Search for the SM Higgs boson in the $(W/Z)H$ channel with $H \to b\bar{b}$ using the ATLAS detector at the LHC

Garoé González Parra

Ph.D. thesis
Institut de Física d'Altes Energies
Universitat Autònoma de Barcelona
Departament de Física
Facultat de Ciències

January 2015

co-Directors:
Prof. Mario Martínez Pérez
Institució Catalana de Recerca i Estudis Avançat
Institut de Física d'Altes Energies
Universitat Autònoma de Barcelona

Dr. Verónica Sorin
Institut de Física d'Altes Energies

Contents

Introduction

The Higgs boson was discovered in 2012 by the ATLAS and CMS experiments at the LHC collider at CERN. After the discovery, the experiments are trying to determine all the properties of this new particles, including a precise measurement of its mass, its quantum numbers and the couplings to the different SM particles. In this context, the search for a Higgs boson decaying into a pair of quarks is a crucial part of the LHC Higgs program and the subject of this thesis.

The $H \to b\bar{b}$ decay channel is the most promising one with a branching ratio of about 58 % for a Higgs boson mass of 125 GeV. The search for the inclusive production of the Higgs in this final state is virtually impossible given the huge (many orders of magnitude larger) multijet background with b-jets in the final state. Therefore, the associate production channel is considered, where the Higgs boson is produced in association with a W or Z vector boson in the final state. The use of the W and Z leptonic decays make possible to maintain the SM background to manageable levels.

This thesis describes the analysis of the search for the Higgs boson in the $(W/Z)H$ channel with $H \to b\bar{b}$ using the 7 TeV and 8 TeV proton-proton data collected by the ATLAS experiment at the LHC. Special emphasis is put in the ZH ($Z \to \nu\nu$ and $H \to b\bar{b}$) channel and the 8 TeV large data sample. The analysis is performed using two different techniques: a cut-and-count method relying in the $H \to b\bar{b}$ invariant mass signal, and a multivariate analysis. This thesis puts emphasis on the former.

The document is organized as follows. Chapter 1 introduces the Standard Model and the BHE mechanism for electroweak symmetry breaking, and discusses the Higgs boson production and decay modes. Chapter 2 is devoted to the phenomenology of proton-proton collisions. The statistical framework, tools and methodology used in the analysis is presented in Chapter 3. Details on the LHC accelerator and the ATLAS experiment are provided in Chapter 4, while Chapter 5 described the reconstruction in the detector of the physics objects used in the analysis. Chapters 6 and 7 detail the Higgs search in the VH channel using 20.3 fb^{-1} of 8 TeV, which is complemented with an appendix on the multivariate analysis. The final results, including a combination with the 7 TeV results, based on 4.7 fb^{-1} of data, are presented in Chapter 8.

The thesis is complemented with Appendix A, which presents the main characteristics of the analysis that targets the vector boson decays $W \to l\nu$ and $Z \to ll$.

The results presented in this thesis led to the following publication by the ATLAS Collaboration:

- *Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated (W/Z)H production with the ATLAS detector.* ATLAS Collaboration, JHEP (2015) 069.

This paper is based on the work developed during 2011 and 2012 data taken periods, which includes one paper and three collaboration public notes:

- *Search for the Standard Model Higgs boson produced in association with a vector boson and decaying to a b-quark pair with the ATLAS detector.* ATLAS Collaboration, Phys.Lett. B718, 369 (2012).
- *Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated W/ZH production with the ATLAS detector.* ATLAS Collaboration, ATLAS-CONF-2013-079.
- *Search for the Standard Model Higgs boson produced in association with a vector boson and decaying to bottom quarks with the ATLAS detector.* ATLAS Collaboration, ATLAS-CONF-2012-161.
- *Search for the Standard Model Higgs boson produced in association with a vector boson and decaying to a b-quark pair using up to $4.7\,\mathrm{fb}^{-1}$ of pp collision data at $\sqrt{s} = 7\,\mathrm{TeV}$ with the ATLAS detector at the LHC.* ATLAS Collaboration, ATLAS-CONF-2012-015.

Chapter 1

Theoretical framework

The Standard Model (SM) of particle physics is the theory that so far best describes nature at the sub-nuclear level. It uses a set of fundamental particles to describe matter and its interactions. The theory has been broadly tested and it has been confirmed with high precision since it was developed in the sixties and seventies. The last piece missing and predicted by the theory was the Higgs field and the associated Higgs boson, responsible of giving mass to the rest of the particle via the Higgs mechanism. The particle was discovered at CERN in 2012 after decades of searching of it.

This Chapter summarizes the main concepts of the SM and the Higgs mechanism, including the discovery and the latest results on the characterization of the Higgs boson.

1.1 The Standard Model of particle physics

The Standard Model divides the elementary particles in “matter” (fermions) and “force mediator” (bosons) particles [1, 2]. It describes the fundamental forces as interactions among particles through the exchange of mediators. These mediators are spin 1 bosons that depending on the force are called photons (γ), to describe electromagnetic interactions; gluons (g), for the strong interactions; and $W^{\pm}$, Z bosons for the weak interactions. All fundamental forces described by the SM are presented in Table 1.1.

Interaction	Symbol	Mass	Electric Charge
Electromagnetic	γ	0	0
Strong	g	0	0
Weak (charged)	$W^{\pm}$	80.38 GeV	±1
Weak (neutral)	Z	91.19 GeV	0

Table 1.1: Table of carrier bosons of the fundamental interactions.

Fermions are classified in quarks and leptons and sub-divided in three generation families. Generations of quarks and leptons are copies with the same properties but their masses, having the 1^{st} generation the lighter particles and the 3^{rd} the heaviest. Each generation of quarks is naturally grouped in pairs: (u,d), (c,s), (t,b). These quarks have color and fractional electric charge, thus they interact through all types of forces (electromagnetic, weak and strong). On the other hand, each family of leptons is formed by a negative charged particle and a neutrino. They do not present color charge, therefore they do not interact via strong force. Instead, charged leptons interact via electromagnetic and weak force, while neutrinos due to their neutral charge and low mass can only be detected via weak force. All leptons, their masses and charges are listed in Table 1.2. Additionally, each quark and lepton has an anti-particle multiplying the presented fermion spectrum by two. The anti-particles are characterized by having the same masses and properties but opposite electric charges and quantum numbers, e.g. an anti-electron, so-called positron, has a charge of +1 and an anti-u has a charge of -2/3 and anti-color. As it is explained in Section 1.1.1, quark and gluons do not appear in nature as single particles but as color-neutral hadronic bound states. These states can be formed by a quark and anti-quark to construct *mesons*, e.g. the pion $\pi^{\pm}$, or three quarks or anti-quarks to construct *baryons*, e.g. the proton or the neutron.

Generation	Name	Symbol	Mass	Charge
	Quarks			
1^{st}	Up	u	2.3 MeV	2/3
	Down	d	4.8 MeV	-1/3
2^{nd}	Charm	c	1.275 GeV	2/3
	Strange	s	95 MeV	-1/3
3^{rd}	Top	t	173.5 GeV	2/3
	Bottom	b	4.65 GeV	-1/3
	Leptons			
1^{st}	Electron	e	0.51 MeV	-1
	Electron neutrino	ν_e	< 2 eV	0
2^{nd}	Muon	μ	105.66 MeV	-1
	Muon neutrino	ν_μ	< 2 eV	0
3^{rd}	Tau	τ	1.77 GeV	-1
	Tau neutrino	ν_τ	< 2 eV	0

Table 1.2: Table of quark and lepton families with their mass and charge according to the Particle Data Group [3].

The SM is a quantum field theory (QFT) [1] based on the gauge symmetry group $SU(3)_C \otimes SU(2)_L \otimes U(1)_Y$. The QFT describes particles as local fields functions of space-time coordinates $\varphi(x)$. Then the dynamics of the fields are described constructing a lagrangian density $\mathcal{L}$, which is function of the field and it first derivatives $\partial_\mu \varphi(x)$:

$$\mathcal{L}(\varphi(x), \partial_\mu \varphi(x)). \tag{1.1}$$

In addition, a gauge theory introduces gauge fields (boson fields) to keep invariant the lagrangian under local transformations. The number of associated boson fields is equal to the number of generators of the symmetry group. In the SM, the gauge symmetry $SU(3)_C$ determines the strong interaction mediated by gluons, while the $SU(2)_L \otimes U(1)_Y$ gauge symmetry governs the electroweak interaction mediated by the photons and the $W^\pm$ and Z bosons.

