IL NUOVO CIMENTO Vol. 33 C, N. 4 Luglio-Agosto 2010
DOI 10.1393/ncc/i2010-10657-2

Colloquia: TOP2010

Strategies of data-driven estimations of $t\bar{t}$ backgrounds in ATLAS

A. Loginov(*)

Physics Department, Yale University - New Haven, CT, USA

(ricevuto il 9 Agosto 2010; approvato il 9 Agosto 2010; pubblicato online il 30 Settembre 2010)

Summary. — An overview of data-driven methods developed in ATLAS to evaluate main sources of backgrounds for the estimation of the $t\bar{t}$ cross section is presented. The techniques have been designed for both the single lepton and for the dilepton channels to establish confidence in background estimates without relying on the simulation.

PACS 14.65.Ha – Top quarks.
PACS 14.80.Bn – Standard-model Higgs bosons.
PACS 13.85.-t – Hadron-induced high- and super-high-energy interactions (energy $> 10\,\text{GeV}$).
PACS 13.85.Qk – Inclusive production with identified leptons, photons, or other nonhadronic particles.

1. – Introduction

The LHC has successfully started to produce pp collision data at the center-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$. The top-antitop ($t\bar{t}$) cross section ($\sigma_{t\bar{t}}$) measurement will test the Standard Model (SM) [1] at the new center-of-mass energy agaist the theoretical predictions, which nowadays are at the percent level [2].

In addition, the abundant sample of $t\bar{t}$ events will be used as a calibration tool for reconstructed leptons (ℓ, electrons e or muons μ), jets (j), missing transverst energy ($E_{\text{T}}^{\text{miss}}$) and b-quark tagging algorithms. On top of that, $t\bar{t}$ events are expected to be an important background in the searches for the Higgs boson or a physics beyond the SM, and it is therefore crucial to understand this process in details.

New physics [3] can affect both the production and decay of the $t\bar{t}$ events, modifying the observed $\sigma_{t\bar{t}}$ differently in different decay channels. In this scenario, it is critical to correctly evaluate all the possible contributions of physics backgrounds coming from SM processes.

(*) ATLAS Collaboration.

In the ATLAS experiment [4], the $\sigma_{t\bar{t}}$ will be first measured in the single lepton ($\ell + jets$) [5,6] and the dilepton ($\ell^+\ell^-$) [6,7] channels, the experimental signatures for the first one being either the $e + E_{\mathrm{T}}^{\mathrm{miss}} + jets$ or the $\mu + E_{\mathrm{T}}^{\mathrm{miss}} + jets$, and for the second one $ee + E_{\mathrm{T}}^{\mathrm{miss}} + jets$, $\mu\mu + E_{\mathrm{T}}^{\mathrm{miss}} + jets$ or $e\mu + E_{\mathrm{T}}^{\mathrm{miss}} + jets$.

The selections for both $\ell + jets$ and $\ell^+\ell^-$ events start by requiring a high-p_{T} lepton lepton (e or μ) trigger. The $\ell + jets$ selection then requires the presence of exactly one e or μ with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$. At least four high-$p_{\mathrm{T}}$ jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ (of which three should have $p_{\mathrm{T}} > 40\,\mathrm{GeV}$) with $|\eta| < 2.5$ are then required [5]. As ATLAS has already demonstrated a good performance of the b-tagging [8], we will also require at least one of the jets to be b-tagged. Finally, to suppress QCD multijet backgrounds ("QCD"), we also require large $E_{\mathrm{T}}^{\mathrm{miss}}$. After the selection, the main background sources are expected to come from the W bosons in association with jets ("W+jets") and from the QCD multijet and photon-jet events with large $E_{\mathrm{T}}^{\mathrm{miss}}$, where we expect one of the jets or photons to be misreconstructed as a lepton [5,6]. We describe data-driven methods to estimate the W+jets in sect. **3** and for the QCD in sect. **2**.

