CMS Draft Analysis Note

The content of this note is intended for CMS internal use and distribution only

2017/07/05
Head Id: 410820
Archive Id: 414156:414759M
Archive Date: 2017/06/16
Archive Tag: trunk

WW cross section measurement using random forests

Thoth Gunter and Michael Schmitt
Northwestern University

Abstract

We present measurements of the fiducial cross sections for the inclusive production of WW using a random forest classifier.

Contents

1 Introduction

Measurements of diboson cross sections are important tests of the electroweak sector of the standard model (SM). The earliest measurements were made at LEP and the Tevatron, and precise measurements have been reported by the LHC collaborations.

Theoretical advances have led to precise predictions. QCD corrections play an important role. The earlier measurements at 7 TeV and 8 TeV confirmed the theoretical predictions for WZ and ZZ final states, but they tended to be a bit above the predictions for the WW final state.

This analysis was motivated by the discrepancy between the measured and theoretical values for the WW cross section. The possibility to avoid a hard cut on the jet multiplicity by using a random forest classifier seemed attractive, and we reported an initial generator-level study in Ref. [1]. According to that study, a superior separation of signal and background can be achieved using a random forest classifier. Subsequently, studies with CMS full simulation showed the performance of Ref. [1] is difficult to achieve; nonetheless, this analysis note reports a very good performance.

Table 1: Summary of prior measurements of diboson cross sections (pb)

Collab.	$\sqrt{s}$	Ref.	measured	predicted
CMS	7 TeV	[2]	$52.4 \pm 2.0 \pm 4.5 \pm 1.2$	47.0 ± 2.0
ATLAS	7 TeV	[3]	$51.9 \pm 2.0 \pm 3.9 \pm 2.0$	$44.7^{+2.1}_{-1.9}$
CMS	8 TeV	[4]	$60.1 \pm 0.9 \pm 3.2 \pm 3.1 \pm 1.6$	$59.8^{+1.3}_{-1.1}$
ATLAS	8 TeV	[5]	$71.1 \pm 1.1^{+5.7}_{-5.0} \pm 1.4$	$63.2^{+1.6}_{-1.4} \pm 1.2$
CMS	13 TeV	[6]	$115.3 \pm 5.8 \pm 6.4 \pm 3.6$	120.3 ± 3.6
ATLAS	13 TeV	[7]	$0.529 \pm 0.020 \pm 0.050 \pm 0.011$	0.478 ± 0.017

1.1 Summary of prior measurements

The diboson (WW, ZZ, WZ) cross sections have been measured by LHC collaborations at $\sqrt{s} =$ 7 TeV, 8 TeV, and 13 TeV. A summary of these measurements is given in Table 1. The ATLAS 13 TeV measurement is for the specific final state $\mathrm{W}^+\mathrm{W}^- \rightarrow e\mu$.

1.2 Analysis strategy

The goal is to measure the inclusive WW cross section and to verify the jet multiplicity distribution. We focus on WW events in which both W bosons decay leptonically; we consider decays to electrons and muons but we do not explicitly identify and select decays to tau leptons. The final state consists, therefore, of two leptons of opposite charge and two neutrinos, and the leptons can be two electrons, two muons, or one electron and one muon. In this note, "same flavor" refers to the two-electron and two-muon channels together, while "different flavor" refers to the electron+muon channel.

The primary sources of background are Drell-Yan (DY) production and $\mathrm{t\bar{t}}$ production, so the selection is geared mainly to suppressing events from these sources. The selection proceeds in three steps:

1. Select events with two good leptons, veto events with a third lepton or with a b-tagged jet.
2. Select events that have high scores from two random forest (RF) classifiers. The first one is designed to suppress DY events, while the second suppresses $\mathrm{t\bar{t}}$ events. This "tight" RF selection is used to measure the fiducial cross section.
3. Loosen the requirements on the random forest scores somewhat; this "loose' RF selection is used to measure the jet multiplicity.

Necessary details are given in the remainder of this note.

2 Data and Simulations

Data recorded during the LHC 8 TeV and 13 TeV run periods were used in this analysis. The data sets used correspond to the 19.7 fb^{-1} (Run A-D) and 35.6 fb^{-1} (Run B-H) run periods respectively. All runs for both 8 and 13 TeV were taken at 25ns bunch spacings. The following subsections describe the data and Monte Carlo sets for each run period in greater detail.

2.1 8 TeV

This analysis utilizes proton proton collision data at 8 TeV with a total integrated integrated luminosity of 19.7 fb^{-1}. The data was reprocessed during the 2013 repocessing campaign with the submission dated *22Jan2013*.

Table 2: Data taking periods and lepton streams used for the 8 TeV analysis.

Data Taking Era	Stream
Run2012A	Single Muon, Single Electron
Run2012B	Single Muon, Single Electron
Run2012C	Single Muon, Single Electron
Run2012D	Single Muon, Single Electron

Single lepton triggers were used for the 8 TeV analysis. The triggers used are summarized in Table 3.

Table 3: HLT paths used in the 8 TeV analysis.

Dataset	HLT path
Single Muon	HLT_IsoMu24_eta2p1_v*
Single Electron	HLT_Ele27_WP80_v*

This analysis leverages several Monte Carlo generators. WW signal samples were produced with PYTHIA and TAUOLA. Backgrounds were produced with a combination of MADGRAPH, PYTHIA and TAUOLA. Table 4 summarizes the Monte Carlo samples.