The SM is described with a lagrangian density in two terms, one including the strong sector described with the quantum chromodynamics (QCD) theory, and the other the electroweak sector, which unifies electromagnetic and weak interactions:

$$\mathcal{L}_{SM} = \mathcal{L}_{QCD} + \mathcal{L}_{EW}. \tag{1.2}$$

1.1.1 Quantum Chromodynamics

The Quantum Chromodynamics theory (QCD) [4, 5, 6] describes the strong interactions between quarks and gluons. Its lagrangian density can be written as

$$\mathcal{L}_{QCD} = \bar{q}_f i\gamma^\mu D_\mu q_f - \frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a, \tag{1.3}$$

where

$$D_\mu = \partial_\mu + i g_s T_a G^a_\mu \tag{1.4}$$

is the covariant derivative, γ^μ are the Dirac matrices, T_a are the Gell-Mann matrices, q_i correspond to the quark field of flavor i and $g_s^2 \equiv 4\pi\alpha_s$ is the coupling constant that gives the strength of the interaction. The gluon field tensor $G^a_{\mu\nu}$ is written as

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f_{abc} G^b_\mu G^c_\nu, \tag{1.5}$$

where G_μ represents the gluon field and f_{abc} is the structure constant of the $SU(3)$ group. In the expressions a, b and c corresponds to the color indices, which run from 1 to 8 and accounts for the different gluon mediators; and the f index is the quark flavor index which runs from 1 to 6. The QCD is a non-abelian theory, therefore self-interaction between the gluon fields are included. The non-abelian nature arises in the last term of Eq. 1.5 which describes three and four gluon vertices interactions.

The presence of self-gluon interactions in QCD explains the behavior of the coupling strength α_s as a function of the energy scale (typically, momentum transfer in the process evaluated). Figure 1.1 presents this α_s dependence. If the strong force is evaluated at very high energies or equivalently, at very short distances, the quarks and gluons interact weakly. This effect is known as asymptotic freedom and explains why quarks and gluons behave as free particles at very short distances. In addition, it allows to apply perturbative theory calculations of QCD. On the other hand, at low energies, or equivalently large distances, the strong interaction coupling diverges causing that gluons and quarks can not be considered free particles. Chapter 2 explains the use of perturbative theory in the calculation of cross sections for hadron-hadron collisions, and the approximations adopted to describe the hadronization process and the formation of jets using non-perturbative QCD models.

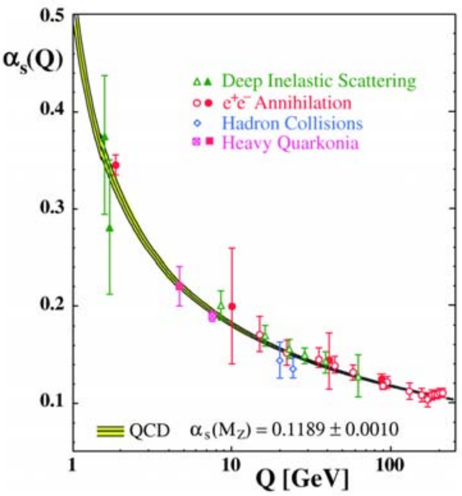


Figure 1.1: Strong coupling α_s as a function of the energy scale Q [7].

1.1.2 Electroweak unification

The electroweak theory describes the weak and the electromagnetic interactions. It unifies the forces in the symmetry group $SU(2)_L \otimes U(1)_Y$. When the symmetry is broken the mediators (photon, $W^{\pm}$ and Z) arise.

The theory defines the fermions as left- and right-handed using the projections on the fermion fields: $\psi_{L,R} = 1/2(1 \mp \gamma^5)\psi$. Left-hand fermions are doublets transforming under $SU(2)_L \otimes U(1)_Y$ and interacting electromagnetically and weakly, while right-hand fermions are singlets transforming only under $U(1)_Y$ and therefore interacting via neutral currents mediated by Z or γ. The fields transform as:

$$\psi_L = \begin{pmatrix} \nu_e \\ e \end{pmatrix}_L, \begin{pmatrix} u \\ d \end{pmatrix}_L, ... \qquad \psi_L \rightarrow \psi'_L = e^{iY\beta(x)+\bar{\sigma}\bar{\alpha}(x)}\psi_L \quad (1.6)$$

$$\psi_R = e_R, u_R, d_R, ... \qquad \psi_R \rightarrow \psi'_R = e^{iY\beta(x)}\psi_R \quad (1.7)$$

where $\bar{\alpha}(x)$ and $\beta(x)$ are the local phases, $\bar{\sigma}$ corresponds to the weak generator (the Pauli matrices) and Y to the hypercharge, defined as $Y = 2(Q - I_3)$, where Q is the electric charge and I_3 the third component of the isospin.

The boson fields associated to the electroweak symmetries are defined as

$$W^l_{\mu\nu} = \partial_\mu W^l_\nu - \partial_\nu W^l_\mu - g\epsilon_{ljk}W^j_\mu W^k_\nu \quad (1.8)$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu, \quad (1.9)$$

where g corresponds to the coupling to the electroweak W^l boson field ($l = 1, 2, 3$).

With the fermions and the boson fields defined, the electroweak lagrangian density can be written as

$$\mathcal{L}_{EW} = \bar{\psi} i\gamma^\mu D_\mu \psi - \frac{1}{4}W^l_{\mu\nu}W^{\mu\nu}_l - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}, \quad (1.10)$$

where its covariant derivative is defined as

$$D_\mu = \partial_\mu + i\frac{g}{2}\bar{\sigma}W^l_\mu + i\frac{g'}{2}B_\mu Y. \quad (1.11)$$

The g' represents the coupling to the boson field B. By expanding the first term of the lagrangian in Eq. 1.10, it is shown the differences between the interactions of the vector bosons with the left- and right-handed fermion

$$\mathcal{L}_{EW,int} = -\bar{\psi}_L\gamma^\mu\left(\frac{g}{2}\bar{\sigma}W^l_\mu + \frac{g'}{2}B_\mu Y\right)\psi_L - \bar{\psi}_R\gamma^\mu\left(\frac{g'}{2}B_\mu Y\right)\psi_R. \quad (1.12)$$

1.1.3 The Brout-Englert-Higgs mass mechanism

The previous formalism predicts the existence of 4 massless mediators to describe the electroweak force (W^l and B). Experiments discovered that $W^\pm$ and Z bosons, associated with the weak force are massive [8, 9, 10, 11]. Moreover, fermions are described as massless particles, against observation. Peter Higgs, in parallel with Robert Brout, François Englert and others developed a model for massive weak mediator fields through a spontaneous symmetry breaking (SSB) that preserves gauge invariance. It postulates an additional field, generally called, the Higgs field [12, 13, 14]. As a consequence of the introduction of this field in the theory, masses of the rest of the particles are explained by interactions between them and the Higgs field.

The simplest case to describe the field is introducing a complex doublet

$$\phi = \begin{pmatrix} \varphi^+ \\ \varphi^0 \end{pmatrix} = \begin{pmatrix} \varphi_1 + i\varphi_2 \\ \varphi_3 + i\varphi_4 \end{pmatrix} \tag{1.13}$$

with a particular potential, like

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2, \tag{1.14}$$

where μ and λ are real constants. When searching for its minimum different choices can be made. The case $\lambda < 0$ is unphysical, leading to no stable minima, while for $\lambda > 0$ two possible solutions arise. The first one, $\mu^2 > 0$ gives a vacuum state $\phi = 0$ where the symmetry stays unbroken. The second solution, $\mu^2 < 0$ gives the potential a non-vanishing vacuum expectation value in a set of infinite degenerate states. Figure 1.2 shows the potential as a function of the field ϕ for the two solutions. One of the possible vacuum states is

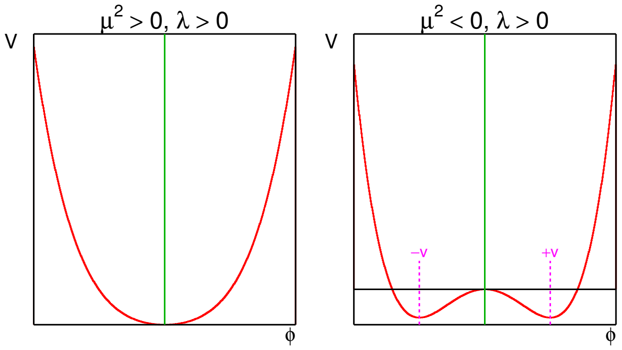


Figure 1.2: The potential V of the scalar field ϕ in the case $\lambda > 0$ and (left) $\mu^2 > 0$ or (right) $\mu^2 < 0$ [15].