The $\ell^+\ell^-$ selection requires two oppositely-charged leptons (ee, $\mu\mu$ or $e\mu$) each satisfying $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, at least one of which must be associated to a leptonic high-level trigger object. At least two jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.5$ are required, but no b-tagging requirements are imposed. For the ee and $\mu\mu$ channels, to suppress backgrounds from Drell-Yan and QCD events, we require large $E_{\mathrm{T}}^{\mathrm{miss}}$ [7,6] and the invariant mass of the $\ell^+\ell^-$ the two leptons to be outside of the Z mass. For the eμ channel, where the background from $Z/\gamma^* \to ee$ and $Z/\gamma^* \to \mu\mu$ is expected to be much smaller, we do not need to apply such a large $E_{\mathrm{T}}^{\mathrm{miss}}$ requirement [7,6], instead we can require large total transverse energy H^T, defined as a scalar sum of transverse energies of the two leptons and all the selected jets. After the selection, the largest backgrounds are expected from W bosons produced in association with jets or photons ("W+jets", where $W \to e\nu$ and $W \to \mu\nu$, and a jet or a photon is misidentified as a lepton) and from the Z/γ^* ("Drell-Yan") associated production with jets and large $E_{\mathrm{T}}^{\mathrm{miss}}$ (mostly due to mismeasurement errors). We describe data-driven methods to estimate the contribution from mismeasured leptons in sect. **2** and for the Drell-Yan in sect. **4**.

To measure the $t\bar{t}$ cross-section, in the early days we need to establish confidence in background estimates without relying on MC description of material, detector performance and complicated backgrounds, such as $W + jets$, $Z/\gamma^* + jets$, QCD. In case of the QCD, another problem is that it would have been hard to simulate a sufficient number of events to perform detailed MC studies. The goal of the studies of the backgrounds in data is to estimate and when possible to reduce backgrounds by optimizing the objects selection. If the Monte Carlo (MC) was perfect, there would no need, but at some level, detector simulation, matrix elements, particle distribution functions (PDFs), will never be perfect, and proving they are good enough can be very hard.

Smaller backgrounds for both the single lepton and the dilepton channels include the diboson production, the $Z \to \tau\tau + jets$ and the single top production. These backgrounds are estimated from the MC and described elsewhere [7,5,6].

To estimate and reject non-collision backgrounds, such as beam-induced (muons coming from beam halo, beam-gas collisions, and also cosmic rays) we have few handles, such as MBTS (MinBias Trigger Scintillators) timing, Liquid Argon Calorimeter timing, Transition Radiation Tracker (TRT) EndCap timing (See ref. [4].) Based on the studies performed we are reasonably confident from our data analysis so far, that they will not be a big background for top analyses.

In the following we describe in more details the main background sources in the $l+jets$ and $\ell^+\ell^-$ $t\bar{t}$ decay channels, together with an overview of the data-driven methods used in ATLAS [4] to evaluate them.

2. – Jets misidentified as leptons

Although the ATLAS detector has an excellent lepton identification power, there are non-negligible contributions from processes where a lepton originates from a misreconstructed jet. The dominating misreconstruction mechanisms are a semileptonic decay of a b jet, a decay-in-flight of a $\pi^{\pm}$ or a K meson, a reconstruction of a π^0 as an electron, a reconstruction of a direct photon or a photon conversion as an electron.

In the case of the single lepton channel, a multi-jet production ("QCD") is the dominant process by which events with a misreconstructed lepton appear in the signal sample. For the dileptons, the dominant processes with events with a true reconstructed lepton and a misreconstructed lepton is the $W+jets$ production, which includes $t\bar{t} \to W+jets$.

In this section we describe data-driven methods to estimate the contribution of events with a misrecomstructed lepton to the $\ell+jets$ and the $\ell^+\ell^-$ channels.

2˙1. *The matrix method.* – The matrix method exploits differences in lepton identification-related properties between prompt isolated leptons from W and Z decays (referred to as "real" leptons below) and those where the leptons are either non-isolated or result from misidentification of photons/jets (referred to as "fake" leptons below).

We define two samples differing only in the lepton identification criteria: a "tight" sample and a "loose" sample, the former being a subset of the latter. For the samples, we require the kinematic $\ell+jets$ (subsubsect. **2˙**1.1) or the $\ell^+\ell^-$ (subsubsect. **2˙**1.2) selection criteria using either "real" or "fake" lepton identification. The tight selection typically employs the final lepton identification criteria used in the analysis, whereas the loose selection is adjusted in order to satisfy basic requirements for the method to work, which are outline below.

