, this variables do not always match, even if referring to the same feature.

First step: match the variables that we will use in the cuts afterwards.

Validation of variables

Variable correspondence

MUONS

Variable	HiggsDemoAnalyzer.cc	NanoAnalyzer_ZeroBias.cc
PF candidate	itMuon.isPFMuon()	Muon_isPFcand
PF isolation valid*	itMuon.isPFIsolationValid()	
Global muon	(itMuon.globalTrack()).IsNonull()	Muon_isGlobal
pT	itMuon.pt()	Muon_pt
eta	itMuon.eta()	Muon_eta
Impact parameter significance	SIP3d_mu	Muon_sip3d
Distance in xy to the vertex	itMuon.globalTrack())->dxy(point)	Muon_dxy
Distance in z to the vertex	itMuon.globalTrack())->dz(point)	Muon_dz
Relative isolation	relPFIso_mu	Muon_pfRelIso04_all

Table: Muon variables.

* We will see that not having this variable defined in the nanoAODplus is not relevant.

point = position of the beamspot

Validation of variables

Variable correspondence

SIP3d_mu AND relPFIso_mu DEFINITIONS

relPFIso_mu =	((itMuon.pfIsolationR04()).sumChargedHadronPt + (itMuon.pfIsolationR04()).sumNeutralHadronEt + (itMuon.pfIsolationR04()).sumPhotonEt) / itMuon.pt()
IP3d_mu =	sqrt((itMuon.globalTrack()->dxy(point) * itMuon.globalTrack()->dxy(point)) + (itMuon.globalTrack()->dz(point) * itMuon.globalTrack()->dz(point)))
ErrIP3d_mu =	sqrt((itMuon.globalTrack()->d0Error() * itMuon.globalTrack()->d0Error()) + (itMuon.globalTrack()->dzError() * itMuon.globalTrack()->dzError()))
SIP3d_mu =	IP3d_mu / ErrIP3d_mu

Validation of variables

Variable correspondence

ELECTRONS

Variable	HiggsDemoAnalyzer.cc	NanoAnalyzer_ZeroBias.cc
PF preselection	itElectron.passingPflowPreselection()	Electron_isPFcand
pT	itElectron.pt()	Electron_pt
Supercluster eta	(itElectron.superCluster())->eta())	Electron_deltaEtaSC + Electron_eta
Misshits	misshits	Electron_lostHits
Impact parameter significance	SIP3d_e	Electron_sip3d
Distance in xy to the vertex	itElectron.gsfTrack()->dxy(point)	Electron_dxy
Distance in z to the vertex	itElectron.gsfTrack()->dz(point)	Electron_dz
Within barrel acceptance	itElectron.isEB	Electron_isEB
Within endcap acceptance	itElectron.isEE	Electron_isEE
Relative isolation	relPFIso_e	Electron_pfRelIso03_all

Table: Electron variables.

Not defined in previous versions of NanoAnalyzer.

point = position of the beamspot

Validation of variables

Variable correspondence

misshits, SIP3d_e AND relPFIso_e DEFINITIONS

misshits =	((itElectron.gsfTrack())->trackerExpectedHitsInner()).numberOfHits()
relPFIso_e =	((itElectron.pfIsolationR04()).sumChargedHadronPt + (itElectron.pfIsolationR04()).sumNeutralHadronEt + (itElectron.pfIsolationR04()).sumPhotonEt) / itElectron.pt()
IP3d_e =	sqrt((itElectron.gfsTrack()->dxy(point) * itElectron.gfsTrack()->dxy(point)) + (itElectron.gfsTrack()->dz(point) * itElectron.gfsTrack()-¿dz(point)))
ErrIP3d_e =	sqrt((itElectron.gfsTrack()->d0Error() * itElectron.gfsTrack()->d0Error()) + (itElectron.gfsTrack()->dzError() * itElectron.gfsTrack()->dzError()))
SIP3d_e =	IP3d_e / ErrIP3d_e

