July 13, 2010

Performance of QCD background estimation methods for single lepton SUSY searches

O. Buchmueller[a)], Z. Hatherell[a)], G. Karapostoli[a)], A. Sparrow[a)], P. Sphicas[b,c)], A. Tapper[a)]

a) Imperial College London, United Kingdom

b) CERN, Geneva, Switzerland

c) University of Athens, Greece

Abstract

The performance of data-driven QCD background estimation methods to be used for single lepton SUSY searches is studied. The methods are contructed using Monte Carlo datasets, and the performance is studied with an X pb^{-1} sample of 7 TeV LHC collisions. The data provides many events with one electron and at least one jet, with which we validate the method of predicting the disibution of the α_T jet-balancing kinematic variable. This is undertaken by inverting the cuts on $\Delta\phi$ and $\Delta\eta$ Electron ID variables. Another method utilises Isolation templates in control regions to predict the distribution in the signal region.

Contents

1 Introduction

We present a study of two methods for predicting the QCD background contribution to the search for SUSY using a signature of one (and only one) electron plus jets.

The first method employs the Isolation distribution and its description in terms of two components, one from the combination of hadronic jets and heavy-flavor (c, b) jets and one from photon conversions. The background due to hadronic jets is modeled using a control sample formed using an anti-selection on the $\Delta\phi$ and $\Delta\eta$ matching cuts. The background due to photon conversions is modeled using a control sample formed using explicitly reconstructed conversions. We then demonstrate that the two control samples can be used to predict successfully the backgrounds remaining after a tight electron selection.

The second method utilizes the kinematic α_T jet-balancing method which has been recently developed within CMS, as a generic approach to discover New Physics (most favorably supersymmetry) in the single-lepton plus missing energy channel. It is possible to extend the α_T method from the all-hadronic channels, in order to reduce the QCD background and gain reliable control over severe jet mismeasurements. In this note we use the anti-selection Control Sample to obtain a description of the background due to QCD processes and demonstrate that the estimated value and shape of the background agrees quite well with the actual value and shape from the Monte Carlo.

The methods are then "commissioned" with CMS data in the context of the first 1pb^{-1} of pp collisions at a center-of-mass energy of 7 TeV.

2 Analysis Framework

The coding structure used for this analysis has been developed in CMSSW_3_6_1 releases on top of the Susy-Analysis software package [3], which is itself an extension of the Physics Analysis Toolkit (PAT) [4]. A detailed description of the code can be found here [5]. The PAT provides post-processing of reconstructed event data, in order to eliminate the information and condense the number of physics objects in an event for simplified physics analysis purposes. The framework comprises three layers. The initial layer reprocesses RECO or AOD data with the aim to refine the reconstructed object collections (remove duplicates for example).

The second layer formulates the cleaned data into simple object collections, such as PAT::Jet, PAT::Electron, PAT::Muon etc. which are sorted in uncorrected transverse energy. At this stage, the data are available for use in analysis. A third layer may optionally used as well, providing utilities such as cross cleaning between various object collections. The output of the above PAT processing steps is a ROOT ntuple which is further analysed with private code as described here [6].

3 Monte Carlo and Data samples

The analysis uses QCD Monte Carlo data samples for QCD background processes as well as the W + jets process -for studying signal contamination effects-, produced with the Summer09 simulation production for Physics at 7 TeV [1], with CMS. The Standard Model QCD background processes considered are listed below:

- QCD EM enriched in complete bins exclusive of $\hat{p_T}$ ([20,30],[30,80],[80,170]) were produced with the event generator Pythia6.
- QCD BC$\rightarrow e$ also in complete exclusive bins of $\hat{p_T}$ ([20,30],[30,80],[80,170]) produced by Pythia6.
- QCD Jets in inclusive bin of $\hat{p_T} > 30$ GeV and 80 GeV, produced with the event generator Pythia6. These samples are exceptionally used in the Isolation template method next.
- $W \rightarrow e\nu$ sample simulated by Pythia6.

The numbers of events available in these datasets as well as the equivalent integrated luminosity they correspond to, are detailed in Table 1. The luminosity figures give an indication of the statistics used in the study, although next the resulting plots have been normalized to 1pb^{-1} of integrated luminosity, unless stated otherwise.

The background estimation methods to be presented are finally commissioned using the following Data [7] and Monte Carlo (for direct data-MC comparisons) samples:

Data Set	N events	σ (pb)	Equivalent luminosity (pb^{-1})
QCD BCtoE [20< $\hat{p_T}$ <30]	1100000	108330	10.15
QCD BCtoE [30< $\hat{p_T}$ <80]	1000000	138762	7.21
QCD BCtoE [80< $\hat{p_T}$ <170	1208000	9422.4	128.21
QCD EM Enriched [20< $\hat{p_T}$ <30]	9714886	1719150	5.65
QCD EM Enriched [30< $\hat{p_T}$ <80]	9683936	3498700	2.77
QCD EM Enriched [80< $\hat{p_T}$ <170]	5494911	134088	40.98
QCDJets $\hat{p_T}$ >170	3171950	25470	124.54
W + jets	10034822	17830	415.18

Table 1: *The Monte Carlo datasets used to investigate the Delta ID Inversion method in QCD backgrounds. The available Lumniosity is shown, although plots produced are normalised to 1pb^{-1} for the purpose of understanding the near-term reach of CMS.*

- 7 TeV data:
  - JetMETTau Secondary Dataset (SD)

    `/JetMETTau/Run2010A-PromptReco-v1/RECO`
    `/JetMETTau/Run2010A-PromptReco-v2/RECO`
    `/MinimumBias/Commissioning10-SD_JetMETTau-v9/RECO`
  - EG SD

    `/EG/Run2010A-PromptReco-v1/RECO`
    `/EG/Run2010A-PromptReco-v2/RECO`
    `/MinimumBias/Commissioning10-SD_EG-v9/RECO`
- 7 TeV Monte Carlo:

  `/QCD_Pt-15_7TeV-pythia8/Spring10-START3X_V26B-v1/GEN-SIM-RECO`

4 Event selection

4.1 Pre-selection for single electron analysis

The current analysis follows a series of requirements on the basic physics objects to define events with 1-electron in the final state. The baseline selection is:

- The High Level Trigger (HLT) requirement is HLT_Jet15U, which is currently the lowest threshold unprescaled trigger.
- Exactly one electron of the following definition:
  - Reconstructed with the PixelMatchGsfElectron algorithm.
  - $p_T > 10$ GeV, 20 GeV[1)]
  - $|\eta| < 2.4$
  - Passes the Cut Based ID formed by simple, yet robust, variables (these are the H/E, the super-cluster (SC) - track matching variables $\Delta\phi$,$\Delta\eta$, and shower shape variable $\sigma_{i\eta i\eta}$ and the combined relative Isolation[2)]).
- The event is vetoed if there are any muons of the following definition:
  - $p_T > 15$ GeV
  - $|\eta| < 2.1$

[1)] The Analysis is targeting to commission the background estimation methods for electron P_T threshold at 10 GeV; however, a 20 GeV threshold is always checked to allow synchronization with the Egamma POG recommendations and EWK PAG analyses of CMS.

[2)] The cut based ID selection cuts are chosen to correspond to an 80% efficiency in the $W \rightarrow e\nu$ analysis.

  - Passes ID requirement: GlobalMuon

- Jets are reconstructed with ak5jet algorithm run on standard Calorimeter Jets. The jet selection is as follows:
  - $p_T > 20$ GeV (corrected energy)
  - $|\eta| < 5$
  - EMF <0.9 (in the MC only); a jet can be further rejected if it is found close to a tight and isolated electron within ΔR=0.3 and the ratio of electron to jet energy is $p_T(e)/p_T(jet) > 0.7$.
  - in real data, in order to reject noise from the calorimeters, jets are furthermore required to pass *loose jet ID* (identification) criteria which are summarised as:
    * $|\eta| > 2.6$ or EMF > 0.01
    * fHPD (fraction of energy contributed by the highest hybrid photo-diode readout in the HCAL) < 0.98
    * n90Hits (number of RecHits contributing 90% of the jet energy) > 1

According to the standard recommendations for a global *event cleaning* when running on the Data, the following requirements have been imposed as well:

- L1 technical trigger bit 0 is active: to ensure consistent timing with LHC bunch crossing.
- HLT PhysicsDeclared bit is ON; which indicates that all CMS systems were operational with stable beams in the accelerator.
- At least one good vertex (excluding fake) with number of degrees of freedom, $N_{\text{dof}} >= 5$, and vertex position along the beam direction of $|z_{\text{vtx}}| < 15$ cm.
- Remove events with many fake tracks (also known as monster events) by requiring the ratio of HighPurity tracks over the total number of tracks to be greater than 25% in events that have 10 or more tracks.

4.2 Simple Cut-Based Electron Selection

Efficient electron selection is important in physics analyses with electron final states to enhance the selection of signal. This is especially crucial when the E_T threshold is low, as is characteristic of many SUSY searches, as the background and fake rate increase. In the era of LHC start-up, it is essential to use simple and robust variables.

The electron selection used in this study follows the recommendations of the e-gamma group[8]. Simple cuts are made on a small number of robust variables suitable for early data taking at the LHC. Different cuts are applied to electrons in the ECAL barrel to those in the ECAL endcap. Aside from this no categorisation is applied.

The Electron Selection variables can be described in three groups:

- Typical electron ID variables ($\sigma_{i\eta i\eta}$, $\Delta\phi$, $\Delta\eta$ and H/E).
  - $\sigma_{i\eta,i\eta}$ measures the RMS shower width in the eta direction.
  - $\Delta\phi_{\text{in}}$ and $\Delta\eta_{\text{in}}$ give the geometric match, in ϕ and η respectively between the GSF track trajectory and the ECAL supercluster.
  - Tracker, ECAL and HCAL isolation formed respectively from the sume of ECAL RecHits, HCAL RecHits and track p_{T} in a cone of $\Delta R < 0.4$. The centre of the cone is taken to be the supercluster for the calorimeter isolations and the track direction at the vertex for the tracker.
  - H/E is the ratio of the energy deposited in the HCAL behind the electron seed to the energy of the supercluster.
- Isolation variables (Tracker Isolation, HCAL Isolation and ECAL Isolation)
- Conversion Rejection Tools; photon conversions are rejected initially by requiring that the track associated to the electron has a hit in the first pixel layer pf tracker pixels. Additional rejection power against converted photons is achieved by using the variables described here [9].