$$\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}, \tag{1.15}$$

where $v = \sqrt{-\mu^2/\lambda}$ is the vacuum expectation value. If fluctuations that break the symmetry around these ground state are considered, a boson associated with the Higgs field arises

$$\phi_0(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{1.16}$$

The Higgs boson is chosen to have a hypercharge of $Y = 1$. Therefore the vacuum breaks $SU(2)_L \otimes U(1)_Y$, but leaves $U(1)_{EM}$ invariant, leaving only the photon massless. The contribution to the lagrangian density of the Higgs field is written as

$$\mathcal{L}_H = (D_\mu\phi)^\dagger(D_\mu\phi) - V(\phi^\dagger\phi). \tag{1.17}$$

The interaction between the weak mediators and the Higgs field gives the masses to the $W^\pm$ and Z bosons. Combining the Higgs lagrangian density with Eq. 1.16 and 1.11, the electroweak boson fields can be defined as linear combination of the W^l and B fields

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp iW_\mu^2) \tag{1.18}$$

$$Z_\mu = \frac{g}{\sqrt{g^2+g'^2}}W_\mu^3 - \frac{g'}{\sqrt{g^2+g'^2}}B_\mu \tag{1.19}$$

$$A_\mu = \frac{g'}{\sqrt{g^2+g'^2}}W_\mu^3 + \frac{g}{\sqrt{g^2+g'^2}}B_\mu, \tag{1.20}$$

where A_μ corresponds to the electromagnetic field associated with the photon. The masses associated with the electroweak mediators arise as new terms in the lagrangian once the Higgs field is included giving

$$m_W = \frac{gv}{2} \tag{1.21}$$

$$m_Z = \frac{v}{2}\sqrt{g^2+g'^2} \tag{1.22}$$

$$m_\gamma = 0 \tag{1.23}$$

$$\tag{1.24}$$

Fermion masses and interaction fermion-Higgs terms will also arise by adding the Higgs field into the lagrangian. These terms are proportional to the Yukawa couplings λ_f, which in the particular case of the electrons is

$$\mathcal{L}_e = \lambda_e(\bar{l}_L\phi e_R + \bar{e}_R\phi^\dagger l_L), \tag{1.25}$$

where l_L refers only to the first family of left-handed lepton. Using the definition of Higgs field ϕ presented in Eq. 1.16, the lagrangian density takes the form

$$\mathcal{L}_e = -\frac{\lambda_e v}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L) - \frac{\lambda_e h}{\sqrt{2}}(\bar{e}_L e_R + \bar{e}_R e_L). \tag{1.26}$$

A general definition of the Yukawa coupling for all fermions is written as

$$\lambda_f = \sqrt{2}\frac{m_f}{v}. \tag{1.27}$$

The Higgs boson mass is related to λ and v as

$$m_H = \sqrt{2\lambda v^2}. \tag{1.28}$$

1.2 Higgs boson at the LHC

This Section describes the production of the Higgs boson at hadron colliders, its different decay modes, and the most relevant results released by the experiments about its discovery and characterization.

1.2.1 Production modes

At the collision energies reached by hadron colliders, the dominant Higgs production modes access are: the gluon-gluon fusion (ggF), the vector boson fusion (VBF), the associated production with W or Z (VH) and the production in association with a $t\bar{t}$ pair (ttH). Feynman diagrams and cross sections for pp collision at $\sqrt{s} = 8\,\text{TeV}$ are presented in Figures 1.3 and 1.4, respectively.

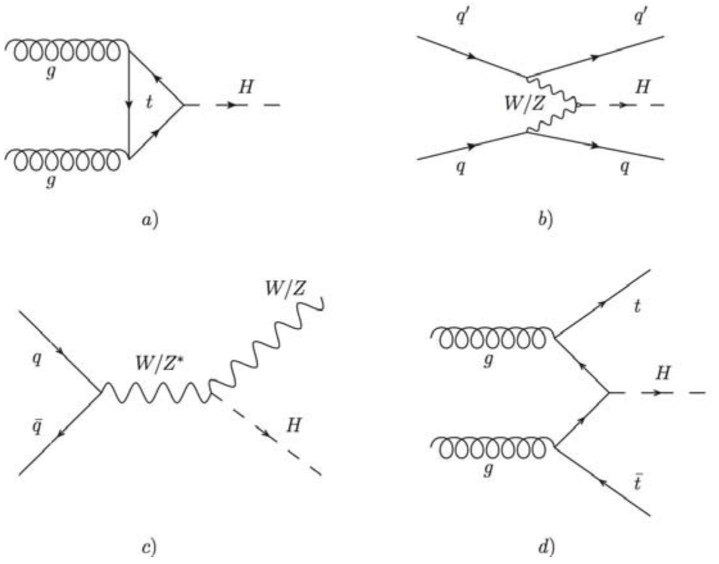


Figure 1.3: Feynman diagrams of the SM Higgs production: a) gluon fusion, b) vector boson fusion, c) vector boson associated and d) $t\bar{t}$ associated production.

1.2.2 Decay modes

The Higgs boson is an unstable particle that has a very short lifetime ($1.56 \times 10^{22}\,\text{s}$). Its main decay modes include decays into pairs of fermions or gauge bosons. Since photons and gluons are massless, they do not couple directly to the Higgs boson. Nevertheless, $\gamma\gamma$ or gg final states can be generated via loops involving massive particles. The couplings between the gauge bosons and the Higgs goes like M_V^2/v, while the coupling to fermions are

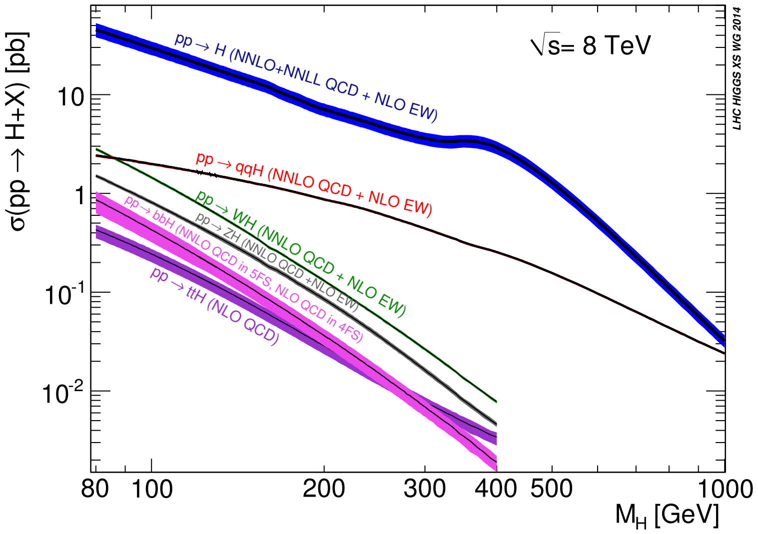


Figure 1.4: Standard Model Higgs boson production cross sections $\sqrt{s} = 8\,\text{TeV}$. The bands indicate the total theoretical uncertainties.

directly proportional to the masses. Therefore the Higgs boson branching ratio (BR) depends on the masses of the final state or the virtual (if the decay occurs via loops) particles. Figure 1.5 shows the Higgs BR as a function of its mass. At a Higgs mass of $m_H = 125\,\text{GeV}$ the dominant decays are: $b\bar{b}$ (58 %), WW (21 %), gg (9 %), $\tau\bar{\tau}$ (6 %), $c\bar{c}$ (3 %) and ZZ (3 %). For the high mass range, $m_H > 140\,\text{GeV}$, the decays to W and Z pairs become dominant.

1.2.3 Higgs boson searches

The sensitivities associated with the different Higgs decay modes depend on the understanding and the rate of the backgrounds. Therefore, physics analyses exploit the kinematics and the unique signatures of the various decay channels to efficiently suppress the backgrounds. Although the highest production rate is given by the process ggF→ $H \to b\bar{b}$, the huge $b\bar{b}$ background produced at hadron colliders makes essentially impossible the search and study of the Higgs boson via this process. Despite their relative low production rate, the most favorable decay modes to discover the Higgs boson at hadron colliders, due to their low background rate and their high mass resolution, are the resonant channels $H \to \gamma\gamma$ and $H \to ZZ$, followed by $H \to WW$.