Table 4: 8 TeV Monte Carlo signal and background sets.

Process	Dataset Name	Events	$\sigma \times BR[pb]$
Drell-Yan	DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball	30M	3531.9
Top Anti-Top	TTJets_FullLeptMGDecays_8TeV-madgraph-tauola	12M	23.81
Top Anti-Top	TTJets_SemiLeptMGDecays_8TeV-madgraph-tauola	24M	107.66
Top W	T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	500K	11.177
Anti-Top W	Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	493K	11.177
Top s channel	T_s-channel_TuneZ2star_8TeV-powheg-tauola	260K	3.179
Anti-Top s channel	Tbar_s-channel_TuneZ2star_8TeV-powheg-tauola	100K	1.76
Top t channel	T_t-channel_TuneZ2star_8TeV-powheg-tauola	3M	56.4
Anti-Top t channel	Tbar_t-channel_TuneZ2star_8TeV-powheg-tauola	2M	30.7
$WZ \rightarrow 3l\nu$	WZJetsTo3LNu_TuneZ2_8TeV-madgraph-tauola	2M	1.07
$WZ \rightarrow 2l2q$	WZJetsTo2L2Q_TuneZ2star_8TeV-madgraph-tauola	3M	2.24
$ZZ \rightarrow 2l2q$	ZZJetsTo2L2Q_TuneZ2star_8TeV-madgraph-tauola	2M	2.47
$ZZ \rightarrow 2l2\nu$	ZZJetsTo2L2Nu_TuneZ2star_8TeV-madgraph-tauola	954K	0.71
Higgs	GluGluToHToWWTo2LAndTau2Nu_M-125_8TeV-powheg-pythia6	299K	0.43
$WG \rightarrow l\nu\gamma$	WGToLNuG_TuneZ2star_8TeV-madgraph-tauola	5M	553.9
$WG \rightarrow 2e\nu$	WGstarToLNu2E_TuneZ2star_8TeV-madgraph-tauola	314K	5.873
$qq \rightarrow WW$	WW_TuneZ2star_8TeV_pythia6_tauola	10M	57.2
$gg \rightarrow WW \rightarrow 4l$	GluGluToWWTo4L_TuneZ2star_8TeV-gg2ww-pythia6	109K	0.18

2.2 13 TeV

This analysis utilizes proton proton collision data at 13 TeV with a total integrated integrated luminosity of 35.9 fb^{-1}. The data was reprocessed during the 2017 repocessing campaign with the submission dated *03Feb2017*.

Single lepton triggers were used for the 13 TeV analysis. The triggers used are summarized in Table 6.

This analysis leverages several Monte Carlo generators. WW signal samples were produced with PYTHIA and TAUOLA. Separate simulated sets were used for partons and gluons production methods. Backgrounds were produced with a combination of MADGRAPH, PYTHIA and TAUOLA. Table 7 summarizes the Monte Carlo samples.

Table 5: Data taking periods and lepton streams used for the 13 TeV analysis.

Data Taking Era	Stream
Run2012B-v1	Single Muon, Single Electron
Run2012B-v2	Single Muon, Single Electron
Run2012C-v1	Single Muon, Single Electron
Run2012D-v1	Single Muon, Single Electron
Run2012E-v1	Single Muon, Single Electron
Run2012F-v1	Single Muon, Single Electron
Run2012G-v1	Single Muon, Single Electron
Run2012H-v2	Single Muon, Single Electron
Run2012H-v3	Single Muon, Single Electron

Table 6: HLT paths used in the 13 TeV analysis.

Dataset	HLT path
Single Muon	HLT_IsoMu24_v*, HLT_IsoTkMu24_v*
Single Electron	HLT_Ele27_WPTight_Gsf_v*

3 Event Selection

We describe the reconstructed objects used in the analysis and the pre-selection requirements. The most powerful selection criteria are based on the random forest classifier, described in Section 4.

The analysis proceeds in two steps:

1. **Preselection:** Relatively loose criteria are imposed that select events with two leptons and no b-tagged jets.
2. **Classification with the random forest:** Several event observables, called “features”, are used as input to two random forest classifiers. One is designed to reject Drell-Yan events, and the other is designed to reject $t\bar{t}$ events. Most of the rejection of background is achieved by applying cuts to the scores of the two random forest classifiers.

This section is concerned mainly with the first step and with the event observables used by the random forest classifiers.

3.1 Lepton reconstruction and selection

We intend to measure the WW cross section in the dilepton channel, therefore we need two “good” leptons. “good” leptons are defined by a set of identification and isolation cuts developed by the Muon and Electron POGs [8–11]).

We identify leptons via the Particle-Flow algorithm. We apply identification and isolation cuts and define a transverse momentum (pt) cut of 25 (30) GeV for the leading muon (electron). To remove poor leptons a minimum lepton p_{T} threshold was set 20 GeV. Events must also pass either of the single lepton triggers. Tables 8 – 11 summarizes the criteria used. The identification and isolation criteria was selected from CMS Tight recommendations for 8 and 13 TeV respectively.