Validation of variables

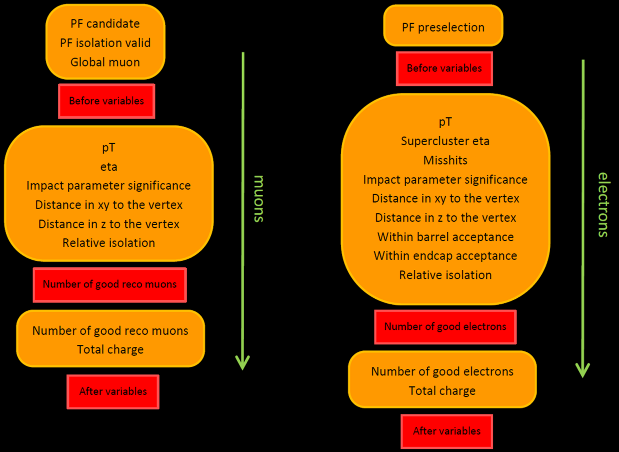
Variable correspondence

- After applying the cuts in this variables we can define the number of 'good muons', or 'good electrons' (muons and electrons that pass all the cuts).
- We can then use this new variables to select different types of event ($ZZ \to \mu\mu$, $ZZ \to \mu\mu ee$, $ZZ \to eeee$, ...), along with the charge of the particles.

Validation of variables

Selection cuts

The selection cuts are applied in three stages. Depending on the stage we will have 'before variables' and 'after variables'.



Validation of variables

Selection cuts

MUON SELECTION

Variable	Cut
PF candidate	true
PF isolation valid*	true
Global muon	true
pT	> 5
$\|$eta$\|$	< 2.4
$\|$Impact parameter significance$\|$	< 4
$\|$Distance in xy to the vertex$\|$	< 0.5
$\|$Distance in z to the vertex$\|$	< 1
Relative isolation	< 0.4

Table: Muon cuts.

* Only in Higgs example.

To select events of the type $ZZ \to 2\ell$ or $ZZ \to 4\ell$ we use Number of good reco muons = 2 or 4. The total charge must be 0. In the four lepton case, the total charge of each pair of leptons must also be 0.

Validation of variables

Selection cuts

ELECTRON SELECTION

Variable	**Cut**
PF preselection	true
pT*	> 7
|Supercluster eta|	< 2.5
Misshits	≤ 1
|Impact parameter significance|	< 4
|Distance in xy to the vertex|	< 0.5
|Distance in z to the vertex|	< 1
Within barrel acceptance**	true
Within endcap acceptance**	true
Relative isolation	< 0.4

Table: Electron cuts.

* The nanoAODplus already contains the cut pT$>$ 5 This will make a difference in the before variables, but it is irrelevant in the after variables.

** Either one of them must be true, but not both at the same time.

To select events of the type $ZZ \to 2\ell$ or $ZZ \to 4\ell$ we use Number of good electrons = 2 or 4. The total charge must be 0. In the four lepton case, the total charge of each pair of leptons must also be 0.

Validation of variables

Selection cuts

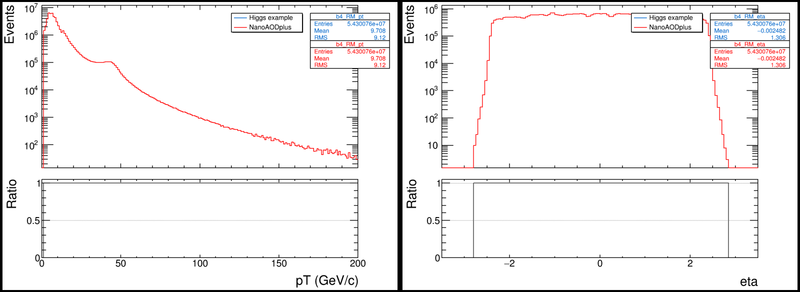
- If the before variables do not match, the after variables will be different.
- Our goal is to reproduce the variables in the Higgs example as similarly as possible, but using the same definitions as in the official nanoAOD from Run 2.
- This can lead to some differences.

Next step: find the similarities and differences and understand why they are produced.

