IL NUOVO CIMENTO Vol. 33 C, N. 4 Luglio-Agosto 2010
DOI 10.1393/ncc/i2010-10667-0

Colloquia: TOP2010

Background estimation strategies in CMS

M. Barrett on behalf of the CMS Collaboration

Experimental Particle Physics Group, Centre for Sensors and Instrumentation
Department of Electronic and Computer Engineering, School of Engineering and Design
Brunel University - Uxbridge, UB8 3PH, UK

(ricevuto il 28 Luglio 2010; approvato il 23 Agosto 2010; pubblicato online l'11 Ottobre 2010)

Summary. — The simulation or data-driven estimations of the diverse background processes to top quark decays are a key activity to be performed with the first data. This document describes some of these strategies and the possible achievements with an integrated luminosity of $20\,\mathrm{pb}^{-1}$.

PACS 14.65.Ha – Top quarks.
PACS 29.85.Fj – Data analysis.

1. – Introduction

The study of top quarks produced at the LHC will form an important part of the early physics programme of the CMS experiment [1]. The study of the decays of top quarks involves reconstructing many different decay products, including electrons, muons, jets and missing transverse energy. The reconstruction of these requires every part of the detector to be utilised and understood, and thus such studies will be important early measurements for the whole CMS physics programme.

There are many backgrounds to be considered when searching for top quarks. The expected contribution of each type of background can be estimated using simulated events, generated to resemble the types of events expected in data, incorporating detector performance. However the simulated events have many uncertainties associated with them, such as those arising from extrapolating known physical processes from a lower energy regime to that to be explored at the LHC.

Therefore it is desirable to have methods available to estimate the backgrounds based on the data themselves. Some of these data-driven methods are described here, together with how they might perform with $20\,\mathrm{pb}^{-1}$ of data. The studies presented are based on simulated samples generated for LHC collisions with an energy of 10 TeV, though the methods themselves should be suitable for application to collisions at 7 TeV. None of these methods have however been tested on data, and they are therefore designed to be as robust as possible, with multiple methods for estimating the same contributions as cross-checks.

Of the three main classifications of $t\bar{t}$, each referring to the decay products W bosons originating from top quark decays, the fully hadronic channel, and channels explicitly requiring a tau lepton, are not covered here. The dileptonic channels have the lowest backgrounds, and is covered later in this document. The semileptonic channels, where the lepton is either an electron or a muon, have some important background contributions which must be taken into account, and these are the main focus of this document. The two channels are referred to as the electron+jets channel and the muon+jets channel.

2. – Lepton + jets channel

2˙1. *Selecting lepton+jets events.* – For early data, simple and robust selection criteria are desirable. The lepton+jets channels consist of a single isolated lepton accompanied by four or more jets. The detailed criteria used for selecting electron+jets [2] events or muon+jets [3] events are summarised here:

- At least four jets with: transverse momentum, $p_T > 30\,\mathrm{GeV}/c$; pseudorapidity, $|\eta| < 2.4$; and in the electron+jets channel, veto any jets which satisfy $\Delta R(e, jet) < 0.3$, where $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$, and ϕ is the azimuthal angle.
- Exactly 1 lepton:
- Electrons are selected using a single electron trigger with a threshold transverse energy of $E_T > 15\,\mathrm{GeV}$, and the following criteria: $E_T > 30\,\mathrm{GeV}$; $|\eta| < 2.5$ (excluding the region $1.442 < |\eta| < 1.560$; transverse impact parameter, $d_0 < 200\,\mu\mathrm{m}$; and relative isolation < 0.1 (where relative isolation is defined as the sum of the individual isolations, I, in the tracker and electromagnetic (Ecal) and hadronic (Hcal) calorimeters, divided by the electron's transverse energy: $\mathrm{relIso} = (I_{\mathrm{tracker}} + I_{\mathrm{Ecal}} + I_{\mathrm{Hcal}})/E_T$.
- Muons are selected using a single muon trigger, with a threshold transverse momentum of $p_T > 9\,\mathrm{GeV}/c$, and the following criteria: $p_T > 20\,\mathrm{GeV}/c$; $|\eta| < 2.1$, $d_0 < 200\,\mu\mathrm{m}$, and $\mathrm{RelIso} < 0.05$ (relIso is defined in the same manner as for the electron, except that p_T is used instead of E_T).
- Events are vetoed if they have extra isolated electrons or muons (looser selections are applied to the extra leptons).
- For early data no missing transverse momentum criterion is applied, and no attempt is made to identify if jets have originated from the decay of b quarks.