Table 7: 13 TeV Monte Carlo signal and background sets.

Process	Dataset Name	Events	$\sigma \times BR[pb]$
Drell-Yan	DYJetsToLL_M-50_TuneCUETP8M1_13TeV-amcatnloFXFX-pythia8	121M	5765.4
Top Anti-Top	TTTo2L2Nu_TuneCUETP8M2_ttHtranche3_13TeV-powheg-pythia8	79M	87.31
Top Anti-Top	TTToSemilepton_TuneCUETP8M2_ttHtranche3_13TeV-powheg-pythia8	91M	364.35
Top W	ST_tW_top_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M2T4	992K	35.85
Anti-Top W	ST_tW_antitop_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M2T4	987K	35.85
Top s channel	ST_s-channel_4f_leptonDecays_13TeV-amcatnlo-pythia8_TuneCUETP8M1	1M	3.36
Anti-Top s channel	—	—	—
Top t channel	ST_t-channel_top_4f_inclusiveDecays_TuneCUETP8M2T4_13TeV-powhegV2-madspin	6M	136.02
Anti-Top t channel	ST_tW_antitop_5f_inclusiveDecays_13TeV-powheg-pythia8_TuneCUETP8M2T4	3M	80.95
$WZ \rightarrow 3l\nu$	WZTo3LNu_TuneCUETP8M1_13TeV-powheg-pythia8	2M	5.29
$WZ \rightarrow 2l2q$	WZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	26M	5.595
$ZZ \rightarrow 2l2q$	ZZTo2L2Q_13TeV_amcatnloFXFX_madspin_pythia8	15M	3.22
$ZZ \rightarrow 2l2\nu$	ZZTo2L2Nu_13TeV_powheg_pythia8	9M	0.564
Higgs	—	—	—
$WG \rightarrow l\nu\gamma$	—	—	—
$WG \rightarrow 2e\nu$	—	—	—
$qq \rightarrow WW$	WWTo2L2Nu_13TeV-powheg	2M	118.7 * (3*.108)**2
$gg \rightarrow WW \rightarrow 4l$	—	—	—

Table 8: 8 TeV muon identification and isolation summary.

Identification cuts	
feature	cut
$\lvert\eta\rvert$	< 2.1
χ^2	< 10
Number of Valid Hits	> 0
Number of Matched Stations	> 1
Number of Tracking Layer Measurements	> 5
Dz	< 0.5
Dxy	< 0.2
Isolation cuts	
Sum Particle Flow Iso	< 0.12
track Isolation	< 0.1

3.2 Jets and $E_{\mathrm{T}}^{\mathrm{miss}}$

3.2.1 Jets

While we select events using a lepton criteria, jet and $E_{\mathrm{T}}^{\mathrm{miss}}$ objects are important in discriminating background events. Jets are reconstructed using the particle flow algorithm. The particle flow algorithm reconstructed physics objects using hits and energy deposits within the various subdetectors.

The 8 TeV analysis used the Anti-k_T clusting algorithm with a ΔR cut of 0.5. The 13TeV analysis used the Anti-k_T clusting algorithm with a ΔR cut of 0.4.

ΔR is the anglar distance in ϕ and η between physics objects. A ΔR cut defines an angular cone, shich we are to reject additional particles. We accept jets with an $|\eta| < 4.7$ and $p_{\mathrm{T}} > 30$. We also apply identification cuts defined by the Jet Met POG [12].

The number and energy of jets are impacted by the pileup [13]. We apply a jet energy correction to take this into account.

To determine this correction the contribution to jet energy from pileup is estimated from data. There are two steps to applying the correction. The first is the L1$FastJet$ correction which subtracts the a mean energy density from the jet p_{T} [14]. This correction, in addition to removing the pileup component it also removed some of the underlaying event contribution. The second

Table 9: 13 TeV muon identification and isolation summary.

Identification cuts	
feature	cut
$\lvert\eta\rvert$	< 2.4
χ^2	< 10
Number of Valid Hits	> 0
Number of Matched Stations	> 1
Number of Tracking Layer Measurements	> 5
Dz	< 0.5
Dxy	< 0.2
Isolation Cuts	
Sum Particle Flow Iso	< 0.12

Table 10: 8 TeV electron identification and isolation summary.

Identification cuts		
	EB cut	EE cut
$\lvert\eta\rvert$	< 1.479	$1.479 \leq \lvert\eta\rvert < 2.4$
SCDeltaEta	< 0.004	< 0.0066
SCDeltaPhi	< 0.03	< 0.03
SigmaIEtaIEta	< 0.01	< 0.03
Hadon Over Em	< 0.12	< 0.10
Dz	< 0.1	< 0.1
Dxy	< 0.02	< 0.02
Isolation cuts		
SumPFIso	< 0.1	< 0.1

is a jet energy scale corrections attempts to correct for this **??**.