Validation of variables

Data 2011

DOUBLE MUON 2011

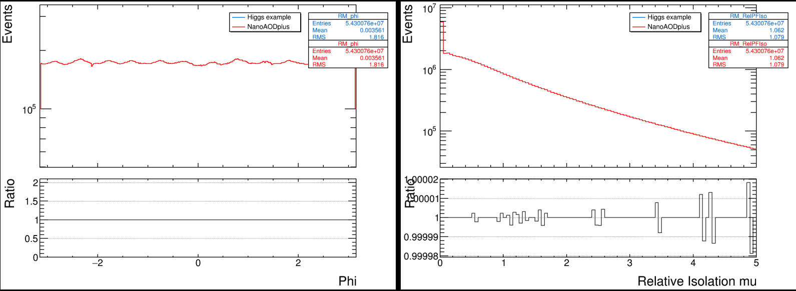


Muon pT (left) and eta (right) before.
Perfect match.

Validation of variables

Data 2011

DOUBLE MUON 2011



Muon phi (left) and relPFIso (right) before.
Perfect match.

Even though phi is not used in the cuts, we need it to calculate the invariant masses.

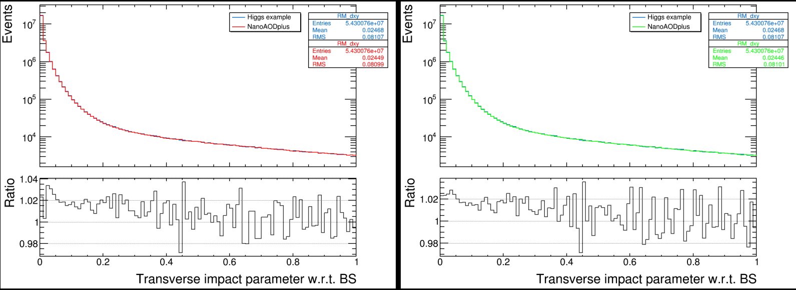
Validation of variables

Data 2011

DOUBLE MUON 2011

- We can define two types of variable for muon SIP3d, muon dxy and muon dz: best and non-best.

- Since the muons are sorted according to decreasing pT, the non-best variables use only the primary vertex of the first muon. On the other hand, the best variables include the primary vertices of all muons (because they are not necessary the same).

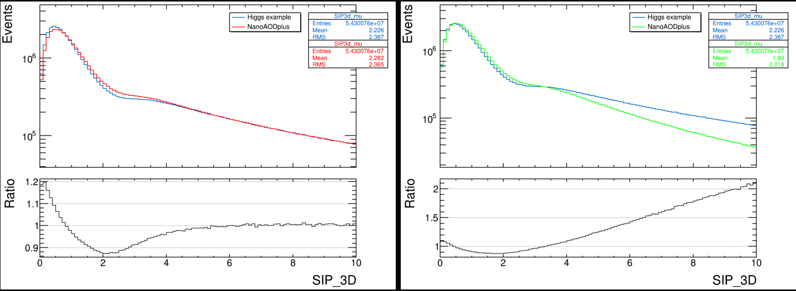


Muon dxy (left) and dxyBest (right) before.

Validation of variables

Data 2011

DOUBLE MUON 2011

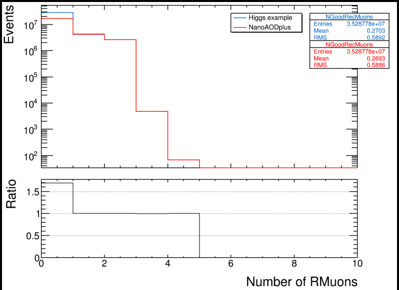


Muon SIP3d (left) and SIP3dBest (right) before.

Validation of variables

Data 2011

DOUBLE MUON 2011

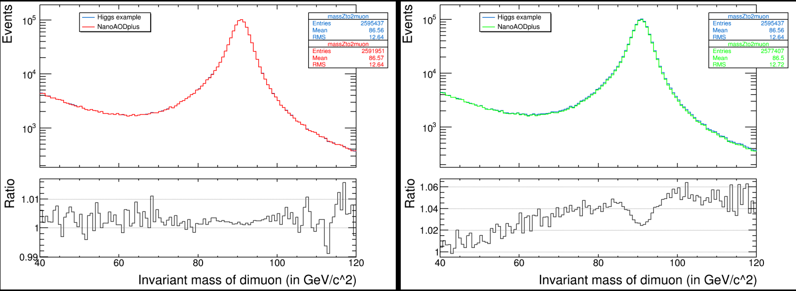


Number of good muons. There is a discrepancy in the first bin due to the type of variable (**float**), that can be solved by using **double** instead. The statistics box is correct.