We apply this electron selection with values at a working point corresponding to 80% efficiency for the W$\rightarrow e$ analysis. One of the advantages of this new cut based selection is thiat it allows the inversion of one or more of the variables, a common tool in background-subtraction and signal extraction methods.

5 QCD Background Estimation Methods

In the context of a search for SUSY using a signature containing an electron, there are three sources of "background electrons", namely "non-prompt" electrons:

1. Jets which are either mismeasured or have very atypical hadronization, yielding "electrons" which pass the basic identification and selection requirements. We refer to these as "Jet electrons" (Jet-e) in what follows.
2. Photon conversions in the tracker material. We refer to these as "Conversion electrons" (Conv-e) in what follows.
3. Electrons from semileptonic decays of c and $b-$ flavored hadrons. We refer to these as "Heavy Flavor electrons" (HF-e) in what follows.

There are numerous methods for estimating both the amount of background and the shape of this background as a function of various variables (the transverse momentum, pseudorapidity, and in some cases more complicated topological variables) that remains in a data sample after a specific set of selection cuts. All these methods employ "control samples" which are selected via appropriate alternate requirements which do not affect the shape of the variable in question (e.g. the transverse momentum of the lepton). As an example, we mention the well-known "Isolation inversion" method, where one uses events failing the isolation requirement and an extrapolation into the "selection region", i.e. where events pass the isolation requirement, to estimate the amount of background in the latter [10]. A crucial issue in most methods is the demonstration that the extrapolation from the "control region" to the "selection region" is correct or, at least, its deficiencies are small and can be estimated reliably. There are two elements that enter this extrapolation from a control region to a selection region in the case of electrons:

- The actual knowledge of the full dependence of any single source of background electrons on the variable in question. As an example, concentrating on only the Jet-e background, there remains the issue of how to obtain reliably the full shape of the isolation of the Jet-e background which passes all selection cuts from some other "control sample" which passes a slightly different set of cuts.
- The presence of the three different sources of background implies that each of these backgrounds may well exhibit a different dependence on the variable in question. As an example, the isolation distribution from Jet-e and Conv-e need not be (and is in fact not) the same. This implies that it may be necessary to determine explicitly the amount of each background in the selection region, and to then form the total expected background as the sum of the three individual components.

For the sake of concreteness, in what follows we will refer to the "Isolation variable", even though the discussion applies equally well to essentially most variables.

Of the two above issues, the first is usually tackled initially using Monte Carlo: one uses generator-level information to obtain the Isolation distribution for any specific background (e.g. Jet-e) and then compares this shape with that extracted from the "control sample". Assuming that this comparison is favorable, i.e. that the shape from the control sample can adequately describe the shape in the selection sample, the next and final step is to compare this shape with the one extracted from data using the "control selection".

Tackling the second issue is more complicated though. First, the presence of three sources of background (instead of two) implies more degrees of freedom in the fitting (and in general in describing) any particular variable – e.g. the Isolation variable. Second, the relative amount of each source needs to be determined for the final selection sample.

In this section, we concentrate on the following three key observations to address all of the above:

1. A good method for identifying a control sample for the Jet-e background is the reversal of the matching cuts $\Delta\eta$ and $\Delta\phi$. These two variables exhibit near independence from other selection variables, especially the isolation variable
2. The isolation distributions of the Jet-e and HF-e backgrounds seem to be similar and to be collectively described well by the same control sample (via the matching-cut reversal).
3. A fairly pure sample of Conv-e background can be identified using the existing conversion-identification tools. This sample can then be used to extract a template for other variables in question, e.g. the Isolation distribution or the α_T distribution.

In what follows, we exploit these three observations to obtain the shapes of the total background in both the isolation and α_T distributions. Moreover, a fit to the isolation distribution using different shapes (templates) for the Jet-e and Conv-e backgrounds yields an estimate of the absolute number of background events in the selection region.

5.1 Modeling the electron backgrounds

In what follows, we will consider more closely two different definitions of "Isolation", namely "Calo Isolation", which is defined using only the sum of the energies in the ECAL and HCAL, and "Combined Isolation" which includes also the transverse momenta of charged-particles tracks around the electron. Since the search for SUSY may necessitate using low-P_T leptons, we investigate the behavior for two different thresholds of 10 and 20 GeV on the electron P_T.

The isolation distribution using only calorimeter isolation is displayed in Figure 1, whereas the combined (relative) isolation is shown in Figure 2. It can be seen that two backgrounds, namely the Jet-e and JF-e have distributions which are fairly similar for both the Calo-Iso and Comb-Iso.

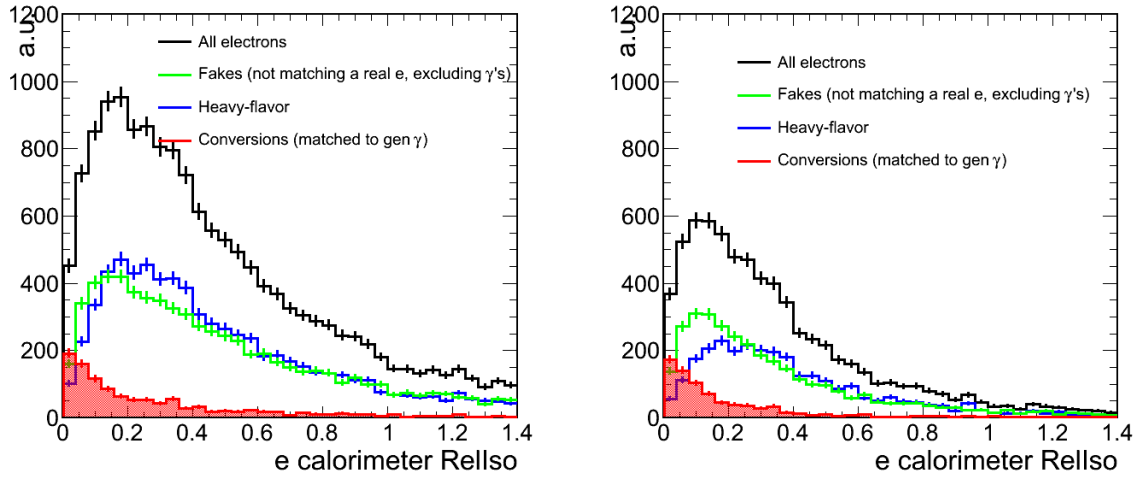

Figure 1: *The Calorimeter Isolation distribution from Monte Carlo simulation (QCD pythia $\hat{p}_T > 80$). On the left for a threshold of 10 GeV on the electron and the right for a 20 GeV threshold. The three sources of background, namely Jet-e, HF-e and Conv-e are shown separately, along with the sum of the three.*

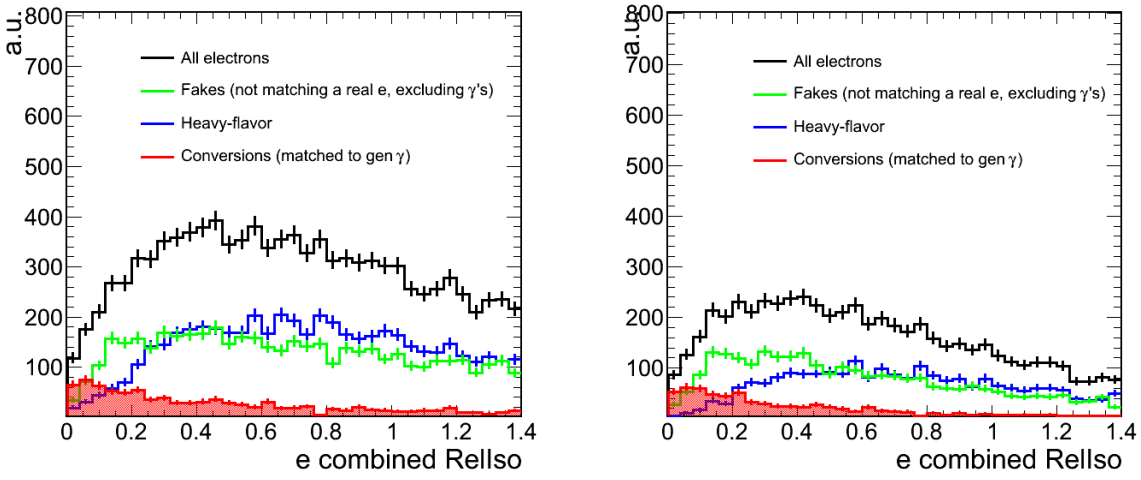

Figure 2: *The Combined Isolation distribution from Monte Carlo simulation (QCD pythia $\hat{p}_T > 80$). On the left for a threshold of 10 GeV on the electron and the right for a 20 GeV threshold. The three sources of background, namely Jet-e, HF-e and Conv-e are shown separately, along with the sum of the three.*

5.1.1 Isolation Distribution for the Jet-e and HF-e backgrounds

The first step is to define a control sample for the combination of Jet-e and HF-e and to see how well it can describe the Isolation distributions (both CaloIso and CombIso) in the selection region. As stated previously, this is done by inverting the $\Delta\eta$(trk-SC) and $\Delta\phi$ (trk-SC) id cuts in the electron selection. The selected events in this method pass the pre-selection described in Section 4, while the anti-selected are those events which pass the selection with an electron that passes all selection criteria *except* the $\Delta\phi$ (trk-SC) and $\Delta\eta$(trk-SC) ones. The resulting distributions are shown in Figure 3 for the CaloIso and in Figure 4 for the CombIso.