3.2.2 Missing Transverse Energy ($E_{\mathrm{T}}^{\mathrm{miss}}$)

Missing transverse energy is a powerful tool used to discriminate against Drell Yan events. For this analysis we use particle flow $E_{\mathrm{T}}^{\mathrm{miss}}$. In addition to the standard $E_{\mathrm{T}}^{\mathrm{miss}}$ variable we construct additional $E_{\mathrm{T}}^{\mathrm{miss}}$ variables to further distringuish Drell Yan events from signal events. These variables are projected missing transverse energy and ratio of missing transverse energy to the transverse energy of the event. Projected missing energy is defined in equation 1 Drell Yan events that have two charged leptons should effectively have no $E_{\mathrm{T}}^{\mathrm{miss}}$. The $E_{\mathrm{T}}^{\mathrm{miss}}$ that we see in these events are resolutions effects and often have a ϕ close to the ϕ of one of the leptons. The $E_{\mathrm{T}}^{\mathrm{miss}}$ projected variable highlights this relationship. The ratio of missing transverse energy to the transverse energy of the charged leptons is a variable that highlights the magnitude difference between fake $E_{\mathrm{T}}^{\mathrm{miss}}$ ($E_{\mathrm{T}}^{\mathrm{miss}}$ from resolution effects) and the real $E_{\mathrm{T}}^{\mathrm{miss}}$ one expects to see in a dilepton event.

During the 8TeV analysis period Drell Yan Monte Carlo produced showed a strong disagreement with data. Specifically Monte Carlo events were in excess at low MET while at the substantially decreasing in the real $E_{\mathrm{T}}^{\mathrm{miss}}$ turn on region.(SHOULD PHRASE THIS DIFFERENTLY) We apply a scaling factor to the Drell Yan distribution of 1.057 to correct the disagreement. We determine the scaling factor by minimizing the data and Monte Carlo $E_{\mathrm{T}}^{\mathrm{miss}}$ distributions in the Z peak. Plots **??** show that this correction aligns data and Monte Carlo.

Table 11: 13 TeV electron id and iso ADD CAPTION ADD energy Inversion and nMissing Hits

Identification cuts	EB cut	EE cut
$\|\eta\|$	< 1.479	$1.566 \leq \|\eta\| < 2.4$
SCDeltaEta	< 0.009	< 0.00729
SCDeltaPhi	< 0.0336	< 0.0918
SigmaIEtaIEta	< 0.0101	< 0.0279
Hadon Over Em	< 0.0597	< 0.0615
Dz	< 0.466	< 0.417
Dxy	< 0.011	< 0.0351
Isolation cuts SumPFIso	< 0.1	< 0.1

Table 12: 8 TeV Jet selection.

	$\|\eta\| < 2.4$	$2.4 \leq \|\eta\| < 4.7$
Neutral Hadron Fraction	> 0.9	-
Neutral Em Fraction	> 0.9	-
Number of Constituents	$> 1.$	-
Charged Em Fraction	> 0.99	0.99
Charged Hadron Fraction	> 0	0
Pt	$> 30.$	
DR mu el	< 0.4	
DR me	< 0.3	
vtxCut	$< .3$	

$$\text{projected}E_{\mathrm{T}}^{\text{miss}} = \begin{cases} E_{\mathrm{T}}^{\text{miss}} & if\Delta\phi_{\text{min}} \geq \frac{\pi}{2} \\ E_{\mathrm{T}}^{\text{miss}} \sin(\Delta\phi_{\text{min}}) & if\Delta\phi_{min} < \frac{\pi}{2} \end{cases} \tag{1}$$

where $\Delta\phi_{\text{min}} = \min(\Delta\phi(l_1, E_{\mathrm{T}}^{\text{miss}}), \Delta\phi(l_2, E_{\mathrm{T}}^{\text{miss}}))$.

3.3 b-jet tagging

b-jets are selected from the collection of jets that pass the jet selection criteria. We identify b-jets using the combined secondary vertex algorithm version 2 (csv2), tight (medium) selection criteria for 8 (13) TeV. The csv, C, algorithm combines several topological and kinematical secondary vertex related variables as well as information from track impact parameters to discriminate between jets that originate from b-quarks **??**.

Table 13: 8 TeV and 13 TeV b-jet tagging.

	cut
8 TeV	$C < 0.898$
13 TeV	$C < 0.8484$

Top background events are the main producers of b-jets. The top-qark decays into a W boson and b-quark pair the vast majority time, therefore it is important to remove as many b-jets as possible. For this analysis we reject any event with a b-jet.

3.4 Pre-selection of events

This measurement considers all three fully leptonic WW final states: e^+e^-, $\mu^+\mu^-$ and $e^\pm\mu^\mp$. The W→ $\tau\nu_\tau$ is counted as a signal, though the selection is not optimized for it. The triggers require a high momentum lepton of p_{T} greather than 24 and 27 GeV for muon and electron single lepton triggers respectively. In addition we require a momentum to be greater than 25(30) GeV for the leading muon(electron). The additional momentum cut is placed higher than the trigger threshold to fall in the high efficiency region of the triggers.

The preselection is designed to be simple and to retain as many signal events as possible. Those cuts are defined as follows:

- Exactly two good oppositely charged leptons.
  - Passes ID and isolation cuts.
  - Leptons with $|\eta| < 2.4$
  - Leading leptons with $p_{\mathrm{T}} > 25$ ($p_{\mathrm{T}} > 30$) muon (electron).
  - Sub leading lepton with $p_{\mathrm{T}} > 20$.
- 15 GeV invariant mass cut around the Z peak for same flavor leptons pairs. ($74 < mll < 106$)
- Zero bjets.
- Invariant mass > 30 GeV to remove low reasonances.