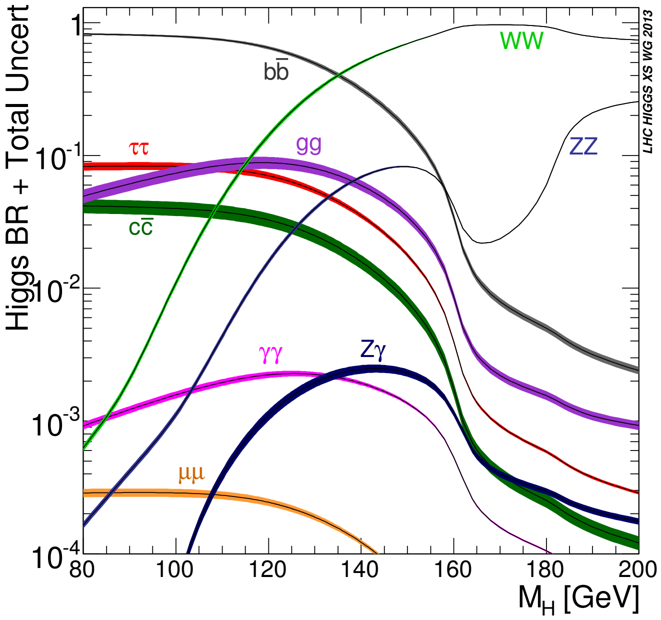


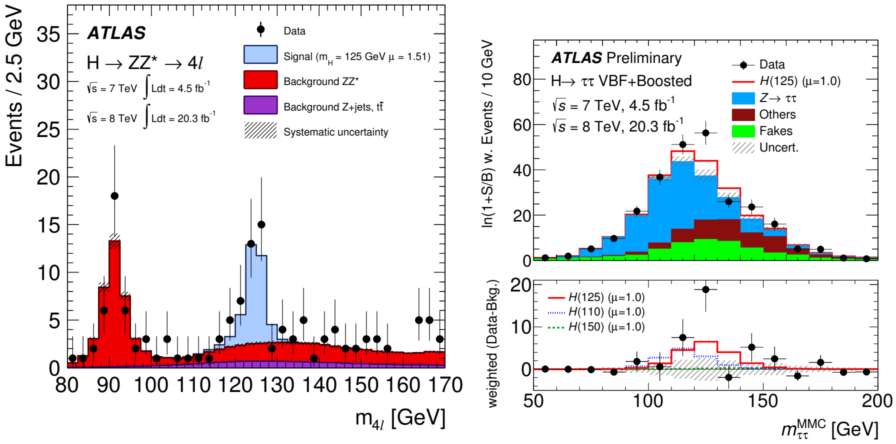
Figure 1.5: Standard Model Higgs boson decay branching ratios as a function of the boson mass.

1.2.4 Results on the Higgs boson search at the LHC

In July, 2012 CERN announced the discovery of a new particle compatible with the SM Higgs boson. The resonance was observed by both ATLAS [16] and CMS [17] experiments independently. Since then, experiments have been characterizing the new particle by measuring its mass, spin and coupling to other particles.

The resonance of the Higgs boson has been observed in several decay channels. Figure 1.6a shows the invariant mass of four leptons (m_{4l}) where a resonance from the process $H \rightarrow ZZ \rightarrow 4l$ is observed. The particle has been also observed decaying into fermions, Figure 1.6b presents the Higgs signal from the $H \rightarrow \tau\bar{\tau}$ process in the invariant $\tau\bar{\tau}$ mass distribution.

The mass of the boson has been measured using the decay channels with better energy resolution: $\gamma\gamma$ and ZZ. Figure 1.7 presents the measured cross section (relative to the SM Higgs expectation) versus the Higgs mass. ATLAS obtains a result of $m_H = 125.36 \pm 0.37(stat) \pm 0.18(syst)$ GeV [18], compatible with the CMS result $m_H = 125.7 \pm 0.3(stat) \pm 0.3(syst)$ GeV [20]. Using the same decay modes, measurements on the spin and parity of the particle has shown a strong preference for SM Higgs boson expectation values $J^P = 0^+$ [21].



(a) (b)

Figure 1.6: (a) Distributions of the four-lepton (m_{4l}) and (b) the $\tau\bar{\tau}$ ($m_{\tau\tau}$) invariant masses for the selected candidates (filled circles) compared to the expected signal and background contributions (filled histograms) [18, 19].

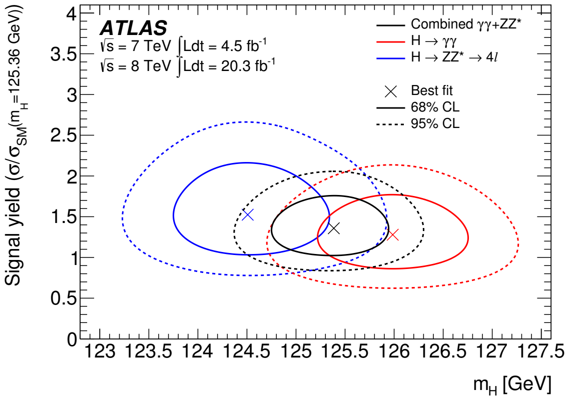


Figure 1.7: Likelihood contours as a function of the normalized signal yield σ/σ_{SM} and m_H for the $H \to \gamma\gamma$ and $H \to ZZ \to 4l$ channels, and their combination [18].

Chapter 2

Phenomenology of *pp* collisions

This Chapter gives a brief overview of the proton structure, the phenomenology of the *pp* collisions, and the different steps performed in their simulation. A small description of the main Monte Carlo generator programs, related with this analysis, is also presented.

2.1 Proton structure

The proton is not an elementary particle, it is formed by quarks and gluons, usually referred as partons. To describe their density inside the proton, it is used the parton distribution function (PDF). This function describes the probability of finding a parton of type i carrying a fraction of the proton momentum x

$$\int_0^1 x \sum_i f_i(x)dx = 1, \tag{2.1}$$

where f corresponds to the PDF and the sum runs over all parton types (g, u, $\bar{u}$, d,...). The PDFs can not be calculated analytically using QCD and its estimation relies on experimental data. This data comes from a broad type of experiments such as hadron-hadron collisions or deep inelastic scattering, among others.

The PDF evolution with the transferred momentum Q^2 can be predicted in perturbation series of α_s, known as perturbative QCD (pQCD), using the DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) function [22]:

$$\frac{df_i(x,Q^2)}{dQ^2} = \sum_j \frac{\alpha_s(Q^2)}{2\pi} \int_x^1 \frac{dz}{z} P_{j\to i}(z) f(x/z, Q^2), \tag{2.2}$$

where $P_{j\to i}$ is the splitting function that gives the probability that a process like $j \to i$ takes place. In this process, a parton of type j radiates a quark or a gluon and becomes a parton i carrying a fraction z of its longitudinal momentum. Figure 2.1 shows some of the PDFs for the different partons as a function of x for two different values of Q^2. It is observed that the u and d valence quarks, that form the proton bound state, carry around 0.5 of the proton momentum, while the rest is carried by virtual gluons and, to a lesser extent, by sea quarks from gluon splitting.

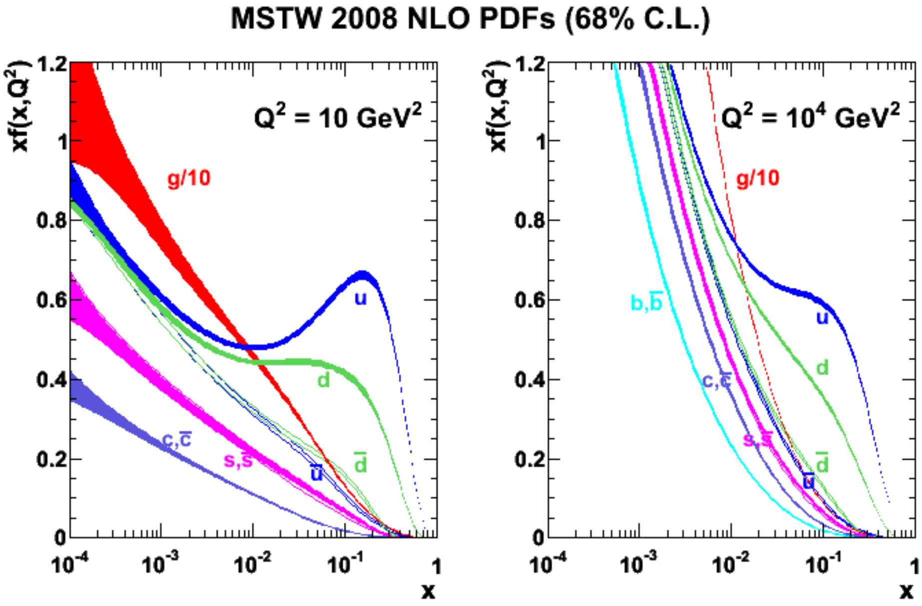


Figure 2.1: Proton PDFs and their associated error at transfer momentum $Q^2 = 10\,\mathrm{GeV}$ ($Q^2 = 10^4\,\mathrm{GeV}$) on the left (right).

2.2 The QCD factorization theorem and the *pp* cross section calculation

Cross section calculations in hadron-hadron collisions are possible thanks to the QCD factorization theorem. This theorem states that a cross section can be factorized into short- and long-distance effects delimited by a factorization scale μ_F. Short-distance effects are calculable in pQCD, while for long-distance, corresponding to soft and collinear parton emissions, in the non-perturbative QCD regime. Figure 2.2 illustrates a hadron-hadron collision where the scale μ is used to separate these effects.