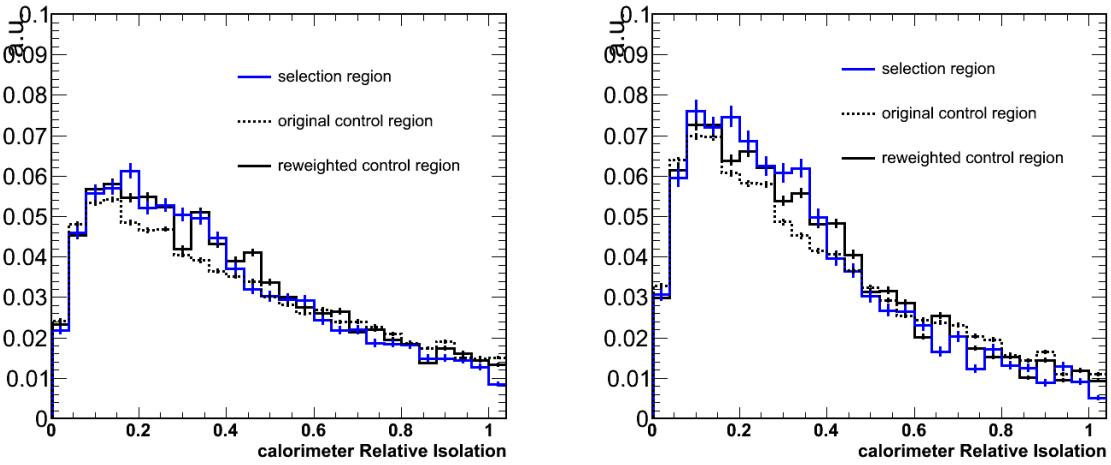

Figure 3: *The Calorimeter Isolation distribution from Monte Carlo simulation of the combined Jet-e+HF-e background. On the left for a threshold of 10 GeV on the electron and the right for a 20 GeV threshold. The solid blue line is the total Jet-e+HF-e background in the selection region, whereas the dashed line is the distribution from the control sample, defined via the anti-selection on the matching cuts. The solid black line is the result of re-weighting the control sample for the jet spectra – as described in the text.*

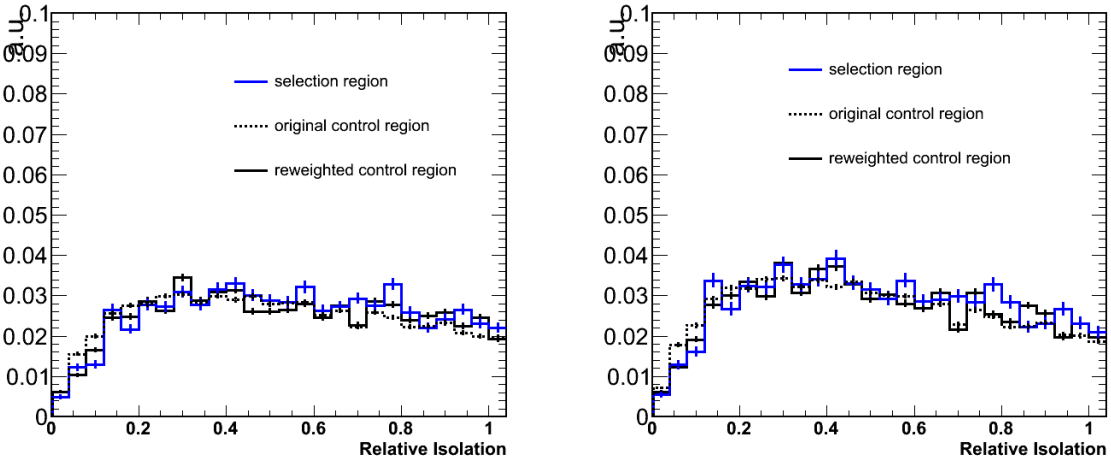

Figure 4: *Same as Figure 3 only this time for the Combined Isolation distribution.*

It can be seen that the shape of the Combined Isolation distribution from the Control Region is quite similar to that of the distribution from the Selection Region – a very encouraging result which indicates that the anti-selection of the matching cuts yields a good method for modeling non-conversion electrons. A closer look into various properties of these events, however, yields a slight difference in the Calorimeter Isolation distributions. As can be seen in Figure 3, for small values of the CaloIso variable the two distributions, i.e. from the Control and Selection regions have a small difference (around CaloIso$\approx 0.1 - 0.4$). This is more easily seen in Figure 5 where the ratio of the Calorimeter Isolation distributions from the Selection and Control regions is displayed.

We have investigated possible sources of this difference, from the surrounding objects, and in particular any jet that (possibly) accompanies the electron candidate. There are small differences in the shapes of the associated jet

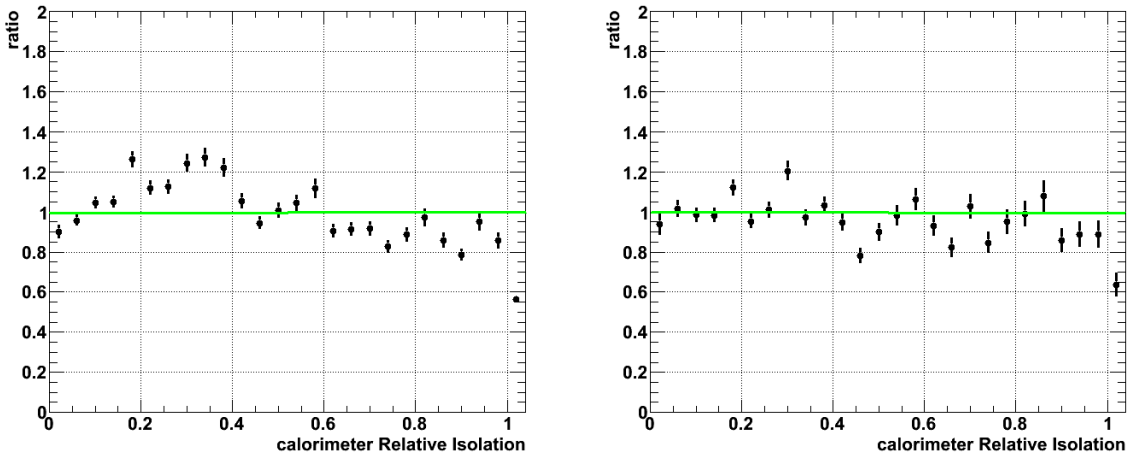

Figure 5: *The ratio of the Calorimeter Isolation distributions from the Selection Region and the Control Region. Left plot: the ratio of the two "raw" distributions (the dashed curves) seen in Figure 3. Right plot: after the re-weighting correction, i.e. the ratio of the two black lines in Figure 3.*

P_T spectrum as well as from the distance in $\eta - \phi$ space (ΔR) between the jet and the electron. These can be seen in Figure 6.

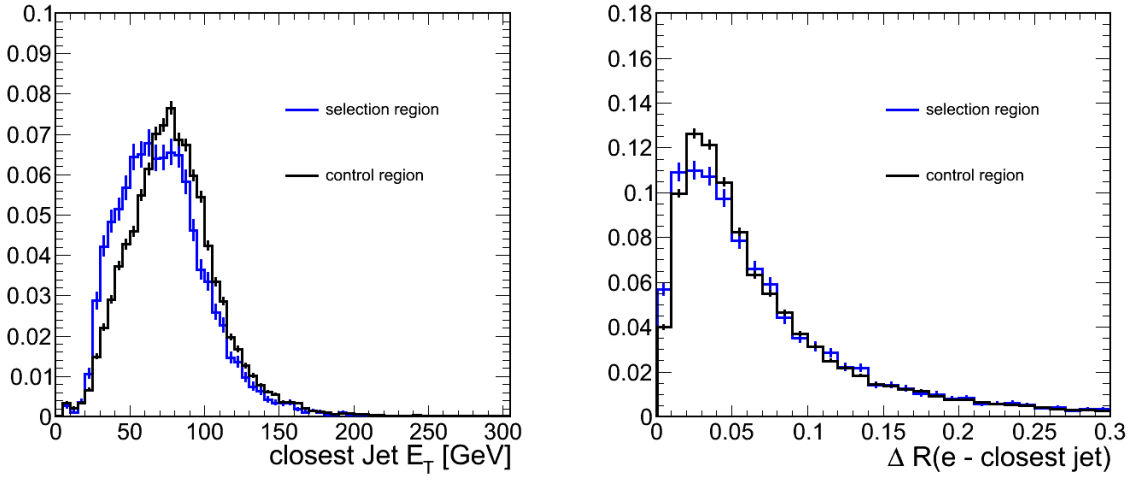

Figure 6: *Distributions of the transverse momentum (P_T) and distance to the electron ΔR) of the nearest jet.*

Presumably, these differences arise from a small correlation between the matching variables and the density of the overall hadronic energy surrounding or near the electron candidate. Since the difference is small, we have attempted to correct the "predicted" shape, i.e. the shape from the Control Region, by applying a weight which depends on the associated Jet P_T and the ΔR between the jet and the electron. The solid blue line in 3 is the result of this re-weighting. It can be seen that the corrected Calorimeter Isolation distribution from the Control Region now describes the distribution from the Selection Region quite well. This is more readily seen in Figure 5 (right).

We have also applied the same correction to the Combined Isolation distribution, which did not exhibit a visible difference between the Control and Selection Regions, to ensure that the correction did not adversely affect this original agreement. As can be seen in Figure 4 the corrected distribution still describes the Selection Region very well.

5.1.2 Isolation distribution from the Conv-e background

Electrons from photon conversions are identified using the standard criteria of the Conversion Finder tools [9]. A suitable Control sample for modeling the Conv-e component in the Isolation distribution can be obtained by electrons that pass the Conversion tools (i.e. using an anti-veto on the Conversion rejection requirements). There are two algorithms to identify electrons from converted photons:

- *Missing expected hits*: the algorithm asks that there be > 0 expected layers with a missing hit before the first valid hit on the electron's track. The number of missing expected hits in front of the innermost valid hit is available via the electron's gsfTrack Hit Pattern.
- *Partner track finding*: the algorithm looks for the electron's partner track from a converted photon in the generalTrack collection. The track is identified as a conversion partner if:
  - the track has opposite charge to the electron track.
  - Approximately the same $\delta \cot(\theta)$, in this case: $|\delta \cot(\theta)| < 0.02$.
  - small distance (dist) in the $r - \phi$ plane, in this case: $|\text{dist}| < 0.02$.

The Conv-e component of the Isolation in the Selection region is formed using only electrons which match a generated photon at MC generator ("MC-truth") level. The shape of this component is compared with the one obtained from the Conv-e control region as described previously. The resulting distributions are shown in Figure 7 for the CaloIso and in Figure 8 for the CombIso. It can be seen that the shape of both distributions in the Selection Region is described well by the Control Region.