3.5 Validation

To verify the Monte Carlo data agreement we study various kinematic distributions. Concurrency between the shape and yields of these distributions validate the agreement. To more rigorously analyze these distributions we develope a set of selections that enhance select physics processes. We are particularly interested in high jet regions primarily populated by top-quark events, high $E_{\mathrm{T}}^{\mathrm{miss}}$ events primarily populated by WW and top-quark , low $E_{\mathrm{T}}^{\mathrm{miss}}$ same flavor region populated by Drell Yan events ,and opposite flavor events that derive from tau decays. We define the selection in these tables 14, 15.

3.5.1 8 TeV Validation

Figures **??**, **??** show Mll, $d\phi$, $E_{\mathrm{T}}^{\mathrm{miss}}$ and number of jets plots. These plots as well as the other preselection plots, which can be found in the appendix, show very good agreement between data and Monte Carlo.

3.5.2 13 TeV Validation

Figures **??**, **??** show Mll, $d\phi$, $E_{\mathrm{T}}^{\mathrm{miss}}$ and number of jets plots. These plots as well as the other preselection plots, which can be found in the appendix, show very good agreement between data and Monte Carlo.

Table 14: Drell Yan and Z → $\tau\tau$ control regions.

Z → $\tau\tau$	
qT	> 30
Number of jets	> 0
$E_{\mathrm{T}}^{\mathrm{miss}} > 30$ and $E_{\mathrm{T}}^{\mathrm{miss}} < 50$	
Mll > 60 and Mll < 120	
different flavor leptons	

Drell Yan selection	
Number of jets	$== 0$
$E_{\mathrm{T}}^{\mathrm{miss}}$	< 60
different flavor leptons	

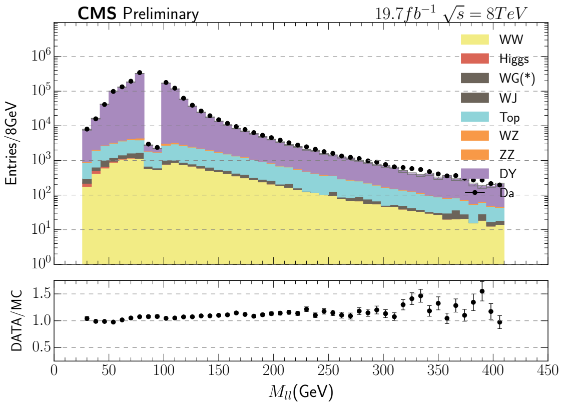

Figure 1: first figure

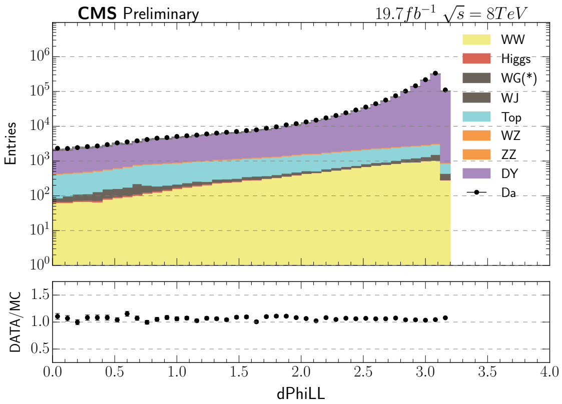

Figure 2: second figure

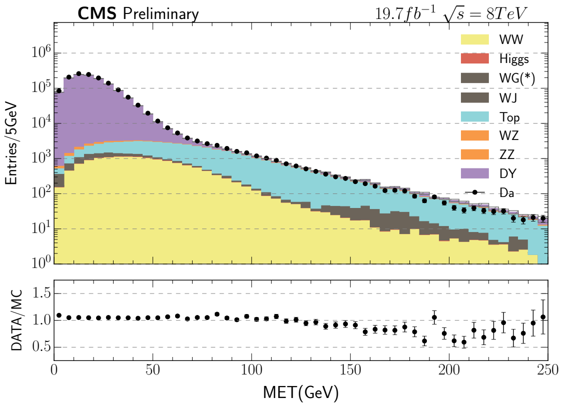

Figure 3: first figure

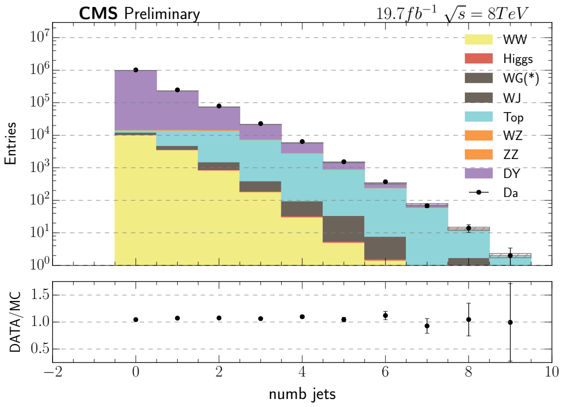

Figure 4: second figure

Table 15: Top-quark and WW control regions.