2˙1.1. Single lepton channel. Assuming that the number of selected events in each sample (N^{loose} and N^{tight}) can be expressed as a linear combination of the numbers of events with real and fake leptons, we define

$$\begin{aligned} N^{\text{loose}} &= N^{\text{loose}}_{\text{real}} + N^{\text{loose}}_{\text{fake}}, \\ N^{\text{tight}} &= \epsilon_{\text{real}} N^{\text{loose}}_{\text{real}} + \epsilon_{\text{fake}} N^{\text{loose}}_{\text{fake}}, \end{aligned} \tag{1}$$

where ϵ_{real} (ϵ_{fake}) is the probability for a real (fake) lepton that satisfies the loose selection criteria, to also satisfy the tight ones

$$\epsilon_{\text{real}} = \frac{N^{\text{tight}}_{\text{real}}}{N^{\text{loose}}_{\text{real}}}, \quad \epsilon_{\text{fake}} = \frac{N^{\text{tight}}_{\text{fake}}}{N^{\text{loose}}_{\text{fake}}}. \tag{2}$$

The efficiencies ϵ_{real} and ϵ_{fake} are estimated in control samples, and therefore the estimated number of events with a fake lepton in the signal $\ell+jets$ sample is

$$N^{\text{tight}}_{\text{fake}} = \frac{\epsilon_{\text{fake}}}{\epsilon_{\text{real}} - \epsilon_{\text{fake}}} (N^{\text{loose}} \epsilon_{\text{real}} - N^{\text{tight}}). \tag{3}$$

For the method to work with a reasonable precision, both the ϵ_{real} and the ϵ_{fake} must be sufficiently different so that the statistical precision of the fake background estimation is not compromised $\left(\frac{\epsilon_{\text{fake}}}{\epsilon_{\text{real}}-\epsilon_{\text{fake}}}\right)$, and should be as independent of the event topology as possible, or parametrized in such a way that they can be determined in control samples and applied to the signal sample.

The ϵ_{real} is measured in inclusive $Z \to \ell^+\ell^-$ events with the tag-and-probe method. The measurement of ϵ_{fake} requires selecting a sample enriched in QCD multijet events, such that a contamination from real leptons from W and Z decays is negligible. To do so, we require low $E_{\text{T}}^{\text{miss}}$ and choose a loose lepton selection which is expected to be dominated by the background (non-isolated e's or μ's, μ's with a large d_0 significance).

2˙1.2. Dilepton channel. For the dileptons, we also define "loose" and "tight" lepton selection criteria and then define the probabilities r and f that a real or fake loose lepton will pass the tight criteria using purified control regions. The composition of the signal samples can then be extracted by inverting eq. (4). We take into account events with both leptons being fake.

$$(4)\quad \begin{bmatrix} N_{TT} \\ N_{TL} \\ N_{LT} \\ N_{LL} \end{bmatrix} = \begin{bmatrix} rr & rf & fr & ff \\ r(1-r) & r(1-f) & f(1-r) & f(1-f) \\ (1-r)r & (1-r)f & (1-f)r & (1-f)f \\ (1-r)(1-r) & (1-r)(1-f) & (1-f)(1-r) & (1-f)(1-f) \end{bmatrix} \begin{bmatrix} N_{RR} \\ N_{RF} \\ N_{FR} \\ N_{FF} \end{bmatrix}.$$

2˙1.3. Systematics. The loose-to-tight efficiencies have a binomial statistical error due to limited control region statistics. We evaluate the systematic uncertainty due to extrapolation from the control region to the signal region by studying the variation of the fake electron constituents: heavy flavor, light flavor and photon conversion. We define three control regions, each with enhanced contributions from one constituent. We also vary lepton definition to make sure the methods are robust. The most important cross-check comes from comparing the matrix method with other methods, outlined below.

2˙2. *ABCD method*. – It is a simplified version of the matrix method discussed in subsect. **2**˙1. It relies on the assumption that the QCD event distribution can be factorized in (x, y)-plane, where x and y are two uncorrelated variables, for instance x can be a lepton isolation or impact parameter significance, and y can be $E_{\text{T}}^{\text{miss}}$.

The ABCD method is illustrated in fig. 1. Neglecting the signal contribution in regions B and D, and assuming that variables x and y are uncorrelated, the number of QCD events in the signal region can be evaluated as $N_A = N_B \times N_C/N_D$. The signal contamination in A can be taken into account measuring the ratio R_W between loose and tight prompt leptons from the Z with tag and probe. The systematics can be evaluated by changing definitions of the ABCD regions and by comparing to other methods.

2˙3. *Jet-triggered events: Antielectron method*. – We give an example of an approach to select a sample from collision data that can potentially model QCD in the $\ell + jets$ channel. The approach uses jet-triggered events and an electron selection orthogonal to the standard electron selection ("anti-electrons"), while other $e + jets$ event selection requirements remain the same.