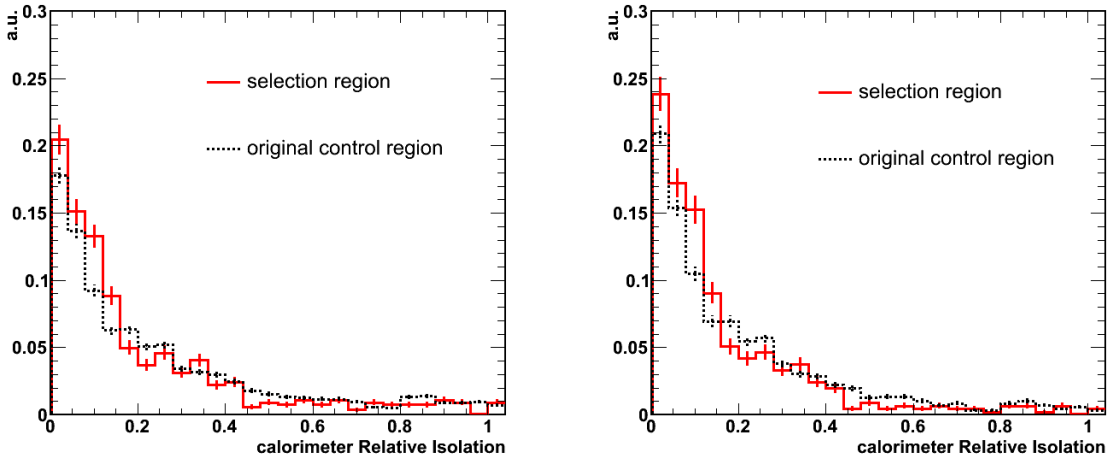

Figure 7: *The Calorimeter Isolation distribution from Monte Carlo simulation of electrons from conversions. On the left for a threshold of 10 GeV on the electron and the right for a 20 GeV threshold. The solid red line is the Conv-e background in the selection region, whereas the dashed line is the distribution from the control sample, defined via the conversion identification requirements described in the text.*

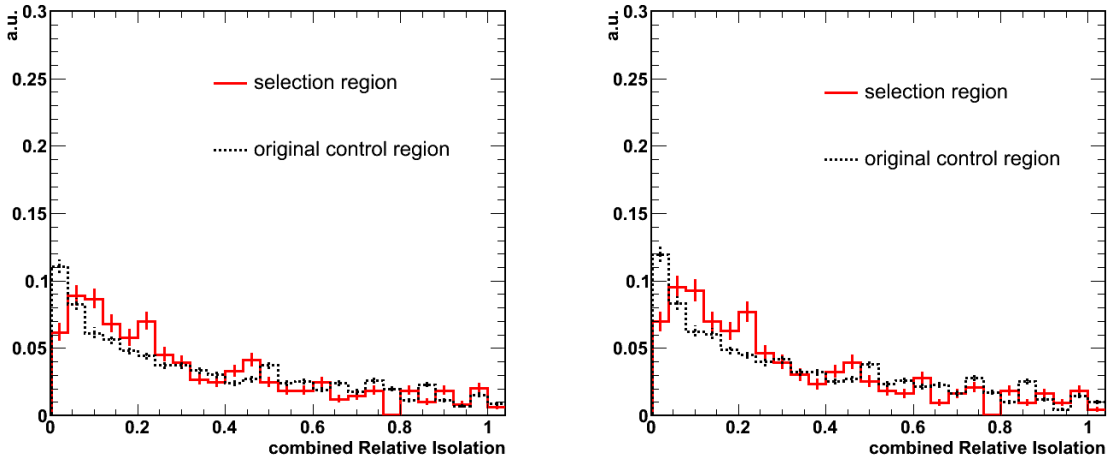

Figure 8: *Same as Figure 7 only this time for the Combined Isolation distribution.*

5.1.3 Describing the Isolation distribution in the Selection Region

Having demonstrated that two independent Control Samples can yield good descriptions of the Isolation distribution in the Selection Region for each background (i.e the combined Jet-e+HF-e and the Conv-e) we next attempt to describe the full Isolation distribution in the Selection Region as a sum of two components, with the template of each component extracted as described above: the combined Jet-e and HF-e background is described from the corrected (re-weighted) anti-selected (for matching) sample, whereas the Conv-e background is described from the sample passing the conversion-identification criteria. We thus fit the total Isolation distribution using these two components, leaving the relative normalization of the two as a free fit parameter. The result is shown in Figure 9 for the calorimeter isolation and shows a very good description of the total background distribution. It can also be observed that the conversion component becomes more relevant at high $P_T(e)$. The corresponding distributions for the Combined Isolation variable are described equally well (see Figure 10).

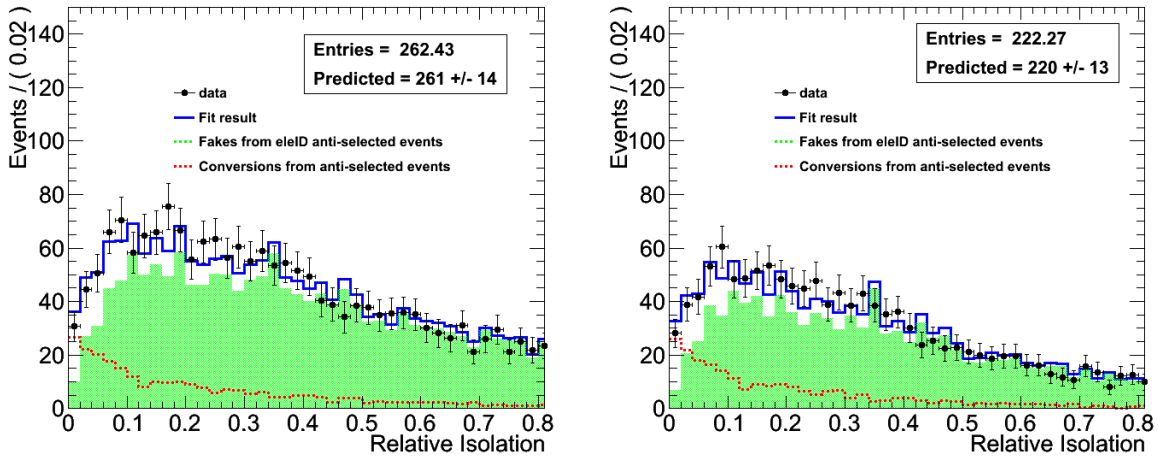

Figure 9: *The Calorimeter Isolation distribution from Monte Carlo simulation of background electrons. On the left for a threshold of 10 GeV on the electron and the right for a 20 GeV threshold. The dashed lines are the two background components as extracted from the two Control Samples, whereas the sum of the "predicted" background is the solid line.*

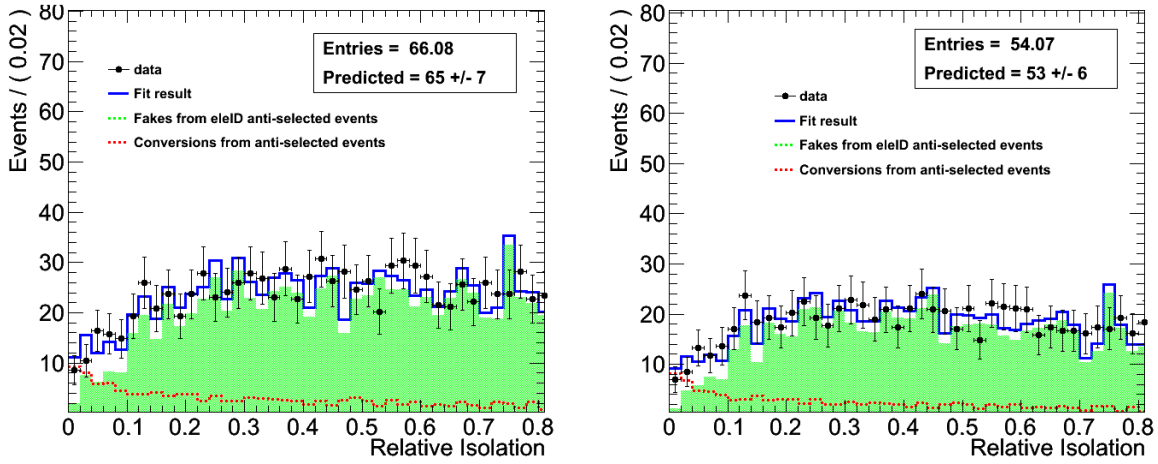

Figure 10: *The combined Isolation distribution from Monte Carlo simulation of background electrons. On the left for a threshold of 10 GeV on the electron and the right for a 20 GeV threshold. The dashed lines are the two background components as extracted from the two Control Samples, whereas the sum of the "predicted" background is the solid line.*

5.1.4 Describing the Isolation distribution in the presence of prompt electrons

In the data, the Isolation distribution in the Selection region will also be populated by prompt electrons sources -in addition to QCD sources - with $W \rightarrow e\nu$ being the most prominent one (and electrons from semi-leptonic $t\bar{t}$

events). We have thus investigated the performance of the Isolation template method in the presence of W events. In this case, one needs to add to the fit function a third component (template) to describe the Isolation shape for prompt electrons. We note here that an Isolation template for prompt electrons can be easily extracted using data-driven ways (e.g. the random cones technique) or even using a MC shape directly.

The selection of electrons with offline $P_T > 10$ GeV implies that one needs to incorporate the low $\hat{p}_T$-bins of QCD MC samples. We therefore include both the QCD inclusive jets samples QCD_pt30 and QCD_pt80, with a weight normalized to 0.1pb^{-1} of integrated luminosity[3)]. The W pythia sample is also normalized accordingly and included in the Selection. We finally repeat the Isolation template method, using the combined relative Isolation distribution, as previously described, in order to extract the number of fake (background) electron events that fall into the Signal Selection region (RelIso < 0.1). Figure 11 (left) shows the combined fit for $P_T(e) > 10$ GeV in the presence of a W signal. The same fit is repeated on the right plot, where now a pfMET < 20 GeV cut has been applied to the Selection, to suppress the $W \rightarrow e\nu$ component. The latter case descreases the error on the measurement due to smaller correlations in the fit parameters.