In the muon+jets channel after selection there are expected to be 277 events from $t\bar{t}$ semimuonic events, and 43 events from other $t\bar{t}$ decays. The contributions from background processes are 140 events from W boson (+jets) production, 10 events from Z boson (+jets) production, 14 events from single top production and 7 events from QCD processes ("QCD background" here covers all processes arising from QCD interactions, that are not specifically covered by another type of simulated event).

In the electron+jets channel after selection there are further refinements used to reduce specific backgrounds. To reduce the contribution from Z bosons, events with extra electrons (with looser selection criteria) can be vetoed; alternatively events with such electrons are only vetoed if the invariant mass of the pair of electrons in the event lies within $15\,\mathrm{GeV}/c^2$ of the nominal Z boson mass.

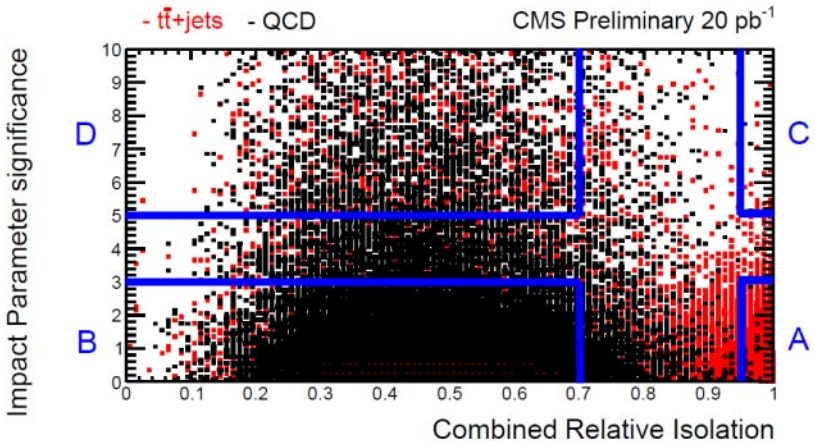

Fig. 1. – The phase space of the impact parameter significance, and combined relative isolation variables, with the regions defined as A, B, C, and D shown, together with the distributions of $t\bar{t}$ and QCD events [3].

To further reduce the QCD background, a missing transverse momentum cut can be used. Another possibility is to reduce the number of electrons selected that arise from photons converting into electron-positron pairs when interacting with the detector. This can be achieved by using only electrons detected in the Ecal barrel, as those detected in the endcaps have to pass through more material, or by using a dedicated conversion finding algorithm.

Using a baseline selection for the electron+jets channel, there would be expected to be 183 events from $t\bar{t}$ events, together with 80 events from W boson production, 28 events from Z boson production, and 9 events from single top production. There are also expected to be 30 QCD events.

2˙2. *ABCD method for estimating QCD.* – The ABCD method is used in the muon+jets channel. The method utilises two variables that are at most weakly correlated, and defines four regions within the phase space of these variables. Figure 1 shows the phase space for the variables impact parameter significance (d_0/σ_{d_0}, where σ_{d_0} is the uncertainty on the measurement of d_0), and combined relative isolation (this is defined as 1/(1+relIso) in terms of the previously defined relIso, to scale it to lie between 0.0 and 1.0).

Region A is the signal region, where signal events dominate, and the objective is to estimate the background contribution to this region using the other three regions. Assuming that the ratio of the number of events is $N_A/N_B = N_C/N_D$, where N_x is the number of events in region x, and that the other regions are dominated by background, then the number of background events in region A becomes: $N_A = N_B \times N_C/N_D$.

Table I shows the results of applying this method to events with either 2, 3 or 4 or more selected jets. The results are in good agreement with the simulated data in this closure test, though with large uncertainties for the higher jet multiplicities. The stability of the method and systematic uncertainties are estimated by varying the boundaries of the different regions. Conservatively a 50% uncertainty is placed on the method.