The sample will mostly consist of QCD background to the $t\bar{t}$ e + jets $E_{\text{T}}^{\text{miss}}$ with the anti-electron $E_{\text{T}}^{\text{miss}}$ shape (after subtracting other backgrounds in the low-MET region),

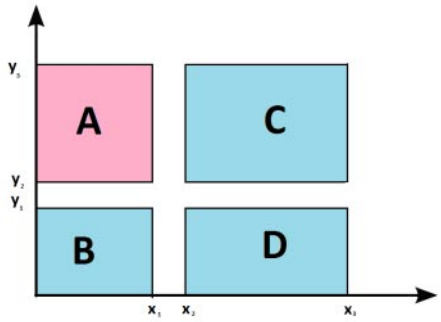

Fig. 1. – The ABCD method. The x and y axes are two uncorrelated variables. The region A is dominated by signal (large $E_{\mathrm{T}}^{\mathrm{miss}}$, low isolation or small impact parameter significance), while all others are dominated by background (either non-isolated leptons or leptons with large impact parameter significance, and/or small $E_{\mathrm{T}}^{\mathrm{miss}}$).

see fig. 2. We can use the same QCD shape for the μ+jets channel as the approach is proved to work in a number of channels at the Tevatron [9].

3. – W+jets background for the single lepton channel

W boson production in association with jets is predicted to be the dominant background to $t\bar{t}$ events in the single lepton channel [5, 6]. In the following, three complementary data-driven background estimates (W charge asymmetry, the W/Z and the $W + n + 1/W + n$ jets ratios) are described, all of which exploiting the fact that cross-section ratios are better predicted than their absolute values.

For each of the methods, the uncertainties include statistics (that is a limited factor for the early data) and systematics (parton distribution functions, jet and lepton energy scale, initial and final state radiation in the MC). The most important cross-checks come from comparing the different methods.

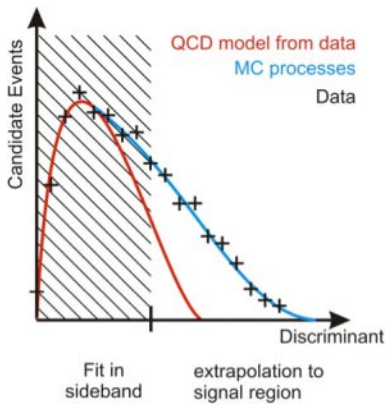

Fig. 2. – The antielectron method. The low-$E_{\mathrm{T}}^{\mathrm{miss}}$ region is fit to the lepton+jets data (after subtracting non-QCD backgrounds) using the $E_{\mathrm{T}}^{\mathrm{miss}}$ shape from the anti-electron sample.

3˙1. *Charge asymmetry method.* – At the LHC, the $t\bar{t}$ and most backgrounds are charge symmetric, while the W^+ and W^- cross-sections σ_{W^+} and σ_{W^-} are different because the PDFs of quarks and anti-quarks are different in a proton. The ratio $r \equiv \sigma_{W^+}/\sigma_{W^-}$ is well understood theoretically [10, 11]. The main theoretical uncertainty on the r is due to the PDF uncertainties and is equal to few % at the LHC energies, which is better than the prediction of the total σ_W. The method is based upon the following formula:

$$N_{W^+} + N_{W^-} = \left(\frac{r_{\mathrm{MC}}+1}{r_{\mathrm{MC}}-1}\right)(N_{W^+} - N_{W^-})_{\mathrm{data}}. \tag{5}$$

The left-hand side is the total W+jets contribution. We obtain the r_{MC} from the W+jets MC and N_{W^+} (N_{W^-}) are the numbers of W+jets events selected with a positively (negatively) charged lepton. The right-hand side is proportional to the charge asymmetry in the data and can be measured directly from the data and has to be corrected for a small contribution to the asymmetry from the electroweak production of single top quarks. This method is statistically limited for the early data taking period compared to the methods described below.