Armed with this result, we next investigate the performance of the template method on fake electron events (in signal region RelIso < 0.1) in the presence of W events, and with increasing a cut on the hadronic activity in the event -namely applying successively an H_T cut. Figure 12 shows a comparison of the number of background events observed in the signal region, in black, versus the number of fake electron events predicted in blue. Also superimposed are the number of total events, - including Ws -, in the signal region shown in red dots. A cut of pfMET < 20 GeV has been applied. Similar plots are repeated for electrons with $P_T > 20$ GeV (see Figures 13 and 14).

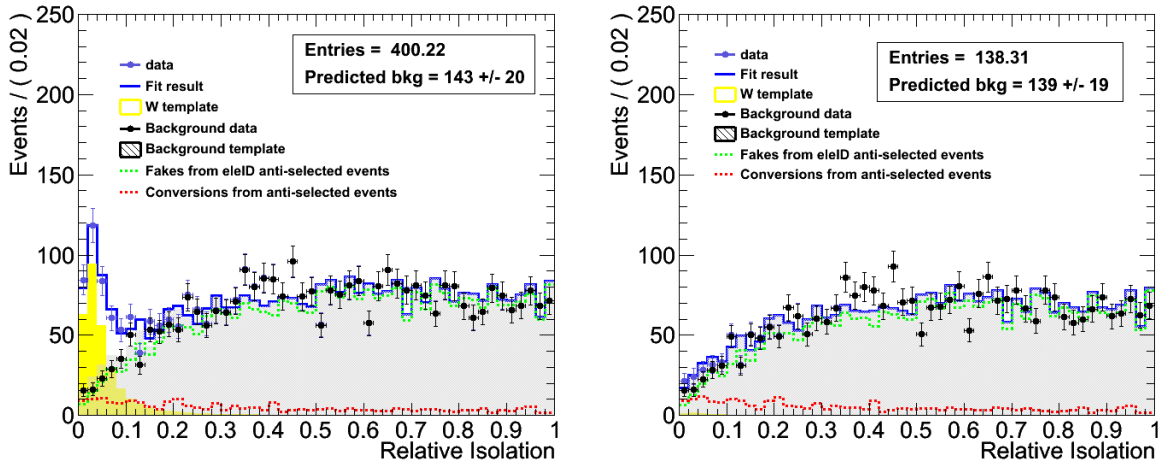

Figure 11: *The combined Isolation distribution from Monte Carlo simulation of background electrons - using the PYTHIA QCD_pt30 and QCD_pt80 samples -, and prompt electrons - using the PYTHIA $W \rightarrow e\nu$ sample, for $p_T(e) > 10$ GeV. On the right plot, a pfMET < 20 GeV anti-cut has been applied. The dashed lines are the two background components as extracted from the two Control Samples, whereas the total "predicted" background is the solid line.*

Given the remarkably good description of the combined background in Monte Carlo simulation, we next test this procedure on CMS data. This is the subject of the next section.

5.2 Predicting the distribution of the α_T kinematic variable by inverting Electron ID Cuts

Following the promising results of the α_T jet-balancing method previously described for the all-hadronic SUSY searches [12], a natural extension of this approach has been developed to the single-lepton SUSY search [13], where a significant presence of QCD multi-jet backgrounds is expected.

The α_T variable is here defined as an N-object system where the set of objects is 1 electron and N-1 jets. This definition reproduces the kinematics of a di-jet system by contructing two pseudo-jets, which balance one another

[3)] The luminosity chosen to normalize the QCD MC samples corresponds roughly to the available statistics of the QCD_pt30 sample; so that QCD_pt30 events are worked with a weight of ≈ 1

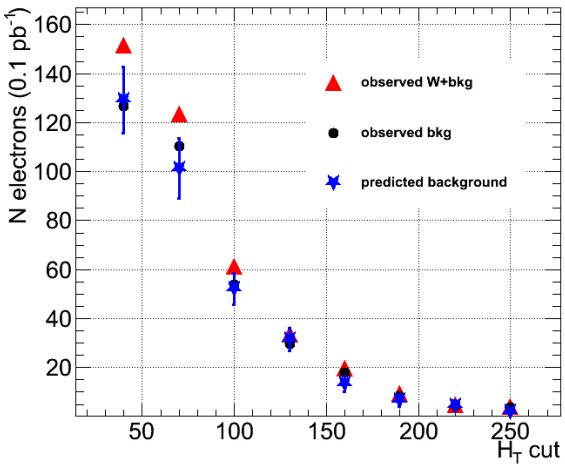

Figure 12: *The number of truth background electrons (in black dots) in signal region, RelIso* < 0.1*, are compared to the fit prediction (in blue stars), as a function of the cut in the hadronic activity of the event (*H_T *cut). A cut of pfMET* < 20 *GeV has been applied to suppress sources of prompt electrons (Ws here). The number of total truth electron events - including residual Ws - is also shown superimposed in red triangles.*

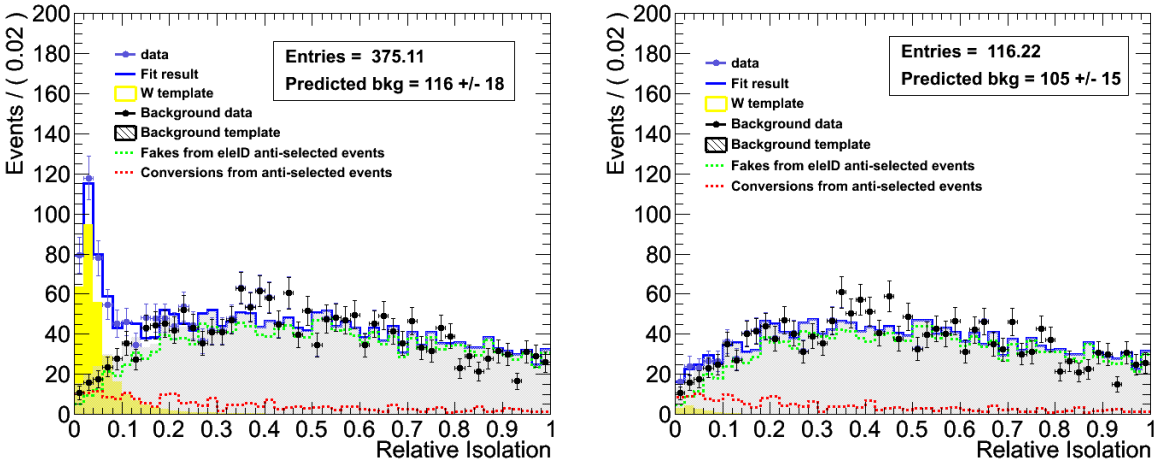

Figure 13: *Same as Figure 11 only this time for electrons with* P_T *threshold at 20 GeV.*

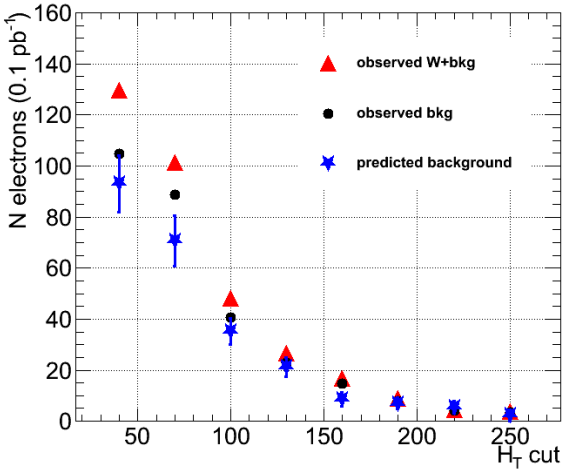

Figure 14: *Same as Figure 12 only this time for electrons with* $P_T > 20$ *GeV.*

in H_T. The two pseudo-jets are formed from the combination of the N objects that minimizes the $\Delta H_T \equiv |H_{T,1} - H_{T,2}|$ of the pseudo-jets, and the resulting α_T is

$$\alpha_T = \frac{1}{2}\frac{H_T - \Delta H_T}{M_T} = \frac{1}{2}\frac{H_T - \Delta H_T}{\sqrt{H_T^2 - MH_T^2}}. \quad (1)$$

This section is dedicated to a first approach of commissioning the alphaT observable and study its behavior in pure fake electron events. It is therefore desirable to collect a suitable control sample which will be dominated by fake electrons and eliminate sources of prompt electron events (like W events).

One way to obtain such a sample is using the anti-selection method on electron ID variables which are less correlated with the missing transverse energy. In this section, we investigate the possibility of inverting the $\Delta\eta$(trk-SC) and $\Delta\phi$ (trk-SC) id cuts in the electron selection. The selected events in this method pass the pre-selection described in Section 4, while the anti-selected are those events which pass the selection with an electron that passes all selection id criteria except the $\Delta\phi$ (trk-SC) and $\Delta\eta$(trk-SC) ones.

In order to establish the validity of the control sample obtained by the anti-selection method above, the perfomance must be compared of the leptonic α_T as obtained from the control sample and the actual QCD events passing the electron criteria defined in the "signal" region. Because SUSY events are expected to have high H_T, it is desirable to understand how the method evolves with increasing H_T cuts. The number of events for 1pb^{-1} passing the pre-selection, and two different cuts in H_T are shown for selected in Table 2 and for anti-selected in Table 3.

Cutflow	QCD EM enriched	QCD BC$\rightarrow e$	QCDJets Pythia $\hat{p_T}$	W	% contamination from W
All events	5351938	256514.4	25470	24170	0.43%
N(e^-) ≥ 1	9372.1	2635.5	19.9	3848.6	32%
N($jets$) ≥ 1	8174.5	2317.3	10.4	3632.48	34.6%
HT > 100 GeV	249.2	55.8	6.2	95.3	30.6%
HT > 180 GeV	7.53	1.2	3.6	2.86	23.2%

Table 2: *Cutflow for selected events.Numbers shown for 1pb^{-1} for both QCD and W samples used for the anaylsis with electron p_T requirement set to 20GeV.*

Cutflow	QCD EM enriched	QCD BC$\rightarrow e$	QCDJets Pythia $\hat{p_T}$	W	% contamination from W
All events	5351938	256514.4	25470	24170	0.43%
N(e^-) ≥ 1	32725.5	729.0	52.9	298.0	0.89%
N($jets$) ≥ 1	27545.3	609.9	28.1	271.5	0.96%
HT > 100 GeV	902.8	24.9	18.3	4.9	0.52%
HT > 180 GeV	20.7	0.8	10.7	0.3	1.01%

Table 3: *Cutflow for anti-selected events.Numbers shown for 1pb^{-1} for both QCD and W samples used for the anaylsis with electron p_T requirement set to 20GeV.*

It can be seen that the anti-selected control sample allows one to study and validate the expected behavior of the α_T in a sample of *pure* fake electron events, where the contamination from prompt electrons should be at a level below 1%, as opposed to the selection region where $W \rightarrow e\nu$ is expected to contribute at the $\approx 10\%$ level.