Top-qark selection	
qT	> 30
Number of jets	> 0
$E_{\mathrm{T}}^{\mathrm{miss}}$	> 60
Number of jets	> 0
different flavor leptons	

WW selection	
Number of jets	$== 0$
$E_{\mathrm{T}}^{\mathrm{miss}}$	> 50
different flavor leptons	

Table 16: Top-quark

Process	Diff Flavor	Same Flavor
WW	718	0
GluGluToWWTo4L	38	0
WW	756	0
DY	387	0
Top	8061	0
WZ	83	0
ZZ	4	0
WG(*)	35	0
Higgs	6	0
WJ	310	0
Total	9642	0
DATA	9792	0

4 Random Forest Classifier

We employ random forest classifiers to obtain a pure sample of WW events. Since random forest classifiers are not commonly used in high energy physics, we begin with a general introduction to this powerful tool.

4.1 General Introduction to Random Forest Classifiers

The random forest is a classification and regression tool developed by Leo Breiman and Adele Cutler. The specifics of the random forest can be found in Breiman's paper Ref. [15]. The random forest is an aggregated estimator that combines many unique tree estimators into one estimator. Simply put the random forest is a collection of tree estimators each trained using a random subset of features and training data. Feature and training sets are chosen with replacement. Through aggegation, random forests are able to utilize the strengths of tree estimators, low bias and conceptual simplicity, while reducing the effects of their weakness, high variance. The following is an introduction to random forest. We will dicuss the tree algorithm, forest aggregation and the sci-kit learn [16] implementation we are using for this analysis. The following selection will discuss the architechture and design of the random forest used in this analysis.

The tree estimator is the backbone of the random forest. The standard implementation of the tree estimator is the Classification And Regression Tree (CART); for this analysis we use the CART implementation. Classification trees (decision trees) are constructed from a hierarchy

Table 17: WW

Process	Diff Flavor	Same Flavor
WW	1830	0
GluGluToWWTo4L	81	0
WW	1911	0
DY	398	0
Top	683	0
WZ	44	0
ZZ	2	0
WG(*)	42	0
Higgs	12	0
WJ	0	0
Total	3093	0
DATA	3252	0

Table 18: Drell Yan

Process	Diff Flavor	Same Flavor
WW	4952	4064
GluGluToWWTo4L	120	103
WW	5072	4168
DY	20154	972038
Top	811	674
WZ	119	214
ZZ	6	203
WG(*)	145	102
Higgs	59	58
WJ	1074	3401
Total	27439	980859
DATA	27241	1026259

of feature cuts. This hierarchy can be imagined as a node graph where each node represents a feature cut. For the CART implementation these cuts are generally binary. This cut value segregates events determining which nodes an event might matricuate through as it progresses through the cut hierarchy of cuts. SOME PLOT WITH GOOD CAPTON Once the event has propagated to the end of the decision tree, the event is scored. The score is a reflection of the training set distribution that passed the same set of cuts.

Decision tree training is done in two steps. These sets are repeated for each node. The steps are:

1. Select the best feature to cut on, from the list of features that have yet to be cut on, in that respective path.
2. Determine and place best cut for the current feature.

One common way to determine the best feature to cut on is to rank them by normalized variance. Normalized variance is the variance of a feature divided by the range of said feature. Features with higher variance can be thought of as the principle component of the set of features. Cut selection is done by minimizing the gini impurity. Gini impurity is a measurement

Table 19: $Z \rightarrow \tau\tau$

Process	Diff Flavor	Same Flavor
WW	39	0
GluGluToWWTo4L	1	0
WW	40	0
DY	524	0
Top	4	0
WZ	1	0
ZZ	0	0
WG(*)	0	0
Higgs	0	0
WJ	15	0
Total	584	0
DATA	585	0

of how often an element would be mislabled if that element was randomly selected from the distribution of labels it was selected from.

The gini impurity is given by:

$$I_g = \sum_{i}^{N_{\text{classes}}} f_i(1 - f_i)$$

where $N_{classes}$ is the number of unique lables, f is the purity of the subsample for the i^{th} label.

Once the feature list has been exhausted, maximum depth reached or minimum number of events has been reached, the decision tree ends and that branch is given a score. That score is based on the label distribution for the subset of events that has passed the hierarchy of cuts.

Random forests are constructed from decision trees, where each tree is constructed from a random subset of feature and a random subset of training data. These random subsets are chosen with replacement. By aggregating many unique trees the random forest is becomes both a low bias and low variance estimator. The score is given by aggregating the scores from each tree. Trees are weighted equally. Unlike other aggregated estimators trees are trained independently of each other. This allows training to occur in parallel.

For this analysis we use the scikit-learn machine learning package. Scikit-learn is a machine learning framework for python. Scikit-learn was chosen for this analysis because of its ease of use, relative transparency, and the ability to parallelize random forest training. We used scikit-learn version 0.18.2.

4.2 Architecture and design

We design and structure our random forest using hyperparameters. These hyperparameters define various aspsects of each tree and the forest as a whole. Hyperparameters are chosen to deminish overtraining while maximizing the accuracy. Table 20 is a sample set of the hyperparameters for the random forest.

We determined the best set of hyperparameters through grid search. Grid search is a technique for exploring multidimensional parameter space. We use accuracy to determine the best set of hyperparameter values. We ran a grid search on both Drell Yan and Top-quark random forest. We decided to use the same hyperparameters set for both forest. for simplicity.