According to the factorization theorem, the cross section for a proton-

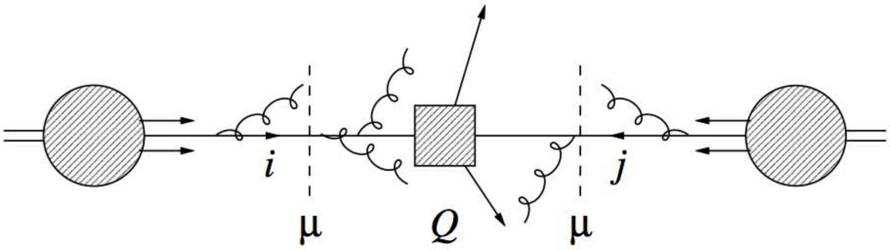


Figure 2.2: Illustrative hadron-hadron interaction where the scale μ delimit its hard process.

proton interaction with momenta $P1$ and $P2$ can be written as

$$\sigma(P_1,P_2) = \sum_{i,j} \int dx_1 dx_2 \; f_i(x_1,\mu_F^2) f_j(x_2,\mu_F^2) \;\times\; \hat{\sigma}_{ij}(p_1,p_2,\alpha(\mu_F^2,\mu_R^2),Q^2/\mu_F^2), \tag{2.3}$$

where the sum runs over all parton types, f_i are their PDFs and $\hat{\sigma}_{ij}$ is the parton-level cross section for incoming partons with momenta $p_1 = x_1 P_1$ and $p_2 = x_2 P_2$. The PDFs are function of the μ_F such that any parton emitted with smaller transverse momentum than the factorization scale is considered part of the hadron structure, and it is absorbed into the parton distributions. The partonic cross section is calculated at a given order in α_s which introduces a dependence on a renormalization scale μ_R, usually chosen to be equal to μ_F.

2.3 Simulation of *pp* collisions

The Monte Carlo (MC) simulation method is applied to describe hadron-hadron collisions. It uses pseudo-random numbers to simulate the full evolution of high energetic particle collisions. Comparisons between the simulated MC and real data events are crucial to understand the physics behind the collisions.

The simulation of a proton-proton collision is performed in steps according to the energy involved in the different processes. It starts simulating a relatively simple interesting sub-process in a collision with a larger momentum transfer. This process is known as hard process and is calculated using pQCD. Soft and collinear emissions are added to the resulting partons of the hard process to form a parton shower. When the energy of the partons is small enough, they are recombined into hadrons. The simulation finishes adding the processes involved between the proton remnants partons from the protons that do not take part in the hard interaction. Figure 2.3 illustrates the full simulation chain.

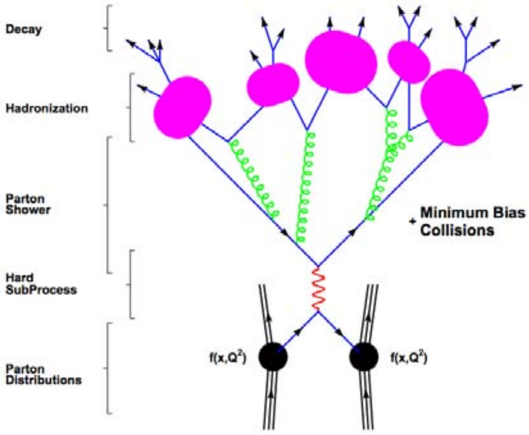


Figure 2.3: Illustrative representation of the different processes involved in a *pp* collision simulation [23].

2.3.1 Hard process

As previously mentioned, the event simulation begins with the hard process. Its cross section is calculated integrating the squared matrix element (ME) over the phase space. Eq. 2.3 can be rewritten to express the hard process cross section as

$$\sigma(P_1, P_2) = \sum_{i,j} \int dx_1 dx_2 \int d\Phi \; f_i(x_1, \mu_F^2) f_j(x_2, \mu_F^2) \; \times \; \frac{1}{2x_1x_2s} |M_{12}|^2(\Phi, \mu_R, \mu_F), \tag{2.4}$$

where it is introduced the dependence with the phase space Φ, the parton flux $\frac{1}{2x_1x_2s}$, being s the center of mass energy, and the ME M_{12} which can be written as the sum of Feynman diagrams used in the computation:

$$M_{12} = \sum_i F_{12}^{(i)}. \tag{2.5}$$

For a typical process, if only tree-level diagrams without internal loops are used in the ME the cross section is computed at Leading Order (LO). If higher order diagrams including 1-loop corrections are used, the computation is performed at Next-to-Leading-Order (NLO). Equivalently, the Next-to-Next-to-Leading-Order (NNLO) includes 2-loop diagrams in the ME.

2.3.2 Parton shower

The parton shower (PS) consists of successive soft collinear emissions of quarks and gluons out of the partons involved in the hard process. It is used to approximately simulate the higher orders processes that can not be described using fixed order in pQCD. It uses two simple phenomenological ideas, the iterative splitting of partons via the processes $q \to gq$, $g \to gg$ and $g \to q\bar{q}$ to form multiparton final states, and a proper cancelation of real and virtual divergences, giving physical results. Using both approximations together, a reliable answer to describe the PS is given. Figure 2.4 illustrates how the iterative splitting of the partons forms a parton shower.

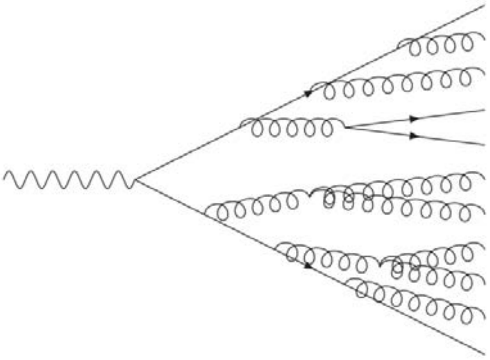


Figure 2.4: A cascade of successive splitting of partons [24].

The splitting of partons is performed using the DGLAP splitting functions describe in Section 2.1. It is iteratively applied until the parton energies reach the hadronization scale ($Q_0 \sim 1\,\mathrm{GeV}$) where confinement effects and hadronization phenomena dominate.

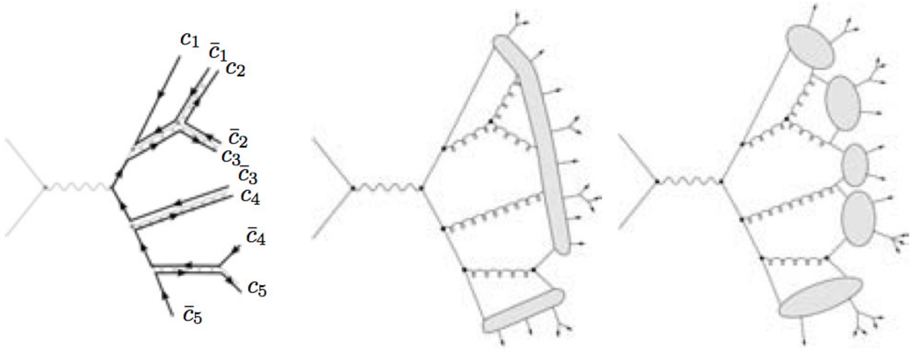
The virtual corrections are taken into account with the Sudakov factors [25]. These factors are derived from the splitting functions and include the corrections in a probability of the parton to evolve in time without splitting.

Depending if an incoming or an outcoming parton from the hard process initiates the shower, the radiation is classified as initial state radiation (ISR) or final state radiation (FSR), respectively. Algorithms used to generate the showers are the same in both cases but for the ISR they are operated backwards in time. This means that the fraction momentum of the initial parton is increased until it matches the PDFs x value. One of the limitations of this approach is that potential interference between ISR and FSR are neglected.

2.3.3 Hadronization

The hadronization is the process by which, after showering, partons at the $Q_0 \sim 1\,\mathrm{GeV}$ scale are recombined in colorless hadrons. The parton-hadron duality assumes that no high momentum transfer is needed in the recombination, as it happens close in phase space. Therefore, hadronization do not change significantly neither the flavor nor the kinematic of the initial partons conserving the event energy flow. The parton-hadron duality is an observed QCD property. Its simulation relies on QCD-inspired phenomenological models where their parameters are tuned using experimental data.

The most used hadronization models are the string and the cluster models (see Figure 2.5). The string model ties a quark with an antiquark (plus a number of intermediate gluons) and it is assumed that a "string" between them produces a linear confinement potential. If the distance between the pair of quarks is so long that the string energy exceeds the mass of a pair of quarks, a color-singlet pair is generated and the string is split into two less energetic strings. The cluster model is based on the pre-confinement of the branching processes. This means that at scales much smaller than the hard process scale, the partons in a shower are clustered in color-singlet groups independently of the nature and scale of the hard process. Therefore final state gluons are forced to split into a quark-antiquark pair. Then neighboring quarks and antiquarks (not from the same gluon) are paired together in color-singlet clusters.