3˙2. *W/Z ratio method.* – The production of jets in association with a W or a Z boson is similar for the two bosons [5,6]. The W background to $t\bar{t}$ can then be estimated using the following formula:

$$(W^{\mathrm{SR}}/W^{\mathrm{CR}})_{\mathrm{data}} = (Z^{\mathrm{SR}}/Z^{\mathrm{CR}})_{\mathrm{data}} \cdot C_{\mathrm{MC}}, \qquad C_{\mathrm{MC}} = \frac{(W^{\mathrm{SR}}/W^{\mathrm{CR}})_{\mathrm{MC}}}{(Z^{\mathrm{SR}}/Z^{\mathrm{CR}})_{\mathrm{MC}}}, \tag{6}$$

where "SR" stands for the signal region (4 jets) and "CR" stand for a control region; the W^{CR} and Z^{CR} represent the number of W and Z candidates measured from the data in a control region at a low jet multiplicity. Z^{SR} is the number of Z candidate events produced in association with 4 jets. For the early data the method is limited by the statistical error on Z^{SR}. An additional uncertainty is given by the level of knowledge of the Monte Carlo based factor C_{MC} in eq. (6). The $W \to \tau\nu$ contribution with the tau decaying leptonically, is estimated from the MC.

3˙3. *Berends-Giele scaling method.* – An alternative method is based on the so-called "Berends-Giele scaling" [12]. Within the SM, the ratio of cross-sections (V + (n + 1) jets/V + n jets), where V = W or Z, is nearly constant as a function of n and it is equal for both W and Z. The number of W events with at least 4 selected jets can be obtained from

$$W^{\geq 4\mathrm{jets}} = W^{2\mathrm{jets}} \cdot \sum_{i=2}^{\infty} (Z^{2\mathrm{jets}}/Z^{1\mathrm{jet}})^i. \tag{7}$$

The $W + 1$ jet control sample used in the ratio method is here replaced by the $W + 2$ jet one, to reduce the uncertainty associated to the extrapolation to high jet multiplicity. While the W/Z ratio method relies on the availability of a $Z + 4$ jet sample, here the lowest statistics control sample is $Z + 2$ jets. This reduces the statistical uncertainty with respect to the ratio method. An additional source of uncertainty here, is the uncertainty on the assumption that ($V + (n + 1)$ jets / $V + n$ jets) ratio is constant as a function of n, which can be confirmed on the MC.

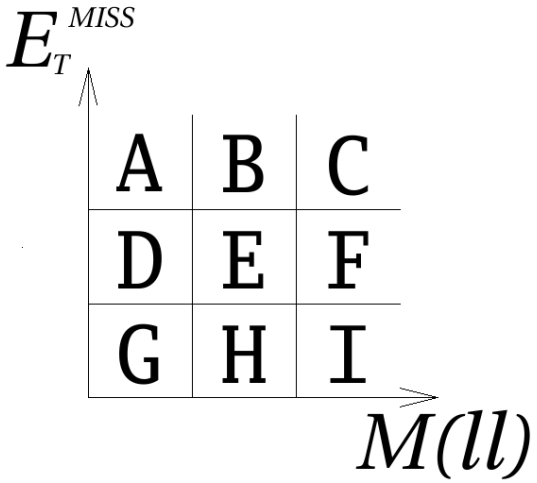

Fig. 3. – Diagram of $E_{\mathrm{T}}^{\mathrm{miss}}$ *vs.* dilepton invariant mass used for the Drell-Yan background estimates. Regions A and C are the signal region, dominated by $t\bar{t}$ for ≥ 2 jets.

4. – Drell-Yan background for the dilepton channel

We expect $Z/\gamma^* + jets$ production to be one of the largest backgrounds for the ee and $\mu\mu$ channels. The inclusive ratio Z/γ^* over $t\bar{t}$ is huge, and therefore $E_{\mathrm{T}}^{\mathrm{miss}}$ and dilepton mass cuts are applied to ee and $\mu\mu$ channels to suppress the $Z/\gamma^* + jets$ background.

For the first measurements in early data, it is assumed that the invariant mass of the lepton pairs is largely uncorrelated with the $E_{\mathrm{T}}^{\mathrm{miss}}$ in Z/γ^* events. If the two observables have no correlation, then a similarity relationship can be established by drawing a grid using these two variables (fig. 3). However, $E_{\mathrm{T}}^{\mathrm{miss}}$ and $m_{\ell\ell}$ do have small correlation, so one must tune the similarity relationship with the MC. It can be assumed that while the yield of high-$E_{\mathrm{T}}^{\mathrm{miss}}$ events in MC may be wrong, the correlation between the two variables will be reasonably modeled, and therefore

$$A_{\mathrm{est}} = G_{\mathrm{data}} \left(\frac{A_{\mathrm{MC}}}{G_{\mathrm{MC}}}\right) \left(\frac{B_{\mathrm{data}}}{H_{\mathrm{data}}}\right) \left(\frac{H_{\mathrm{MC}}}{B_{\mathrm{MC}}}\right), \tag{8}$$