5.2.1 Closure test with pure QCD background sample

The control sample provided by anti-selection on the ($\Delta\eta/\Delta\phi$) electron id variables must first undergo a closure test with a pure QCD sample. This will show if there is any bias in QCD between the distribution of α_T in the selected and the anti-selected regions. The plots in 15 show the normalised shape of distributions, firstly before an H_T cut (top left) and then as increasing cuts in H_T are applied. The selected and anti-selected distributions show good agreement. The evolution of the α_T as H_T cut increases shows the expected reduction in the tail for $\alpha_T >$ 0.55, the region of likely SUSY signal, for both selected and anti-selected events.

In order to demonstrate the power of H_T in α_T tail-reduction, we introduce the variable R_{α_T} which is defined as the ratio of the number of events passing the α_T cut over the number of events failing it:

$$R_{\alpha T} = \frac{N(\alpha_T > 0.55)}{N(\alpha_T > 0.)} \quad (2)$$

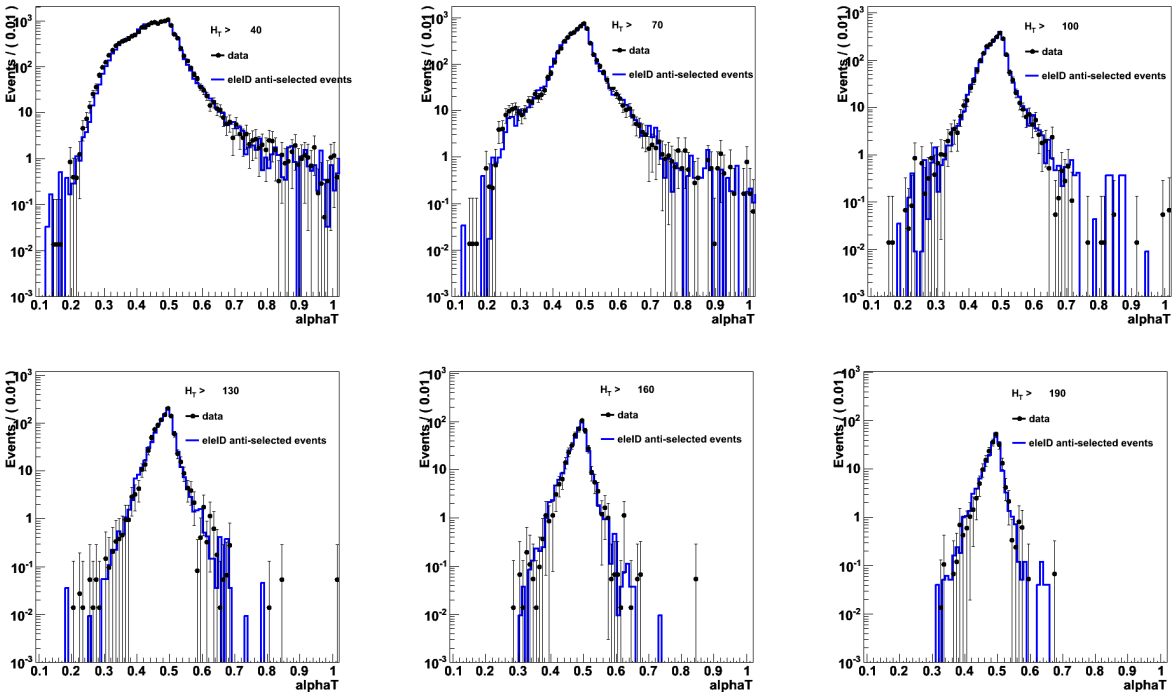

Figure 15: *The α_T distributions for selected (red) and anti-selected events (black) for the QCD multi-jet background, from inversion of the $\Delta\phi$ and $\Delta\eta$ ID Cuts, shown without H_T cut (Top Left) and with progressive HT cuts (left-right, top-bottom). These distributions are normalised to unity for shape comparison. There is good agreement between the selected and anti-selected samples regardless of HT requirement, and the high α_T tails reduce as expected when moving to higher HT cuts.*

The "default" cut value here is the value prompted from the all-hadronic analysis, 0.55. Figure 16 shows a plot of R_{α_T} as a function of the H_T cut applied. As the H_T cut value increases, R_{α_T} is observed to decrease in an (approximately) exponential manner [14]. This result confirms that the noticable reduction in the tail are much more pronounced than those in the peak, therefore this is not a statistical effect only. The selected and anti-selected events remain in good agreement.

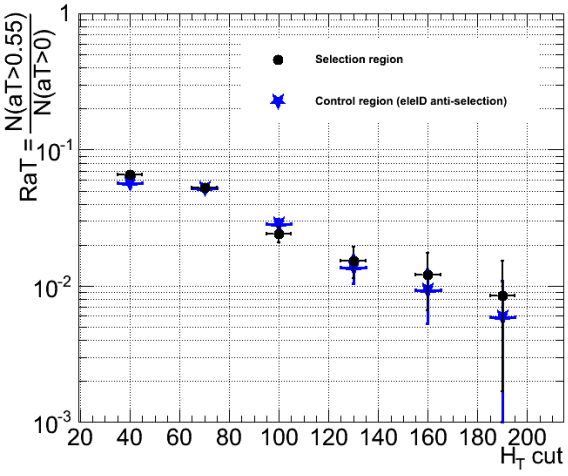

Figure 16: *The R_{α_T} versus the H_T cut applied for the QCD multi-jet background, shown for both selected and anti-selected events in the Delta ID Inversion method.*

5.2.2 Closure test with W + jets contamination in control region

Having acertained the validity of the anti-selection method to predict the QCD contribution in the selected from the QCD in the anti-selected, it is important to test whether the process will work with contamination in the control region. The method is designed for data-driven extimation, and thus must be robust to such contamination.

The closure test is repeated, with the anti-selected now from both the pure QCD sample as before, and W + jets also. The selected remains from pure QCD as a comparison. The normalised ditribution plots shown for this case are in Figure 17, and the plot of R_{α_T} as a function of the H_T cut applied is in Figure 18.

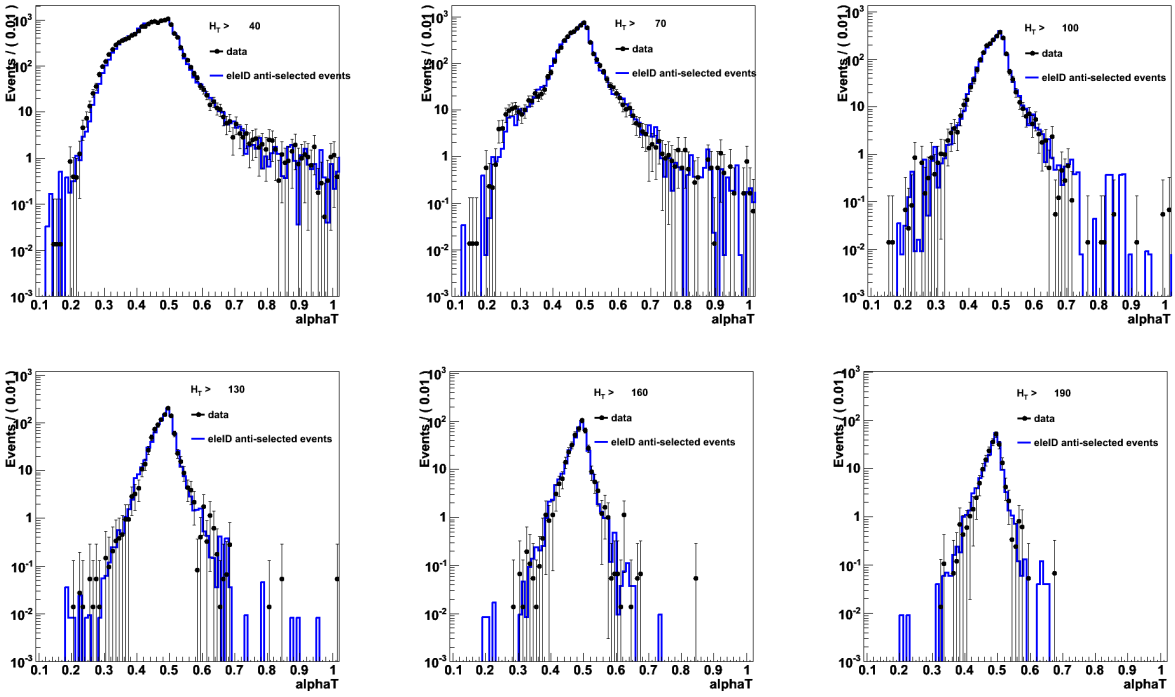

Figure 17: *Same as in Figure 15 only this time with W contamination in the control region.*

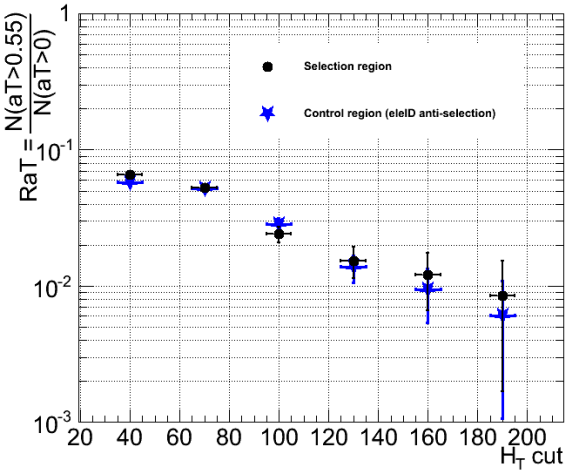

Figure 18: *Same as in Figure 16 only this time with W contamination in the control region.*

Adding in contamination from the W + jets sample has no discernable effect on the shape of the α_T distribution, and the ratio remains to good agreement also. The control sample is thus robust to such a contamination, and still remains a good estimator of the shape of the QCD background for the selected events.