(a) (b) (c)

Figure 2.5: Illustrative color flux parton shower graph (a) completed with hadronization from the string model (b) and the cluster model (c) [26].

2.3.4 Underlying event and pile-up

The underlying event (UE) refers to the interactions between the proton remnants that do not take part in the hard interaction. They happen at low transferred momentum and involves flavor- and color-connection with the hard process, therefore they can not be described using perturbation theory. An important contribution to the UE are multiple parton interactions, relatively hard secondary collisions between the incoming hadron remnants. This is simulated using low-p_{T} minimum bias interactions. The simulation parameters of the UE are tuned from experimental data.

Finally, one has to consider pile-up, due to multiple proton-proton interactions per bunch crossing at the collider. It is reproduced in simulation using minimum bias events as well. These events are then overlaid on top of the hard scattering event.

2.4 Monte Carlo generators

Monte Carlo generators are used to perform the simulation chain described in Section 2.3. The following list briefly describes the ones used in this analysis:

- PYTHIA is a multi-purpose MC generator that uses ME to compute at LO 2-to-n (n$\leq$ 3) parton processes and PS to account for the initial- and final-state radiation. The hadronization step is simulated using the string model. There are two versions of the PYTHIA generator, version 6 [27] written in `Fortran` and version 8 [28] rewritten in `C++`.
- HERWIG is a multi-purpose MC generator program that uses ME to compute at LO $2 \to 2$ parton processes and simulates also the PS. It uses the cluster model for hadronization and it is usually interfaced with JIMMY [29] to simulate the UE. There are two versions of the program, HERWIG [30] written in `fortran` and HERWIG++ [31] written `C++`.
- SHERPA [32] is a multi-purpose MC generator that uses ME to simulate $2 \to n$ parton processes at LO. The different jet multiplicities are generated in a single inclusive sample. It is interfaced with PYTHIA to simulate the PS. To cover the whole phase space with a smooth transition from ME to PS, and in order to avoid overlaps, a ME-PS matching is performed with the Catani-Krauss-Kuhn-Webber (CKKW) method [33]. The SHERPA generator uses the cluster model to simulate the hadronization and has its own UE implementation.
- POWHEG [34, 35, 36] is a MC generator that simulate events using ME at NLO in pQCD and usually interfaced with other generators like PYTHIA or HERWIG for modeling the PS, hadronization and UE.
- ACERMC [37] is another MC event generator specialized in the simulation of the Standard Model background processes in *pp* collisions at the LHC. It uses ME at LO and generally is interfaced with other generators like PYTHIA or HERWIG for the modeling of PS, hadronization and UE.
- ALPGEN [38] is a MC generator specialized in $2 \to n$ ($n \leq 9$) parton processes computed with ME at LO. The various parton multiplicities

processes are generated in individual samples and then combined into an inclusive one. It is interfaced either with PYTHIA or HERWIG for the PS development. It also implements a ME-PS matching (MLM matching [39]) to avoid overlaps.

- MC@NLO [40] is specialized MC generator used to simulate ME processes at NLO in pQCD. It is generally interfaced with HERWIG to simulate the PS.

Chapter 3

Statistical Model

A key aspect of this analysis is to define the statistical treatment in order to interpret the results. This Chapter describes the statistical tools and methodology used in all LHC Higgs searches.

3.1 Hypothesis testing

In particle physics, the procedure used to discovery or exclude a new physic model is normally done via a statistical test. Two hypotheses, one which describes the known processes and another which in addition includes the new phenomena, are tested one against the other. This applies to the Higgs searches as follows:

H_0 or Null Hypothesis corresponds to the Standard Model without the presences of a Higgs boson. It is often referred to as the background-only (B) hypothesis.

H_1 or Test Hypothesis corresponds to the Standard Model with the presences of the Higgs boson. For this reason, it is often referred to as the signal+background (S+B) hypothesis.

The compatibility between the observed data and a hypothesis H is given by a p-value. The computation of this probability is based on a test statistic. In LHC Higgs searches, the likelihood ratio is the test statistic choice due to some good properties. It is constructed from the number of signal and background events in the analysis regions. A hypothesis is rejected if the p-value is observed below an a-priori specified threshold. This lead to the discovery of new particle or the setting of 95 % Confidence-Level (CL) limits to its production rate.

3.2 Likelihood function and profile likelihood ratio

This Section is focused on building the likelihood function from the distributions (histograms) provided by an analysis. In this case, each histograms can be represented as $n = (n_1, ..., n_N)$. The expectation value E of each n_i can be written as:

$$E[n_i] = \mu s_i + b_i, \tag{3.1}$$

where s_i and b_i correspond to the number of signal and background events, respectively, of the ith bin of an histogram, and μ is the parameter of interest (POI), also known as signal strength. This parameter multiplies the Standard Model Higgs boson cross section times branching ratio value. Therefore, a $\mu = 0$ is associated with the background-only hypothesis (H_0) while a $\mu = 1$ denotes the present of a signal, and is associated with signal+background hypothesis (H_1). The number of signal and background events expected in the bin i can be expressed as:

$$s_i = s_{tot} \int_{bin\, i} f_s(x; \theta_s) dx, \tag{3.2}$$

$$b_i = b_{tot} \int_{bin\, i} f_b(x; \theta_b) dx. \tag{3.3}$$

The functions $f_s(x; \theta_s)$ and $f_b(x; \theta_b)$ are the probability density functions (pdfs) of the discriminant variable x, while θ_s and θ_b represent parameters that characterize the shapes of the pdfs. The s_{tot} and b_{tot} account for the total number of signal and background events. In the following definitions $\theta = (\theta_s, \theta_b, b_{tot})$ will be used for simplification to denote all nuisance parameters (NPs). The signal normalization s_{tot} is a fixed value extracted from the signal model. Section 3.3 presents a detail description of the NPs.

In addition to the measured histogram n, auxiliary measurements to help constrain the NPs are considered. These measurements can be represented also as histogram bins $m = (m_1, ..., m_M)$, where the expected values E of m_i can be written

$$E[m_i] = u_i(\theta). \tag{3.4}$$

Depending on the characteristics of θ, different pdfs can be used to describe them and calculate u_i.

All this analysis information is gathered in a likelihood function built as the product of Poisson probabilities for all bins:

$$L(\mu, \theta) = \prod_{i=1}^{N} \frac{(\mu s_i + b_i)^{n_j}}{n_j!} e^{-(\mu s_i + b_i)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k} \tag{3.5}$$

To perform the statistical interpretation, the set of NPs θ is profiled, meaning that they are functions of μ, $\theta(\mu)$. Two estimators are constructed from the same likelihood in the so-called profile likelihood ratio:

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}. \tag{3.6}$$

The numerator is μ dependent, thus $\hat{\hat{\theta}}(\mu)$ correspond to the θ value that maximizes L for a specific μ, also called conditional maximum-likelihood (ML) estimator of θ. Instead, the denominator is not μ dependent and is known as the maximized unconditional likelihood function. There, $\hat{\mu}$ and $\hat{\theta}$ represent the best estimators obtained if no condition is required to the μ value. The presence of a signal is expected to produce only upward fluctuations of the background. Negative values of μ are not taken into account in the statistical interpretation of results by using a modified likelihood ratio:

$$\tilde{\lambda}(\mu) = \begin{cases} \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & \hat{\mu} \geq 0 \\ \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0. \end{cases} \tag{3.7}$$

3.3 Nuisance parameters

The nuisance parameters are any parameter used in the statistical model with the exception of the POI. The parameters account for systematic, statistical and normalization uncertainties on the measurements, with the exception of data statistics, that is the basic Poisson term in the likelihood (Eq. 3.5). To have a factorized form of the likelihood, the θ parameters are taken into account as uncorrelated.

From the fitting point of view, the NPs are used to accommodate any possible variation on the signal (Eq. 3.2) and background (Eq. 3.3) expectation pdfs when the function is maximized. In some cases an extra term is introduced in the likelihood to account for variations of the NPs (Eq. 3.5).

Every NP is described with a pdf $\rho(\theta)$ and, in general, characterized by its best associated estimate $\hat{\theta}$ and extra parameters to describe its width or shape. The election of the pdf depends on the characteristics of the NP. The ones used in the analysis are described as follows:

- The flat pdf is used if no constraint exists on a particular parameter. In general no extra term is included in the likelihood, except if physical boundaries on the NP are set. In that case, the pdf is a rectangle whose limits are the ones set. These NPs are often referred to as freely floating parameters.