The grid search was done by splitting the training set into two sets, a hyperparameter training and testing set. A random forest was created for every hyperparameter combination for each sub training set. Each forest is tested on the hyperparameter testing set.

Table 20: Grid search hyperparameter spectrum

Parameter	Search Values
min samples split	10, 50, 200
n estimators	25, 50, 75
max depth	5, 15, 20
min samples leaf	1, 10, 100

Table 21: Random Forest hyperparameters. The number of trees, maximum tree depth, and minimum number of samples per split were determined through grid search. The square root of number of features, gini impurity are standard settings for CART algorithms.

Parameter	Value
Number of trees	50
Maximum tree depth	20
Minimum number of samples per split	50
Maximum number of features	sqrt Number of Features

The range of scores for the grid search was 0.87707207 - 0.88898398.NOTE STANDARD DEVIATION INSTEAD

In addition to determining the hyperparameters, the features of the random forest were also determined. We began by curating a list of features that hold some information about the main background the signal processes. The list is given here 22.

Table 22: Grid search hyperparameter spectrum

Features	Drell Yan	Top-quarks
lepton flavor	✓	
number of jets		✓
subleading lepton p_{T}	✓	
$E_{\mathrm{T}}^{\mathrm{miss}}$	✓	✓
proj $E_{\mathrm{T}}^{\mathrm{miss}}$	✓	
qT	✓	✓
mll	✓	
mll$E_{\mathrm{T}}^{\mathrm{miss}}$	✓	
$d\phi_{ll,E_{\mathrm{T}}^{\mathrm{miss}}}$	✓	✓
$d\phi_{ll,Jet}$		✓
$d\phi_{E_{\mathrm{T}}^{\mathrm{miss}},Jet}$		✓
$d\phi_{l,l}$	✓	
HT		✓
recoil	✓	✓

We from this set we select features of high importance to keep. The importance is defined as the gini impurity after the best possible cut is made. As it defined the feature with the most important feature is the one with the least importance.

In our studies we found that the we acheived the best performance by using two random forest. One trained against the top background while the other against drell yan. We simultaneously apply the cuts from both forest.

4.3 Validation

To verify the Monte Carlo data agreement we study the random forest scores and various kinematic distributions in regions defined by cuts placed on the scores of these random forests. We define a set of cuts on the random forest scores to enhance the Drell Yan, top-quark and WW regions. Table 23

Table 23: Random forest background and signal control region cuts.

Enhanced Process	S_{DrellYan}	$S_{\mathrm{Top-quarks}}$
Drell Yan	$< .65$	–
Top-quark	> 0.5	$< .4$
WW	$< .6$	$< .6$

4.4 Cut Selection

5 Background Estimates

The main sources of backgrounds are Drell Yan and Top-qark for this analysis. By utilizing a preselection that removed the Z peak and b-jets, then applying cuts to the random forest output we remove the majority of our primary backgrounds. Of the remaining backgrounds top-quark backgrounds remain the largest. For both 8 and 13 TeV the tight random forest selectiona has a purity of about 60%. Tables **??** and **??** summarize the yields for 8 and 13 TeV for the tight selection.

We loosen the Top forest cut to probe the WW jet multiplicity. For the loose selection we found a purity of ??(??) for 8(13) TeV. Tables **??** and **??** summarize the yields.

6 Efficiencies

The cross section must be corrected for lepton inefficiencies. Since the simulation is not perfect, the efficiencies we estimate from Monte Carlo must be corrected to conform to the efficiencies measured with the data.

7 Systematic Uncertainties

There are several sources of systematic uncertainty, discussed here in detail. Lepton scale factor efficiency, b-jet scale factor efficiency, lepton and jet momentum scale, lepton and jet p_{T} resolution, and statistical uncertainties among others will be discussed in the following subsections.

7.0.1 Lepton scale factor efficiency

Lepton scale factor efficiencies are the ratio of data and Monte Carlo lepton efficiency. Lepton efficiency is the ratio of good lepton we select with our lepton criteria over the number of good leptons. We use standard scale factor efficiencies and uncertainties as defined by the muon and election POG for this analysis. We apply the scale factors as weights to each event. Scale factor efficiencies are a function of bins of p_{T} and η. This analysis is based on a dilepton selection, therefore we can not use the lepton scale factor uncertainies in is form. We must account for the this effect. To propagate the scale factor uncertainties we apply the following method.

For each lepton an offset is generated from the uncertainty on the scale factor of that lepton. Generate being defined as taking the uncertainty as the standard deviation of a gaussian. The offset is a random value taken from the aforementioned distribution. We apply the offset to the dilepton weight, and record the mean weight. We repeat this process 100 times, using the distribution of recorded means to determine the dilepton scale factor efficiency uncertainty.

THAT ONE EQ THAT TIES IT ALL UP

7.0.2 bjet scale factor inefficiency

The b-jet scale factor efficiencies are the ratio of the efficiency with which be find b-jets in data and Monte Carlo, much like lepton scale factors. We use scale factors and uncertainties from the official bjet recommendations. For this analysis we attempt to reject events with bjets and are therefore interested in inefficiencies. We define inefficiency as

$$\epsilon = \frac{1 - \mathrm{SF}\epsilon_{\mathrm{MC}}}{1 - \epsilon_{\mathrm{MC}}}$$

.

bjet scale factors are given in functional form therefore calculation fo the Monte Carlo bjet efficiencies were done internally. To calculate we take top-quark Monte Carlo data sets selecting events that have atleast one b jet, as defined by monte carlo truth and determine how many survive the csv selection as a function of p_{T}.