$$C_{\mathrm{est}} = I_{\mathrm{data}} \left(\frac{C_{\mathrm{MC}}}{I_{\mathrm{MC}}}\right) \left(\frac{B_{\mathrm{data}}}{H_{\mathrm{data}}}\right) \left(\frac{H_{\mathrm{MC}}}{B_{\mathrm{MC}}}\right),$$

where the G and I are normalization, the B/H is the extent of Z-region Drell-Yan $E_{\mathrm{T}}^{\mathrm{miss}}$ tail; the A/G is the extent of low-mass Drell-Yan $E_{\mathrm{T}}^{\mathrm{miss}}$ tail; and the C/I is the extent of high-mass Drell-Yan $E_{\mathrm{T}}^{\mathrm{miss}}$ tail.

The total Drell-Yan estimate is then $A_{\mathrm{est}} + C_{\mathrm{est}}$. The choice to split the total yield into the high-mass and low-mass regions allows for a different correlation between the two variables in these regions, and is therefore expected to give a more accurate estimate. Additionally, the data yields in each region should be corrected for signal and other non-Z/γ^* contributions, most of which should be fairly well-modeled in the MC.

The yields in each region of the grid must be corrected for non-Z/γ^* sources, which introduces a dependence on yields for $t\bar{t}$, W+jets, and $Z \to \tau\tau$. We vary criteria used to

define exact values for $E_{\mathrm{T}}^{\mathrm{miss}}$ and $m_{\ell^+\ell^-}$ for the grid on fig. 3 to make sure the method is robust. Because the method relies heavily on data to calibrate the $E_{\mathrm{T}}^{\mathrm{miss}}$ tails, and because any biases may be studied separately in background-dominated $\ell^+\ell^- + 0\ jet$ and $\ell^+\ell^- + 1$ jet events, we have control samples to manage the systematics, where the biggest contributors are $E_{\mathrm{T}}^{\mathrm{miss}}$, jet and lepton energy scales.

5. – Conclusions

We presented an overview of methods used in ATLAS to evaluate the main sources of backgrounds for the $\sigma_{t\bar{t}}$ measurement. We have developed different data-driven methods to ensure our understanding of various backgrounds for both single lepton and dilepton channels.

REFERENCES

[1] Glashow S. L., *Nucl. Phys.*, **22** (1961) 588; Weinberg S., *Phys. Rev. Lett.*, **19** (1967) 1264; Salam A., *Proceedings of 8th Nobel Symposium, Stockholm, 1968*, edited by Svartholm (Almquist and Wiksells, Stockholm) 1968, p. 367.

[2] Moch S. and Uwer P., *Phys. Rev. D*, **78** (2008) 034003, arXiv:0804.1476 [hep-ph]; Langenfeld U., Moch S. and Uwer P., arXiv:0907.2527 [hep-ph].

[3] For instance, a t' decaying in Wq would mimic $t\bar{t}$ single lepton signature. Silva-Marcos J., *J. High Energy Phys.*, **0212** (2002) 036; Sultansoy S. *et al.*, *Acta Phys. Polon. B*, **37** (2006) 2839.

[4] Aad G. *et al.* (ATLAS Collaboration), *JINST*, **3** (2008) S08003.

[5] The ATLAS Collaboration, *Prospects for the Top Pair Production Cross-section at* $\sqrt{s} = 10$ *TeV in the Single Lepton Channel in ATLAS*, ATL-PHYS-PUB-2009-087.

[6] The ATLAS Collaboration, *Expected Performance of the ATLAS Experiment: Detector, Trigger and Physics*, CERN-OPEN-2008-020, pp. 874–878.

[7] The ATLAS Collaboration, *Prospects for measuring top pair production in the dilepton channel with early ATLAS data at* $\sqrt{s} = 10$ *TeV*, ATL-PHYS-PUB-2009-086.

[8] van Vulpen Ivo, these proceedings.

[9] Auerbach B., Ph.D thesis, Yale University, New Haven, CT, USA, expected in 2011.

[10] Martin A. D., Stirling W. J., Thorne R. S. and Watt G., *Eur. Phys. J. C*, **63** (2009) 189285, arXiv:arXiv:0901.0002 [hep-ph].

[11] Kom C. H. and Stirling W. J., arXiv:arXiv:1004.3404 [hep-ph].

[12] Berends F. A., Giele W. T., Kuijf H., Kleiss R. and Stirling W. J., *Phys. Lett. B*, **224** (1989) 237.