- The Gaussian pdf is a common choice for shape systematics uncertainties. It describes uncertainties on parameters that can be both positive and negative:

$$\rho(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\theta - \hat{\theta})^2}{2\sigma^2}\right), \tag{3.8}$$

This pdf is not suitable for positively defined observables.

- The log-normal pdf is used for parameters bounded to positive values:

  $$\rho(\theta) = \frac{1}{\sqrt{2\pi}\ln(\kappa)} \exp\left(-\frac{(\ln(\theta/\hat{\theta}))^2}{2(\ln(\kappa))^2}\right)\frac{1}{\theta}. \tag{3.9}$$

  The parameter κ characterizes the width of the log-normal distribution which tends to a Gaussian form for small values of κ as shown in Figure 3.1 (left). Typically background normalizations can be constrained using such pdfs.

- The Gamma pdf is applied for describing statistical uncertainties associated with the number of selected Monte Carlo events. The event rate n in the a certain region is related to the number of events N in MC using the relation $n = \alpha \cdot N$. The gamma distribution as a function of these variables is expressed as follow:

  $$\rho(n) = \frac{1}{\alpha}\frac{(n/\alpha)^N}{N!} \exp\left(-\frac{n}{\alpha}\right). \tag{3.10}$$

  Figure 3.1 (right) presents the gamma distribution for samples with different number of events.

3.4 Discovery and upper limit

When searching for a new particle, a discovery is claimed if a large enough incompatibility between data and the null hypothesis is observed. If no deviation is observed, an upper limit on the test hypothesis is extracted. The profile likelihood ratio $\tilde{\lambda}(\mu)$ (Eq. 3.7) ranges from $0 < \tilde{\lambda} < 1$ with $\tilde{\lambda}$ close to 1 implying good agreement between the data and the hypothesized value of μ. Although it is more commonly used the form $-2\ln\tilde{\lambda}(\mu)$ for reasons explained in the following Sections.

3.4.1 Test statistics q_0 and q_μ

The procedure to discover a new signal is to reject the background-only hypothesis. The statistic to do so is q_μ when $\mu = 0$ defined as:

$$q_0 = \begin{cases} -2\ln\tilde{\lambda}(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0. \end{cases} \tag{3.11}$$

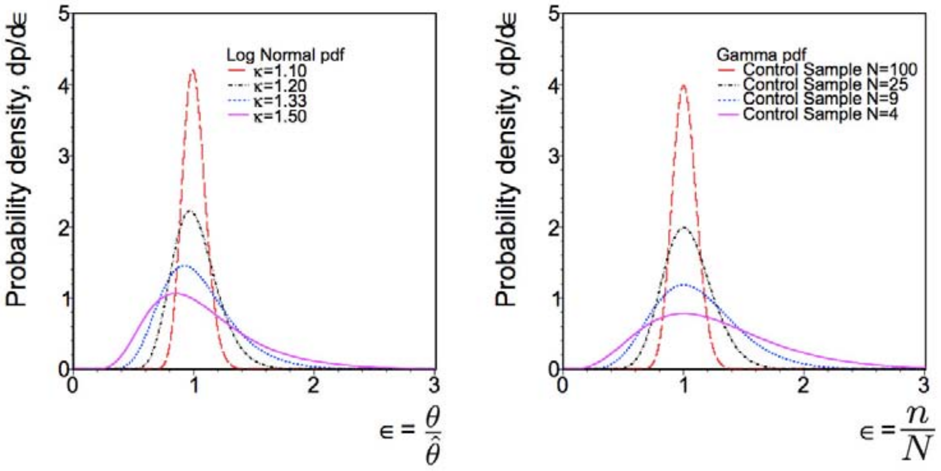


Figure 3.1: (Left) Log-normal distributions with $\kappa = 1.10$, 1.20, 1.33 and 1.50. (Right) Gamma distribution with the number of events in a sample $B = 100, 25, 9$ and 4. Ref. [41].

An increase of event yields above the expected background translates into an increase of $\hat{\mu}$, thus $\hat{\mu} \geq 0$ means that a new signal could be present. If this is the case, larger q_0 values will correspond to an increase level of incompatibility between data and the background-only hypothesis. Since only upward fluctuations of the background are considered, by construction if $\hat{\mu} < 0$ then $q_0 = 0$, indicating a compatibility with the background-only hypothesis.

If no discovery has been made, one can define upper limit on the signal strength parameter μ. The test statistic is defined differently compared to the discovery case:

$$q_\mu = \begin{cases} -2\ln\tilde{\lambda}(\mu) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(0,\hat{\hat{\theta}}(0))} & \hat{\mu} < 0 \\ -2\ln\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} & 0 \leq \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu. \end{cases} \tag{3.12}$$

In this case, q_μ is set to 0 if $\hat{\mu} > \mu$ because no upper limit could be set if the observed signal strength is larger than the tested one. As for the discovery case, higher values of q_μ represent greater incompatibility between the data and the tested value of μ.

3.4.2 Probabilities p_0 and p_μ

The compatibility of data with the background-only hypothesis is given by the p-value:

$$p_0 = \int_{q_{0,obs}}^{\infty} f(q_0|0, \hat{\hat{\theta}}(0))dq_0, \tag{3.13}$$

where $q_{0,obs}$ is the observed value of the test statistic and $f(q_0|\mu)$ denotes the q_0 pdf under the assumption of the signal strength $\mu = 0$ and $\hat{\hat{\theta}}(0)$. In the presence of a signal, the p_0 is generally converted to the corresponding Gaussian significance, Z, defined as the Gaussian distributed variable with upper-tail probability equal to the p_0 which correspond to Z standard deviations above its means. That is $Z = \Phi^{-1}(1 - p_0)$, where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard Gaussian. In particle physics, it has been adopted the convention that if the background-only hypothesis is rejected with a $p_0 = 1.3 \times 10^{-3}$, equivalent to a significance $Z = 3$, an evidence is announced. A discovery is claimed for $Z = 5$, corresponding to $p_0 = 2.9 \times 10^{-7}$.

If no signal is present in data, the signal+background hypothesis can be excluded with a certain Confident-Level, and an upper limits can be set on its cross section. This is done by calculating the probability that the q_μ test statistic for a given signal strength μ assumes a value equal or higher than the observed value $q_{\mu,obs}$ using the p-value expression:

$$p_\mu = \int_{q_{\mu,obs}}^{\infty} f(q_\mu|\mu, \hat{\hat{\theta}}(\mu))dq_\mu. \tag{3.14}$$

The upper limit μ_{up} at 95 % CL is defined as the highest value of the signal strength μ for which the probability p_μ is still higher or equal to 5 %.

3.4.3 CL_S method

When downward fluctuations on the observed number of background events are produced, upper limits computed using p_μ can lead to unphysical exclusion of small μ values to which the search is not a priori sensitive. This is overcome in the analysis by using the CL_S method [42] defined as a ratio of probabilities

$$CL_S = \frac{p_\mu}{1 - p_b} = \frac{p_{s+b}}{1 - p_b}, \tag{3.15}$$

where p_{s+b} and p_b quantify the compatibilities between the data and the signal+background and background-only hypotheses, respectively. The p_b is defined as

$$p_b = 1 - \int_{q_{\mu,obs}}^{\infty} f(q_\mu|0, \hat{\hat{\theta}}(0))dq_\mu. \tag{3.16}$$

A downward background fluctuation in data will lead to small values of $1-p_b$ increasing the CL_S upper limit and avoiding the exclusion of too small cross sections. In LHC Higgs searches CL_S upper limits at 95 % confidence level are obtained by computing the μ for which CL_S gives 0.05.

Figure 3.2 shows the pdfs for the test statistics q^1. It also illustrates the computation of the signal+background and the background-only hypotheses probabilities used in the CL_S method for an observation q_{obs} obtained from data.