6 Commissioning Background Estimation Methods with 7TeV Data

Having shown that the two background estimation methods work well in Monte Carlo simulation, we next take a look at the results using collision data, and, for comparison a Monte Carlo sample generated by PYTHIA8 to be non process-specific.

It is intentional for this analysis to commission low-p_T electrons (of offline thresholds below the standard online threshold of the single electron trigger); and therefore the data sample is chosen from the JetMETTau secondary datasets, which are subject to JetMET triggers. The data sample used in this study is taken from the Secondary Dataset (SD) JetMETTau, and the HLT trigger applied on top of the selection is the HLT_Jet15U.

The 7TeV collision data used to demostrate the performance of the Background Estimation methods described in this Note, amounts to 12.47 nb^{-1} of integrated luminosity.

6.1 Commissioning the Isolation Template method with first data

The data are selected as described in section 4. Events with at least one electron with $p_T > 10$ GeV and passing all standard electron identification, as well with a minimal hadronic activity of $H_T > 20$ GeV - equivalent to the requirement of at least one jet with $p_T > 20$ GeV -, are used to plot the combined Isolation distribution in the Selection region. A three-template fit is used to model the Isolation distribution in the Selection region, with two-template background component taken from the two control regions which defined earlier, and a single W-component to describe the $W \rightarrow e\nu$ shape. The W-shape is used here directly from a MC template ($We\nu$ PYTHIA sample) in order to test the Isolation template method.

Figures 19 show the full fit to the Isolation distribution in the Selection region, with the right plot repeating the exercise with a $pfMET < 20$ GeV cut. In both cases there is a good agreement between the Isolation shape selected and the modeling distributions extracted from the control regions. A first evidence of W-signal contamination is apparent on the plots; with the left one showing an excess of events from the predicted background, and the right one showing the suppression of the signal using the pfMET "anti-cut".

Since, as was shown in section 5.1, we observed a small deviation between the Calorimeter Isolation distributions in the Control and Selection Regions (prior to re-weighting as described above) we would like to confirm that the good description of the full Isolation distribution, i.e. with all backgrounds combined, remains good for different regions in the overall hadronic selection of the events. For this reason, we repeat the fits for different values of the total H_T of the event. The fits yield a number of expected events passing the RelIso< 0.3 requirement – and this number can be compared to the actual number seen in the data. We summarize these results in Figure 20 where a very good agreement between the predicted and observed event yields is seen.

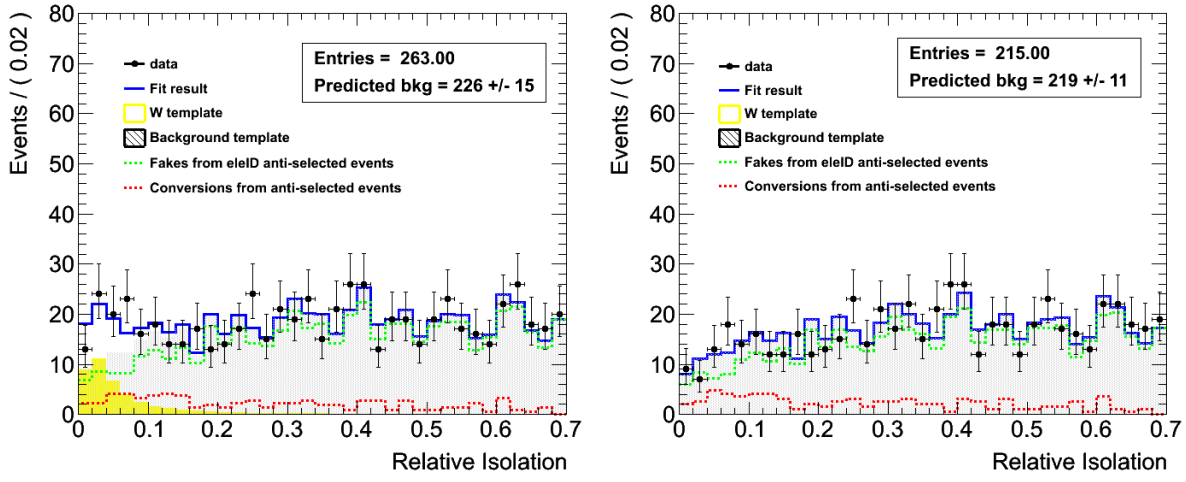

Figure 19: *The combined Isolation distribution for 10 GeV electrons in data (points) and its breakdown into three components, one for the combined background from jets and heavy-flavors (Jet-e and HF-e), another for conversions (Conv-e) and a third for the $W \rightarrow e\nu$. The dashed lines are the two background components as extracted from the two Control Samples, whereas the sum of the "predicted" background is shown in filled grey. On the right plot, a pfMET < 20 GeV anti-cut has been applied.*

The same procedure is applied to events with electrons with $p_T > 20$ GeV, and the result can be in inspected

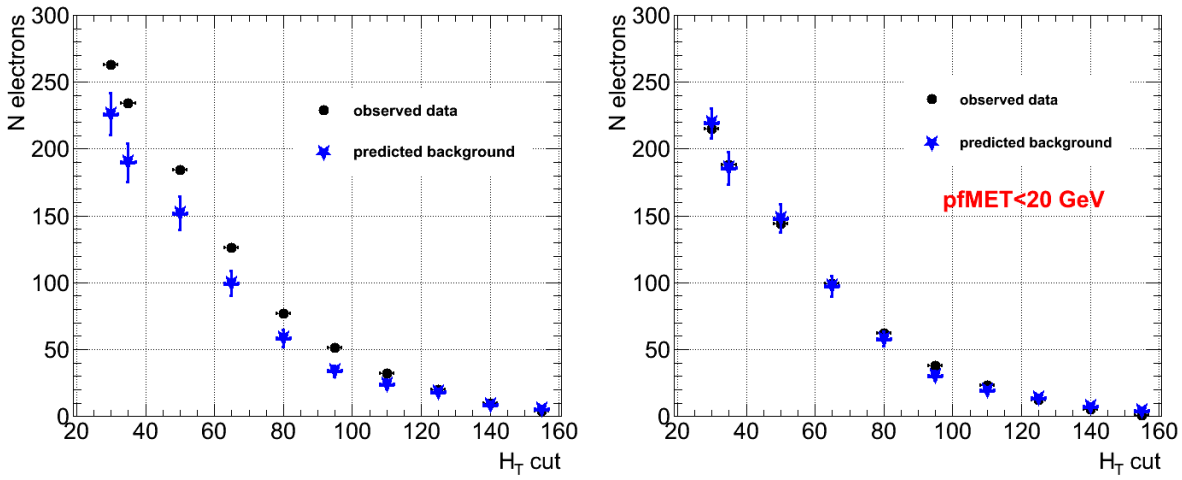

Figure 20: *The observed number of electrons (in black dots) in signal region, RelIso* < 0.3*, are compared to the fit prediction (in blue stars), as a function of the cut in the hadronic activity of the event (*H_T *cut). A pfMET anti-cut at 20 GeV has been applied to suppress sources of prompt electrons (e.g. Ws).*

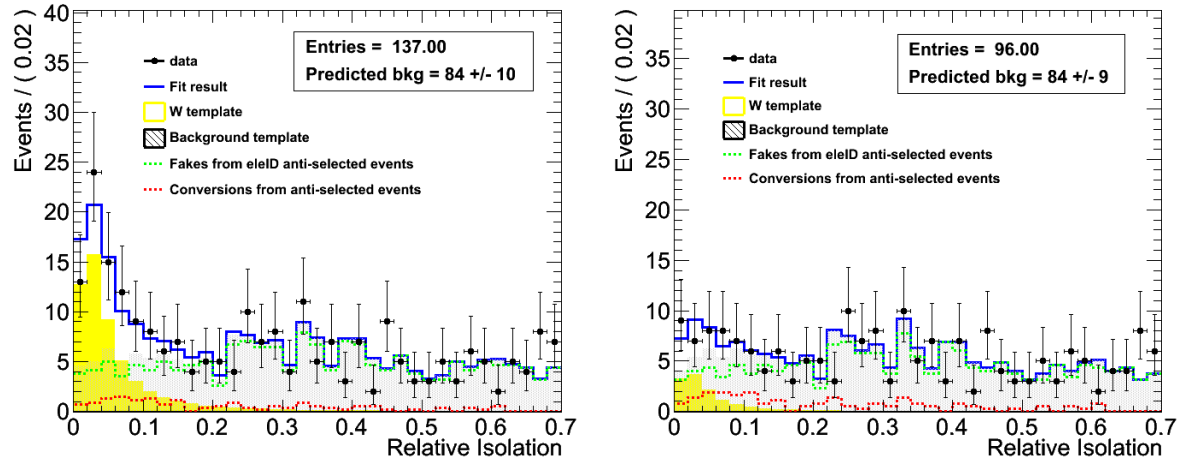

Figure 21: *Same as Figure 19 only this time for electrons with* P_T *threshold at 20 GeV.*

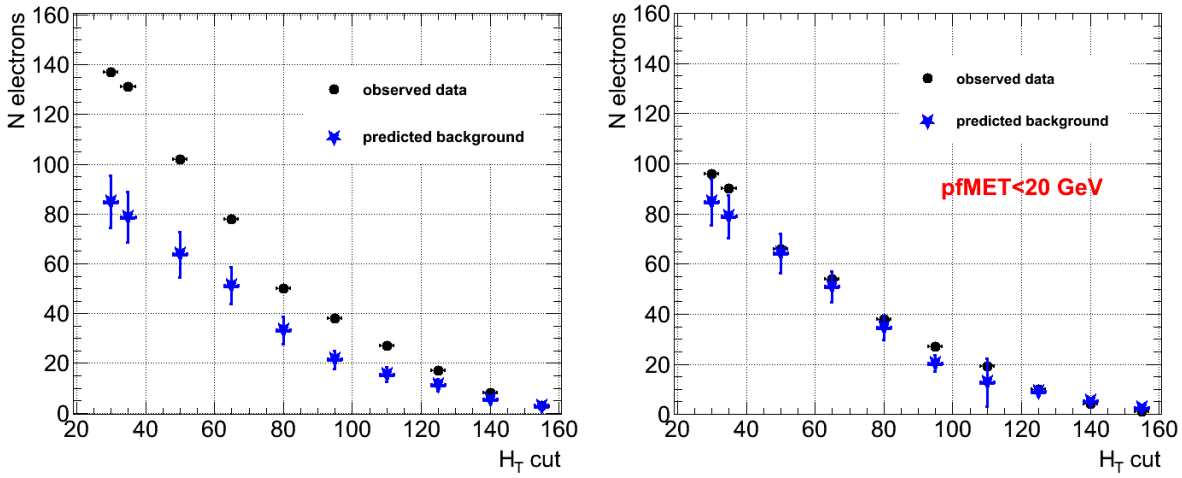

Figure 22: *Same as Figure 20 only this time for electrons with* P_T *threshold at 20 GeV.*

similarly on Figures 21 and 22. The evidence of the W component present in the selection is now more prominent with the level of background being significantly lower with respect to the $p_T(e) > 10$ GeV case. The overall

results show a compatible behavior with the corresponding MC results presented in section 5.1.4.