TABLE I. – *The results of performing the relIso extrapolation for the muon+jets channel* [3]. *The estimated number is the number obtained from the fit.*

Jets	N(QCD) Predicted	N_B	N_C	N_D	N(QCD) Estimated
2	327	86625	61	16240	325 ± 26
3	53	24216	10	5058	48 ± 9
$\geq$ 4j	7	5345	3	1148	12 ± 5

2˙3. *Relative isolation extrapolation for estimating QCD.* – The relative isolation extrapolation is used in both the electron+jets and muon+jets channels. Both analyses cut on the relative isolation of the electron or muon. Low values of this variable are associated with signal $t\bar{t}$ events, and also many backgrounds, whereas high values are dominated by QCD events, as can be seen in fig. 2. This extrapolation method involves fitting the shape of the relIso distribution for higher values, and extrapolating this function to the lower values of the signal region.

In order to establish the function to use for fitting the QCD, a control region is defined, which is QCD dominated, and the whole relIso distribution can be fitted. Functional forms including Gaussian and polynomial distributions were tried, the best fit was achieved with a Landau distribution.

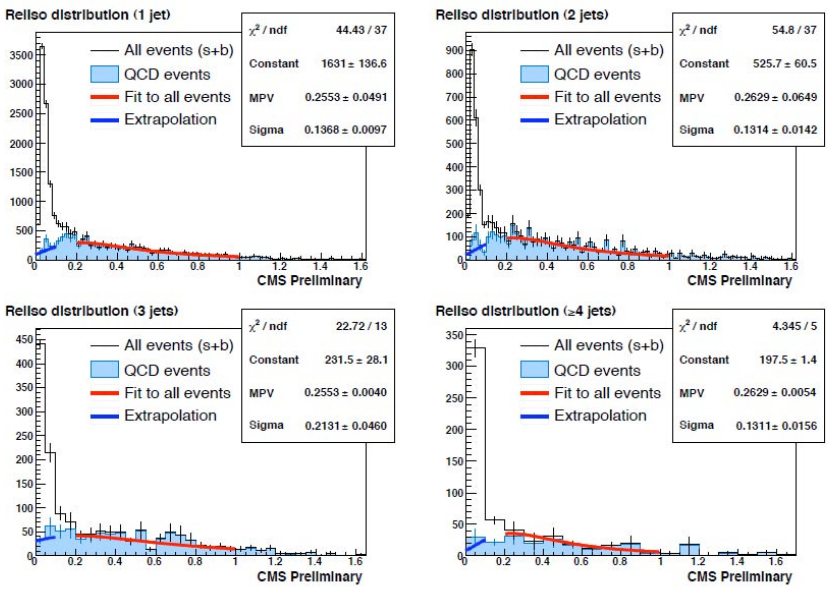

Fig. 2. – The relIso distributions for all events (unfilled histogram), and QCD events only (filled histogram). The fit using a Landau function in the range $0.2 < \text{relIso} < 1.0$ is shown, together with the extrapolation into the signal region ($\text{relIso} < 0.1$). The fit is carried out separately for samples with different jet multiplicities, and the MPV values from the 1 and 2 jet multiplicity fits are used to constrain this parameter for higher multiplicity fits [2].

TABLE II. – *The results of performing the relIso extrapolation for the electron+jets channel* [2]. *Uncertainties associated with the number of simulated events available are shown.*

	Signal region	
	True QCD	Estimate
	$20\,\mathrm{pb}^{-1}$	$20\,\mathrm{pb}^{-1}$
1j	1007 ± 102	815
2j	301 ± 47	227
3j	96 ± 28	71
$\geq$ 4j	30 ± 14	17

This functional form was then used to fit the relIso distributions for different jet multiplicities. For three and four or more jet multiplicity samples, the number of events is quite small, which is deleterious to the stability of the fit. This is ameliorated by constraining the value of the mean peak value (MPV) of the Landau distribution in the fit, by using the MPVs obtained from the one and two jet multiplicity fits.