Validation of variables

Data 2011

DOUBLE MUON 2011



Invariant mass of $Z \to 2\mu$ using non-best (left) and best (right) variables after.

Validation of variables

Data 2011

DOUBLE MUON 2011

- The variables pT, eta, phi and relPFIso before the cuts are the same as in the Higgs example. This means that the variable isPFIsolationValid is already implemented implicitly in the nanoAODplus.

- The variables dxy and SIP3d show a larger discrepancy. This is due to the definition of the primary vertex, which is not the same in both cases. Even if it gives slightly different results, it is preferable to use the same definition as in the official nanoAOD.

- The best matching invariant mass is the one calculated using the non-best variables.

Validation of variables

Data 2011

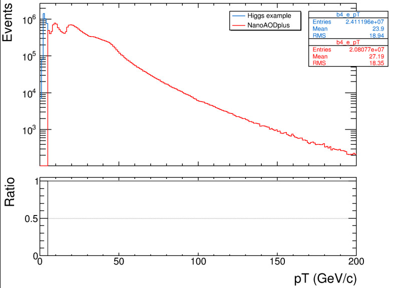
DOUBLE ELECTRON 2011

- The nanoAODplus has a cut in the electron pT. Because of this, we can not compare the before variables using the nanoAODplus leaves.
- Instead of that, we can produce histograms for the variables without the cut using the Nano Analyzer code, and check that they are the same as in Higgs Demo Analyzer $\rightarrow$ they are.
- We will see again that the variables SIP3d and dxy do not match due to the definition of the primary vertex.
- In this case we can check that the after variables do not differ too much.

Validation of variables

Data 2011

DOUBLE ELECTRON 2011

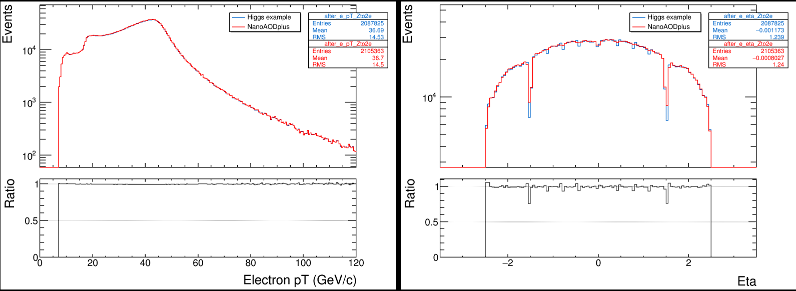


Electron pT before the cuts. Here we can see the cut in 5 GeV in the nanoAODplus.

Validation of variables

Data 2011

DOUBLE ELECTRON 2011

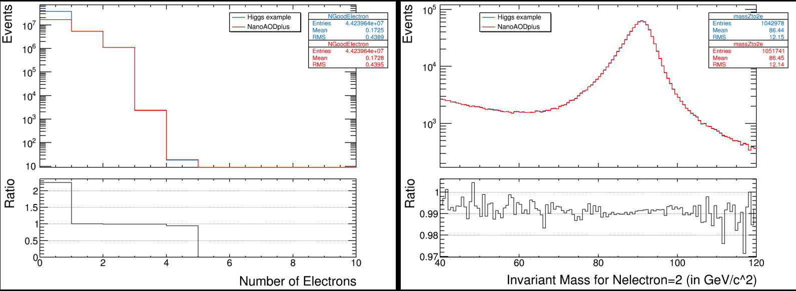


Electron pT (left) and eta (right) after the cuts (in the $Z \to 2e$ case).

Validation of variables

Data 2011

DOUBLE ELECTRON 2011



Number of good electrons (left) and invariant mass of $Z \to 2e$ (right) after. Again we see the discrepancy in the first bin, already explained for the dimuon case.