We determine the uncertainty on the efficiency by throwing toys where the scale factor and Monte Carlo efficiencies are altered by offsets generated from the scale factor and efficiency uncertainties. We calculate the standard deviation of the mean efficiencies from these toys. We take the standard deviation as the uncertainty of the bjet inefficiency.

THAT ONE EQ THAT TIES IT ALL UP

7.0.3 Lepton and jet momentum scale

The lepton and jet momentum scale uncertainty were completed by shifting the transverse momentum of leptons and jets(independantly) and progagating these changes to associated variables. The shift in momentum was taken from muon, electron and jet POG. The scales are 0.2% muon, 0.3%electron and 2.5% jets. We shift the momentum by multiplying the objects momentum by 1 ± scale. We average the difference between the upward and downward shifts and take this as the the lepton(jet) uncertainty.

7.0.4 Lepton and Jet p_{T} resolution

We take the lepton and jet p_{T} resolution from the muon, electron and jet pog and design reports. To determine the uncertainty on the cross section from momentum resolution we shift the momentum by an offset then propagate this to all associated variables. We determine this offset by sampling the gaussian distribution. The gaussian standard deviation is taken as the momentum of the particle multiplied by the resolution. We recalucate the cross section and take the difference between the original cross section and the new cross section as the uncertanty from momentum resolution.

7.0.5 Stat uncertainties

8 Results

We present our measurements of the cross section.

8.1 Fiducial cross section

8.2 Jet multiplicity

8.3 Comparison of 8 TeV and 13 TeV results

9 Summary and Conclusions

Summary and conclusions.

References

[1] J. L. Rainbolt, T. Gunter, and M. Schmitt, "Using Random Forests to Classify W^+W^- and $\mathrm{t\bar{t}}$ Events", `arXiv:1410.8058`.

[2] CMS Collaboration, "Measurement of the $\mathrm{W^+W^-}$ Cross section in pp Collisions at $\sqrt{s} = 7\,\mathrm{TeV}$ and Limits on Anomalous $WW\gamma$ and WWZ couplings", *Eur. Phys. J.* **C73** (2013), no. 10, 2610, `doi:10.1140/epjc/s10052-013-2610-8`, `arXiv:1306.1126`.

[3] ATLAS Collaboration, "Measurement of W^+W^- production in pp collisions at $\sqrt{s} = 7\,\mathrm{TeV}$ with the ATLAS detector and limits on anomalous WWZ and WWγ couplings", *Phys. Rev.* **D87** (2013), no. 11, 112001, `doi:10.1103/PhysRevD.87.112001,10.1103/PhysRevD.88.079906`, `arXiv:1210.2979`. [Erratum: Phys. Rev.D88,no.7,079906(2013)].

[4] CMS Collaboration, "Measurement of the $\mathrm{W^+W^-}$ cross section in pp collisions at $\sqrt{s} = 8\,\mathrm{TeV}$ and limits on anomalous gauge couplings", *Eur. Phys. J.* **C76** (2016), no. 7, 401, `doi:10.1140/epjc/s10052-016-4219-1`, `arXiv:1507.03268`.

[5] ATLAS Collaboration, "Measurement of total and differential W^+W^- production cross sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector and limits on anomalous triple-gauge-boson couplings", *JHEP* **09** (2016) 029, `doi:10.1007/JHEP09(2016)029`, `arXiv:1603.01702`.

[6] CMS Collaboration, "Measurement of the $\mathrm{W^+W^-}$ cross section in pp collisions at $\sqrt{s} = 13\,\mathrm{TeV}$", CMS Physics Analysis Summary CMS-PAS-SMP-16-006, 2016.

[7] ATLAS Collaboration, "Measurement of the W^+W^- production cross section in pp collisions at a centre-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ with the ATLAS experiment", `arXiv:1702.04519`.

[8] https://twiki.cern.ch/twiki/bin/view/CMS/EgammaCutBasedIdentificationC. E. POG, "Egamma Cut Based Identification Run 1",.

[9] https://twiki.cern.ch/twiki/bin/view/CMS/EgammaCutBasedIdentificationRun2C. E. POG, "Egamma Cut Based Identification Run 2",.

[10] C. M. POG, "Baseline muon selections for Run-I",.

[11] C. M. POG, "Baseline muon selections for Run-II",.

[12] C. J. POG, "Jet Identification",.

[13] CMS Collaboration, "Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV", CMS Physics Analysis Summary CMS PAS PFT PFT-10-002, 2010.

[14] M. Cacciari and G. P. Salam, “Pileup subtraction using jet areas”, *Phys. Lett.* **B659** (2008) 119–126, doi:10.1016/j.physletb.2007.09.077, arXiv:0707.1378.

[15] L. Breiman, “Random Forests”, *Machine Learning* **45** (2001), no. 1, 5–32, doi:10.1023/A:1010933404324.

[16] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research* **12** (2011) 2825–2830.