The results of performing the fits for the electron+jets channel can be seen in table II. The fits perform quite well, but the extrapolated results are systematically slightly too low, which is attributable to remaining electrons from photon conversions. Therefore, a conservative systematic uncertainty of 50% is associated with the background estimation method.

The results for the muon+jets channel can be seen in table III. A good performance is again observed, without any systematic bias. An uncertainty of 50% is again placed on this measurement to account for the uncertainties of this method.

2˙4. *M3 method for estimating W+jets*. – The background arising from the production of W bosons, together with jets, is studied by looking at the M3 method. M3 is a variable defined as the invariant mass of the three jets in the event, that together have the highest combined transverse momentum (calculated from the vector sum of the jets' individual momenta). This variable is motivated as it is a simple estimator of the reconstructed top quark mass in the event, and the distribution is expected to peak near this value, particularly for events containing a top quark. The shape of the M3 distribution can be seen for different types of event in the electron+jets channel in fig. 3(left).

In the e+jets channel the fit is performed with four components: signal $\mathrm{t\bar{t}}$, W+jets (including Z+jets, and W/Z+heavy flavour jets), single top, and QCD. The QCD distribution is derived from events in a QCD control region (in the same manner as for the relIso extrapolation method of subsect. **2**˙3). The distributions for the other components are derived from simulated events. The W+jets shape could be investigated using a clean

TABLE III. – *The results of performing the relIso extrapolation for the muon+jets channel* [3]. *The estimated number is the number obtained from the fit. The uncertainty shown for the number of estimated events is the statistical uncertainty from the fit.*

Jets	N(QCD) Predicted	N(QCD) Estimated
2	327	378 ± 82
3	53	47 ± 24
$\geq$ 4j	7	13 ± 7

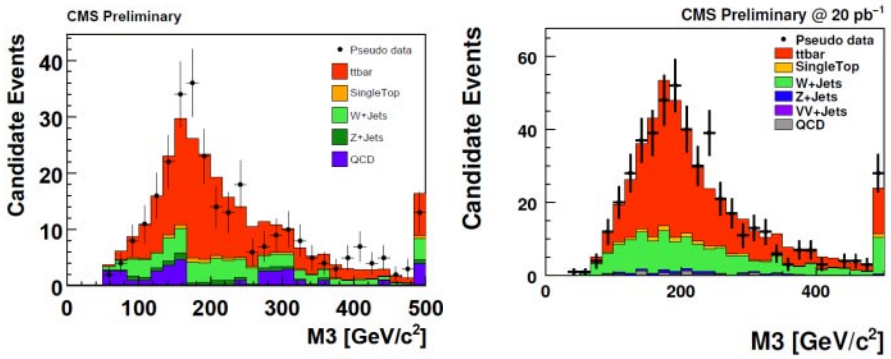

Fig. 3. – The M3 distribution for the electron+jets channel (left) and the muon+jets channel (right). The signal component is shown together with the major backgrounds. The pseudodata are randomly selected events drawn from the different contributions [2, 3].

Z+jets sample, but there will not be enough events in early data. The QCD and single top components are small, and are both constrained to aid the stability of the fit. The number of $t\bar{t}$ events, and the number of W(Z)+jets events can thus be extracted from the fit. Pseudoexperiments are conducted in order to check for bias, and to estimate the uncertainty, which is found to be ±23% on the fit components for $20\,\mathrm{pb}^{-1}$, and would decrease to ±10% with $100\,\mathrm{pb}^{-1}$.

For the muon+jets channel, fits are carried out to M3 (shown in fig. 3(right)) and two additional variables, the pseudorapidity of the muon, η_μ, and M3′. M3′ is extracted using a χ^2 function: $\chi^2 = (m_{j_1 j_2} - m_W)^2/\sigma_{jj}^2 + (m_{j_1 j_2 j_3} - m_t)^2/\sigma_{jjj}^2 + (m_{\mu\nu j_4} - m_t)^2/\sigma_{\mu\nu j}^2$, where m_W and m_t are the nominal W boson and top mass, and m_X and σ_X are the invariant mass and resolutions for the different jet combinations. The χ^2 is calculated for all the combinations of up to seven jets per event. The value of $m_{j_1 j_2 j_3}$ for the lowest calculated value of χ^2 is assigned as M3′. The M3′ distribution is fitted to the various signal and background template shapes to extract the number of $t\bar{t}$, and pseudoexperiments determine the uncertainty on this quantity to be ±12%. This method relies on being able to partially reconstruct the neutrino in the event from missing energy.