The results obtained using the data-driven control samples are quite encouraging that these samples may also be used to obtain the shapes observed in data for more intricate topological variables. This is the subject of the following section.

6.2 Commissioning the cut-based electron ID Inversion method with first data

Due to the limited statistics collected so far, currently the selection is loosened in the following ways:

- The electron offline p_T threshold is lowered to 10 GeV.
- The electronID is applied according to the official recommendations of the CMS Egamma POG. The electron isolation is chosen to be the Relative Calorimeter Isolation, with the value of the cut loosened to 0.3.
- At least one jet with $p_T > 40$ GeV corrected transverse energy.
- The H_T cuts applied are lowered to study the evolution without losing all valid statistics.

Because SUSY events are expected to appear in high H_T regimes, we examine the behaviour of the α_T distribution as a function of H_T. Figure 23 show the α_T distributions for selected and anti-selected events from collision data, in succesive cuts in H_T, whereas a $pfMET < 20$ GeV cut has been applied. The selected and anti-selected distributions show a good agreement. However, the selected distributions are susceptible to W contamination at higher level than the anti-selected ones.

To measure the reduction in high aT events, we plot the ratio $RaT = N(aT > 0.55)/N(aT > 0)$ as a function of the H_T cut in Fig. 24. RaT does show an approximately exponential decrease with H_T, a performance which can be reliably validated using the anti-selected events.

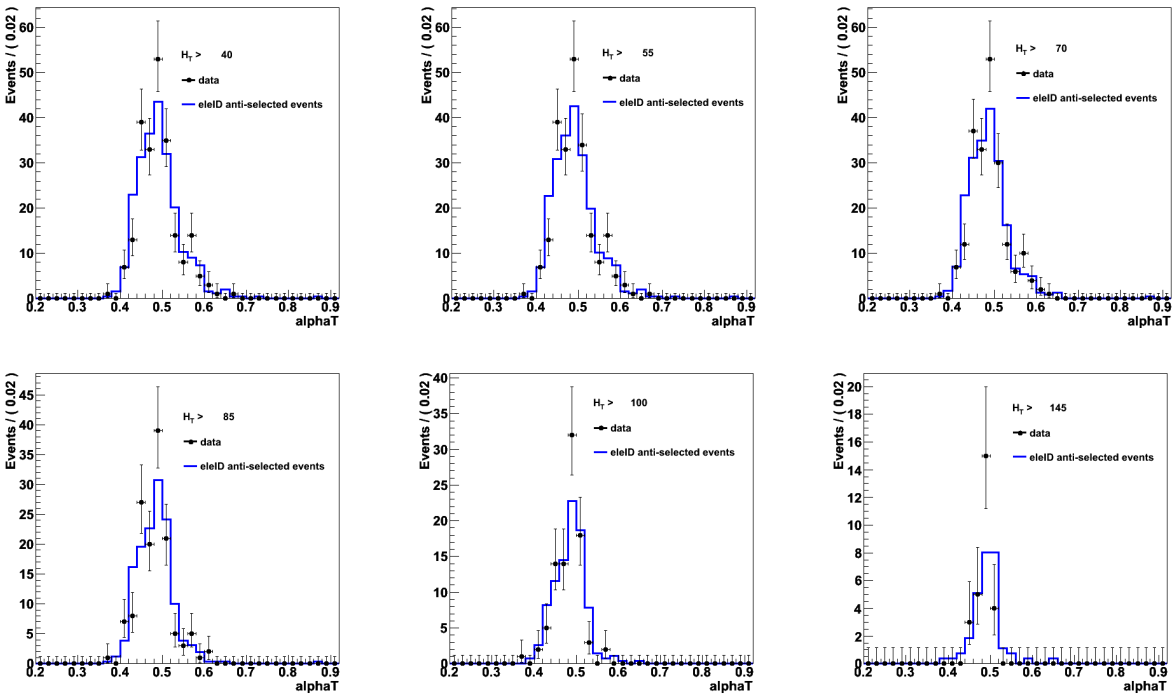

Figure 23: *The α_T distributions for selected (red) and anti-selected events (black) from collision data shown with progressive cuts in H_T. These distributions are normalised to each other for shape comparison. There is good agreement between the selected and anti-selected samples regardless of H_T requirement, and the high α_T tails are reduced as expected when moving to higher H_T cuts.*

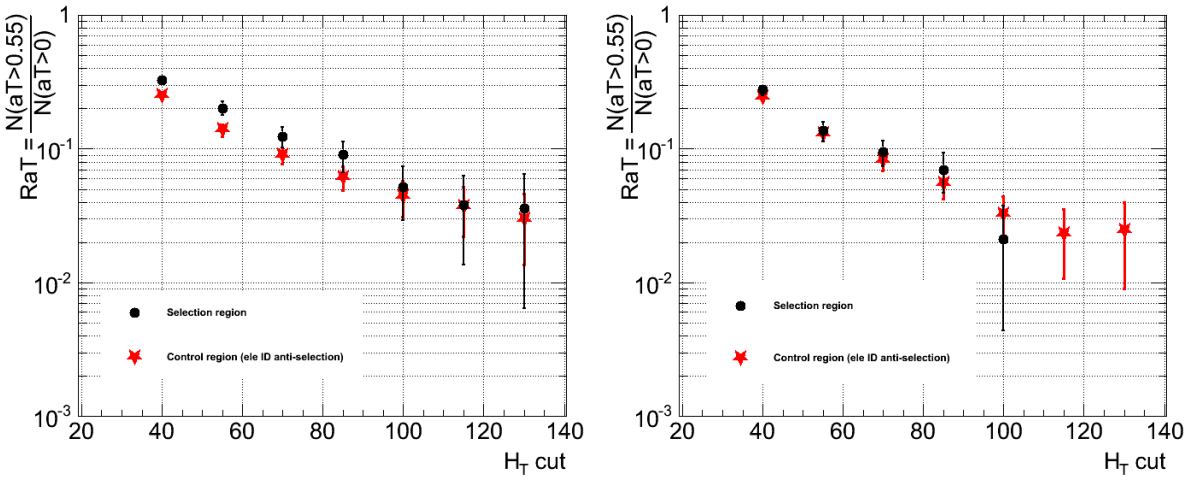

Figure 24: *The R_{α_T} versus the H_T cut applied for collision data, shown for both selected and anti-selected events in the ID Inversion method. The right plot includes a cut at pfMET<20 GeV to eliminate contamination of prompt electron sources (e.g. Ws). The collision data amounts to 12.47 nb^{-1}.*

7 Summary

This study proposes two methods of data-driven QCD background estimation, and presents results from Monte Carlo and the first X pb^{-1} of 7TeV data taken by CMS at the LHC.

Following the promising results of studies into using the α_T kinematic variable in the single electron mode of SUSY searches, it is proposed that a suitable control sample in the distribution of this variable for predicting the QCD background could be obtained by inverting the $\Delta\phi$ and $\Delta\eta$ cuts in the electron selection criteria.

The analysis uses Monte Carlo samples to perform a closure test on this method, first with a pure QCD sample, and then with W + jets contamination in the anti-selected control sample. The method proved unbiased between selected and anti-selected in these tests.

The first look at this method under 7TeV collision data and corresponding Monte Carlo is also shown, with lowered selection to maximise available statistics.

References

[1] https://twiki.cern.ch/twiki/bin/view/CMS/ProductionSummer2009.

[2] J Alwall et al., "MadGraph/MadEvent v4: The New Web generation", JHEP 09 (2007) 028, arXiv:0706.2334.

[3] https://twiki.cern.ch/twiki/bin/view/CMS/SusyPat.

[4] https://twiki.cern.ch/twiki/bin/view/CMS/SWGuidePAT.

[5] https://twiki.cern.ch/twiki/bin/view/CMS/SusyPatLayer1.

[6] https://twiki.cern.ch/twiki/bin/view/CMS/SusyCAFNtupleV000804XX.

[7] https://twiki.cern.ch/twiki/bin/view/CMS/PhysicsSecondaryDatasets.

[8] https://twiki.cern.ch/twiki/bin/view/CMS/SimpleCutBasedEleID.

[9] https://twiki.cern.ch/twiki/bin/view/CMS/ConversionBackgroundRejection.

[10] L. Agostino, A. Chatterjee, R. Patterson, D. Puigh, W. Sun, J. Thom, J. Vaughan, P. Wittich, "Estimation of Fake Electron background using Data-Driven Techniques", CMS Analysis Note 2010/043.

[11] J. Branson, M. Gallinaro, P. Ribeiro, R. Salerno, M. Sani, "'A cut based method for electron identification in CMS", CMS Analysis Note 2008/082.

[12] H. Flaecher, M. Stoye, T. Rommerskirchen, T. Yetkin, T. Whyntie, R. Bainbridge, J. Marrouche, "*Search for SUSY with exclusive n-jet events*", CMS Analysis Note 2008/082.

[13] O. Buchmueller, L. Gouskos, Z. Hatherell, G. Karapostoli, A. Sparrow, P. Sphicas, "An application of the α_T jet-balancing method to the single-lepton mode SUSY searches," CMS Analysis Note 2009/188.

[14] Paul Geffert and David Stuart, "Study of the Correlation between α_T and H_T using Z+jets events", CMS Analysis Note 2009/155.