Validation of variables

Data 2011

DOUBLE ELECTRON 2011

- The variables before the cuts are validated through the histograms without the pT cut from Nano Analyzer.

- The SIP3d and dxy variables do not match due to the primary vertex definition, but the discrepancy is small.

- The variables after the cuts are similar.

Validation of variables

Data 2012

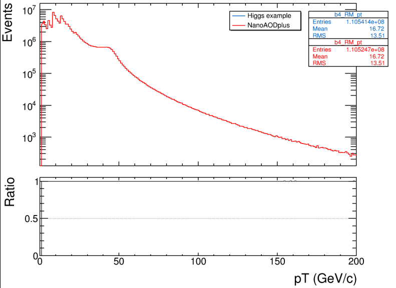
DOUBLE MUON 2012

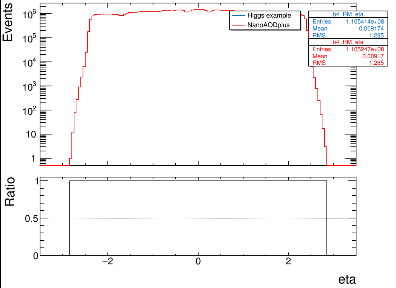
- In this case the before variables are different. This is because the size of the arrays storing the muon data per event (32) are in some cases smaller than the number of muons in that event. This means that not all muons are taken into account.
- This was not a problem in the 2011 data because the center of mass energy was lower (7 TeV) and less muons were produced.
- This feature does not affect significantly the variables, because it is highly improbable to have an event with more than 32 muons.
- We do not find this problem with electrons because the size of the arrays is 100.

Validation of variables

Data 2012

DOUBLE MUON 2012



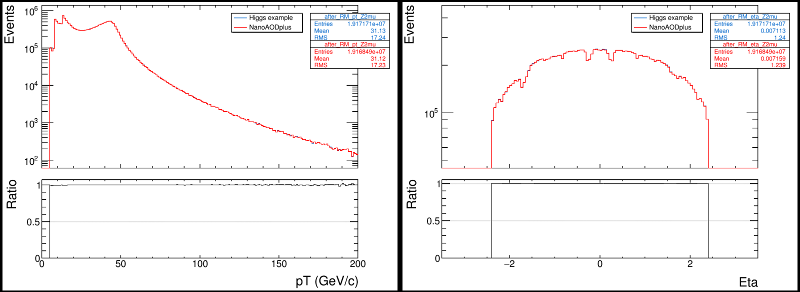


Muon pT (left) and eta (right) before.

Validation of variables

Data 2012

DOUBLE MUON 2012

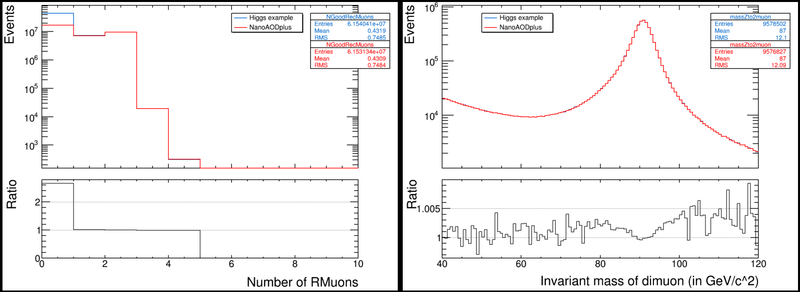


Muon pT (left) and eta (right) after the cuts (in the $Z \rightarrow 2\mu$ case).

Validation of variables

Data 2012

DOUBLE MUON 2012



Number of good muons (left) and invariant mass of $Z \to 2\mu$ (right) after.

Validation of variables

Data 2012

DOUBLE MUON 2012

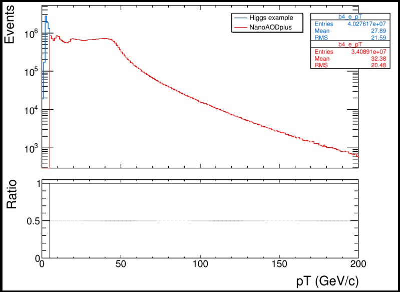
- The variables here are slightly different, probably due to the size of the arrays.
- The SIP3d and dxy are also defined differently.
- Again the first bin in the number of good muons do not match, because of the variable type.
- The after variables are almost equal to the ones in the Higgs example.