In the fits to M3, η_μ, and M3′, there are three components used: $t\bar{t}$, single top and W+jets (which includes Z+jets and QCD), with the single top component being constrained to its theoretical value with the associated uncertainty. From pseudoexperiments the uncertainty is extracted to be ±16% for the M3 fit, and ±18% for the η_μ fit. These uncertainties would decrease to ±10% with $50\,\mathrm{pb}^{-1}$ of data.

2˙5. *Charge asymmetry method for estimating W+jets.* – Assuming that the signal $t\bar{t}$ events are charge symmetric, it is possible to study the "Events leading to a Charge Asymmetry" (ECA), which are assumed to be dominated by W+jets events. The total number of ECAs is given by $(N_+ + N_-)_{\mathrm{data}} = R_\pm(W) \times (N_+ - N_-)_{\mathrm{data}}$, using the assumption that $R_\pm$ is the same for W+jets events and for all ECAs. $R_\pm(W)$ corresponds to the inverse of the W charge asymmetry: $R_\pm(W) = (N_{W^+} + N_{W^-})/(N_{W^+} - N_{W^-})$.

$(N_+ - N_-)_{\mathrm{data}}$ will have a large statistical uncertainty, and this method was studied based on a scenario with $100\,\mathrm{pb}^{-1}$ of data. It can be estimated from pseudodata using muon and antimuon events. $R_\pm$ can be estimated from statistically independent W+jets samples. This method is statistically dominated (the systematic uncertainty is estimated at ±11%), with an uncertainty of ±30% with $100\,\mathrm{pb}^{-1}$.

3. – Dileptonic channel

3˙1. *Selecting dileptonic events.* – A detailed description of the selection and analysis of dileptonic events can be found elsewhere [4], a summary of the selection used is given here:

- single lepton trigger (threshold: $E_T > 15\,\mathrm{GeV}$ for electron trigger, and $p_T > 9\,\mathrm{GeV}/c$ for muon trigger).
- Exactly two leptons with $p_T > 20\,\mathrm{GeV}/c$, $|\eta| < 2.4$, opposite signs, and individual relative isolations $I_{\mathrm{trk}} > 0.9$, and $I_{\mathrm{cal}} > 0.9$ (0.8) for μ (e).
- Two or more jets with $p_T > 30\,\mathrm{GeV}/c$, and $|\eta| < 2.4$.
- Missing transverse energy > 20 or $30\,\mathrm{GeV}$ dependent on the decay channel,

where the individual relative isolations are defined as $p_T/(p_T + I_{\mathrm{trk/cal}})$ where $I_{\mathrm{trk/cal}}$ is the individual tracker or combined calorimeter isolation. There are three dileptonic channels studied: $\mathrm{e^+e^-}$, $\mathrm{e^\pm}\mu^\mp$, and $\mu^+\mu^-$. In the $\mathrm{e^+e^-}$ and $\mu^+\mu^-$ channels there is an additional requirement: vetoing events where the invariant mass of the lepton pair is within $15\,\mathrm{GeV}/c^2$ of the mass of the Z boson. The dominant remaining background is from Drell-Yan(DY) + jets events. For $20\,\mathrm{pb}^{-1}$, there are approximately 12 signal events expected for the $\mathrm{e^+e^-}$ channel, 13 in the $\mu^+\mu^-$ channel, and 36 in the $\mathrm{e^\pm}\mu^\mp$ channel. The DY backgound is expected to be approximately 4 and 5 events in the $\mathrm{e^+e^-}$ and $\mu^+\mu^-$ channels, respectively. Other backgrounds are expected to contribute approximately 1 event in each of these channels, and approximately 3 events in the $\mathrm{e^\pm}\mu^\mp$ channel.

3˙2. *Estimating Drell-Yan+jets contribution.* – DY events are mostly selected due to the mismeasurement of the missing E_T. To estimate the contribution of DY events, dileptonic events with an invariant mass, $76 < m_{\ell\ell} < 106\,\mathrm{GeV}/c$, are used. The ratio $R_{\mathrm{out/in}}$ is estimated from simulation for the number of DY events outside the range, *versus* those inside the range. Then the number of DY events outside the range can be estimated from the number of events observed in the range in data, after correction for non-DY contributions from studying $\mathrm{e^\pm}\mu^\mp$ in the same range. Simulated events are relied on for the ratio, but there is no dependence on jet and missing transverse energy properties. The systematic uncertainty is estimated using different selections, and different simulated samples, and is estimated to be $\pm 30\%$ on the number of DY events.

3˙3. *Estimating the contribution from fake leptons.* – Fake electrons are a background that arises due to jets or components of jets being identified as an isolated lepton. In order to estimate the contribution from fake leptons, events are used that pass a looser selection, principally loosening the isolation (both calorimeter and tracker) requirements. The Fake Ratio (FR) is defined as the number of events that pass the loose selection and also pass the main selection, and is estimated from samples for QCD multijet events. The number of fake leptons passing the main selection can be estimated by weighting the number of events that pass the loose selection, but fail the main selection by $FR/(1 - FR)$. There will be small biases due to double counting, and trigger differences between the samples used. The uncertainty associated with the method is estimated from the statistics of the samples used, and variation between samples of different jet multiplicities, and from these studies and factors a 50% systematic uncertainty is assigned to this method. This

uncertainty is small, given that the number of fakes is expected to be approximately 1 event in the e^+e^- and $\mu^+\mu^-$ channels, and 2.5 events in the $e^\pm\mu^\mp$ channel.

4. – Conclusions

Early measurements of the $t\bar{t}$ cross-section will be possible at CMS using some data-driven methods to estimate some of the important backgrounds. These methods have been developed using simulated events, and are designed to be as simple and robust as possible yet, as they have not been tested on any data. The uncertainties associated with these methods have been estimated for scenarios with $20\,\mathrm{pb}^{-1}$ of data.

In the electron+jets and muon+jets channels, the dominant background is W+jets, which is expected to contribute about 140 events, compared with 277 signal events in the muon+jets channel, and 80 events, compared with 183 signal events for the electron+jets channel. QCD is also an important background, expected to contribute 30 events in the electron+jets channel and 7 events in the muon+jets channel.

The QCD contribution can be estimated using data-driven methods, using a relative isolation extrapolation, or the "ABCD" method. The uncertainty on the number of QCD events is estimated to be about ±50% for these methods. For W+jets, the contribution is estimated using template fits to variables such as M3, with some templates taken from data, and others from simulation, the uncertainty associated with these fits is in the range 15–25%. A method using the charge asymmetry of W+jets events could be used with significantly more data.

In the dilepton channel, the major background in the e^+e^- and $\mu^+\mu^-$ channels is from Drell-Yan events, of which the number of expected events is about 40% of the number of signal events. A data-driven method using events with an invariant mass consistent with the mass of a Z boson can be used to estimate the number of DY events, and has an uncertainty of ±30%. Events arising due to fake leptons, have an expectation which is up to 10% that of the signal. A method using looser selections to estimate the fake rate, can estimate the number of these events with an uncertainty of ±50%.

These techniques are still being developed, and will soon be tested on LHC collision data for the first time.

REFERENCES

[1] CMS Collaboration, *The CMS experiment at the CERN LHC*, *JINST*, **0803** (2008) S08004.
[2] CMS Collaboration, *CMS-PAS-TOP-09-004, Plans for an early measurement of the $t\bar{t}$ cross section in the electron+jets channel at $\sqrt{s} = 10\,\mathrm{TeV}$.*
[3] CMS Collaboration, *CMS-PAS-TOP-09-003, Prospects for the first Measurement of the $t\bar{t}$ Cross Section in the Muon plus Jets Channel at $\sqrt{s} = 10\,\mathrm{TeV}$ with the CMS Detector.*
[4] CMS Collaboration, *CMS-PAS-TOP-09-002, Expectations for observation of top quark pair production in the dilepton final state with the early CMS data at $\sqrt{s} = 10\,\mathrm{TeV}$.*