Validation of variables

Data 2012

DOUBLE ELECTRON 2012

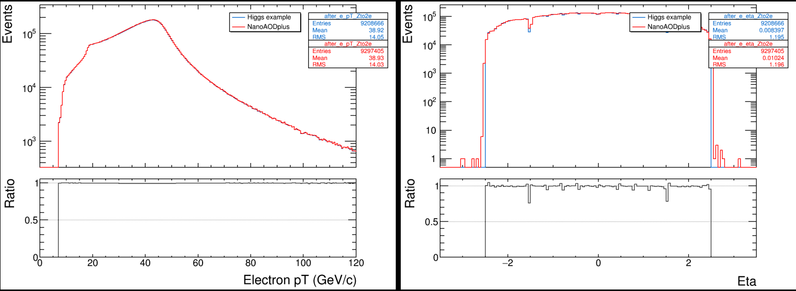
- This case is similar to DOUBLE ELECTRON 2011. Again we have the cut pT$>$ 5 GeV.



Validation of variables

Data 2012

DOUBLE ELECTRON 2012

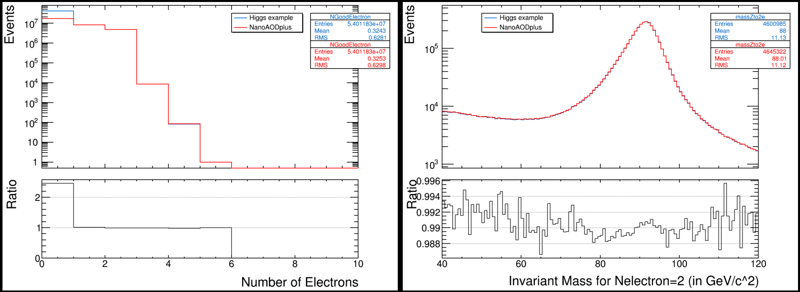


Electron pT (left) and eta (right) after the cuts (in the $Z \to 2e$ case).

Validation of variables

Data 2012

DOUBLE ELECTRON 2012



Number of good electrons (left) and invariant mass of $Z \to 2e$ (right) after.

Validation of variables

Data 2012

DOUBLE ELECTRON 2012

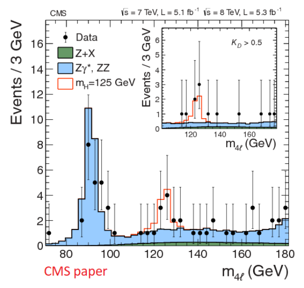
- Again the after variables are pretty similar.

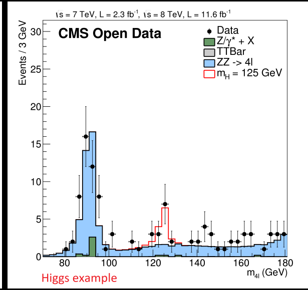
Validation of variables

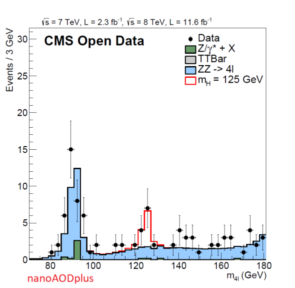
Monte Carlo

- The statistics in the data histograms are not enough to describe the case $ZZ \to 4\ell$.
- By using the Monte Carlo simulation we can see the invariant masses of the four leptons with a sufficiently high number of events.
- Here we have four types of datasets:
  1. $H \to ZZ \to 4\ell$
  2. $ZZ \to 4\mu$
  3. $ZZ \to 4e$
  4. $ZZ \to 2\mu 2e$

Higgs to four leptons plot







Conclusions

- The variables that we have compared either match or are very similar.
- We understand the reasons of the differences in the data histograms.
- There is still some work to do with the Higgs to four leptons plot.