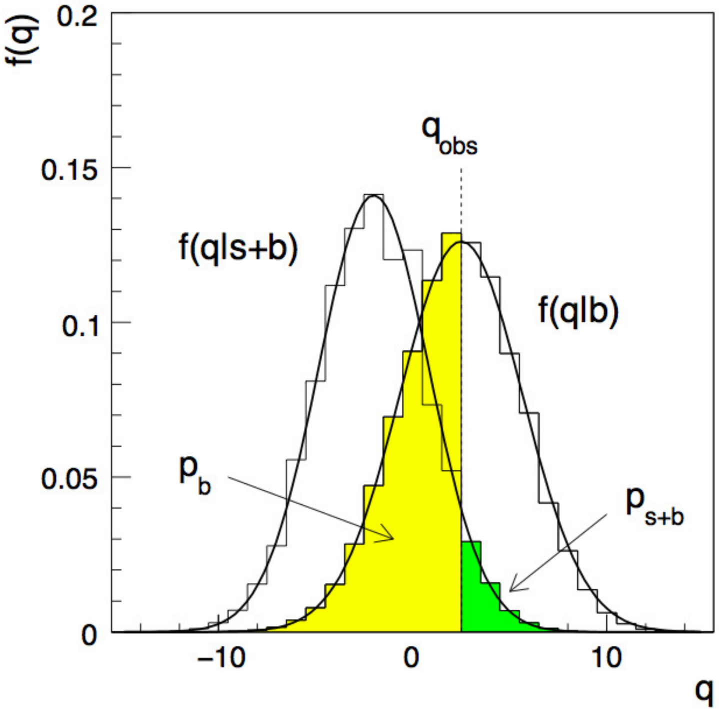


Figure 3.2: The distribution of the statistic q under the hypotheses signal+background and background-only and q_{obs} obtained from data. Green and yellow areas correspond to p_{s+b} and p_b probabilities defined with Eq. 3.14 and Eq. 3.16 integrals respectively. This illustrative Figure uses the Tevatron Collider test statistic which is defined with a different likelihood ratio. Figure obtained from Ref. [43].

---

[1] In this case q represents the Tevatron Collider test statistic which is defined using a different likelihood ratio.

3.5 Approximate sampling distributions and the Asimov data set

The computation of a p-value associated with a hypothesis needs the full distribution of the test statistic as shown in Eq. 3.13 and Eq. 3.14. In particular, when trying to make a discovery, the background only hypothesis needs to be tested using $f(q_0|0,\hat{\hat{\theta}}(0))$. Similarly, when computing an upper limit using the CL_S method, the $f(q_\mu|\mu,\hat{\hat{\theta}}(\mu))$ to describe the q_μ distribution is also needed.

The estimation of q_μ and q_0 distributions could be done with Monte Carlo methods. These methods are computationally heavy specially when calculating upper limits. As an example, if a discovery with $p_0 \sim 10^{-7}$ wants to be claimed, an order 10^8 pseudo-experiments toys would need to be simulated. For this reason, an approximation [43] valid in the large sample limit is normally used to describe the profile likelihood ratio instead. If the assumption is that the data are distributed according to a strength parameter μ', one can write using Wald's approximation [44]

$$-2\ln\tilde{\lambda}(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}), \tag{3.17}$$

where $\hat{\mu}$ follows a Gaussian distribution with a mean μ' and standard deviation σ, and N accounts for the data sample size. The value of σ is estimated from an artificial data set known as 'Asimov data set' [43].

The Asimov data set takes its name after I. Asimov's short story *Franchise*, where one most representative voter was chosen to act as the whole electorate. To construct it, the pseudo-data is forced to be equal to its expectation value $n_{i,A} = E[n_i] = \mu' s_i(\theta) + b_i(\theta)$ and all NPs are set to the values obtained from the best fit to data.

The test statistics q_μ and q_0 can be rewritten using Eq. 3.17 approximating the pdfs $f(q_\mu)$ and $f(q_0)$ to χ^2 distributions [43]. This asymptotic approximation of the profile likelihood ratio works very well and reduces the computational time since calculations are done analytically. An example of distributions obtained with this method is shown in In Figure 3.2 where the histograms are from Monte Carlo and the solid curves are the predictions of the asymptotic approximation.

In Ref. [43] approximate equations are derived for the significance and upper limits using the asymptotic approximation. The significance can be written as

$$Z = \sqrt{q_0}, \tag{3.18}$$

and the upper limit on μ can be written as

$$\mu_{up} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha), \tag{3.19}$$

where σ is the standard deviation on μ, Φ^{-1} the inverse of the cumulative distribution of the standard Gaussian, and $1 - \alpha$ the confident level to be excluded.

3.6 Analysis sensitivity

It is useful to know the sensitivity of the analysis, what can be done calculating its expected significance and upper limit. Expected results are calculated in the same way as the observed results by replacing the data set with the Asimov pseudo-data set.

When computing the expected significance the pseudo-data is usually constructed with a $\mu' = 1$ to obtain the results simulating the presence of a signal. This pseudo-data are used to calculate the value of the test statistic $q_{0,A}$ and with Eq. 3.18 obtained the expected significance of the analysis

$$Z_{0,exp} = \sqrt{q_{0,A}}. \tag{3.20}$$

To set expected upper limits, the pseudo-data is usually constructed with a $\mu' = 0$ to obtain the results simulating the background-only hypothesis. Using Eq. 3.19 the expected upper limit can be computed with the following expression

$$\mu_{up,exp} = \mu' + \sigma\Phi^{-1}(1 - \alpha). \tag{3.21}$$

When presenting expected upper limit results, $\pm 1\sigma$ and $\pm 2\sigma$ bands are usually calculated as well. The expression used is

$$band_{N\sigma} = \mu' + \sigma(\Phi^{-1}(1 - \alpha) \pm N), \tag{3.22}$$

where N correspond to the σ band $(N = -2, -1, 1, ...)$.

Chapter 4

The LHC and the ATLAS detector

4.1 The LHC

The Large Hadron Collider (LHC) is a particle accelerator placed at CERN (Centre Européen pour la Recherché Nucléaire). The machine lies in a circular tunnel of 27 Km long and an underground depth that can reach 175 m under the Franco-Swiss boarder near Geneva. The LHC has four collision points where experiments are placed to record and analyze the particle collisions. The four main detector experiments are: ATLAS [45], CMS [46], ALICE [47] and LHCb [48]. Figure 4.1 shows the LHC ring with its four main experiments.

Protons are accelerated at the LHC up to energies of 4 TeV but it is designed to reach 7 TeV per proton beam when it will be operating at its full potential. Protons are not individually accelerating but collected in bunches of protons. The acceleration and the configuration of the proton beams is performed in various steps. The process starts extracting the protons from a hydrogen gas and injecting them into the first linear accelerator LINAC2 where they are accelerated up to 50 MeV. Then they enter the Proton Synchrotron Booster (PSB) where their energy is increased up to 1.4 GeV and the bunches are formed. Next step is the Proton Synchrotron (PS) which brings the energy to 25 GeV before passing them to the Super Proton Synchrotron (SPS) where the energy is further increased up to 450 GeV. Finally, the bunches are injected into the LHC and the energy is increased before the collisions start. Figure 4.2 illustrates the CERN accelerator complex and the main experiments.

The LHC is constructed with 15 m-long 1232 superconducting dipole magnets which bend the beams to keep them in the ring. Moreover, 392 superconducting quadrupoles 5-7 m long are placed in between dipoles to focus the beams. Additional types of magnets move and squeeze the beams

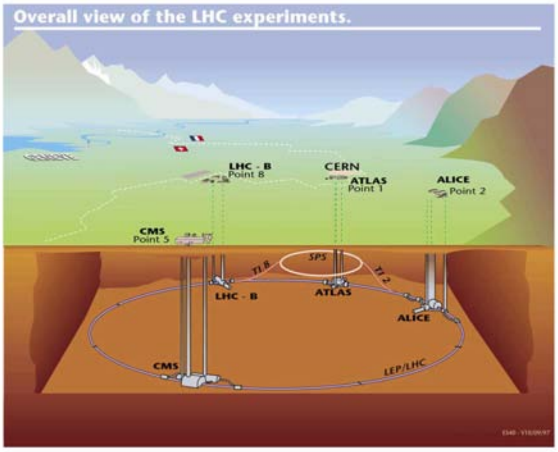


Figure 4.1: Schematic view of the LHC ring under the Franco-Swiss boarder and its four main experiments: ATLAS, CMS, ALICE and LHCb. CERN label represents the headquarters of the organization located in Meyrin, Geneva.

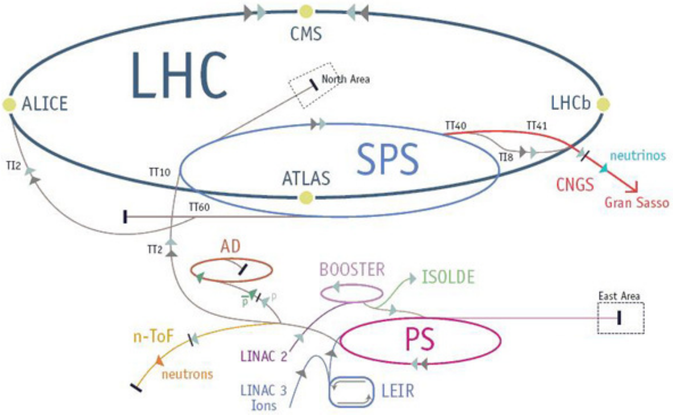


Figure 4.2: A schematic showing the accelerator complex at CERN.

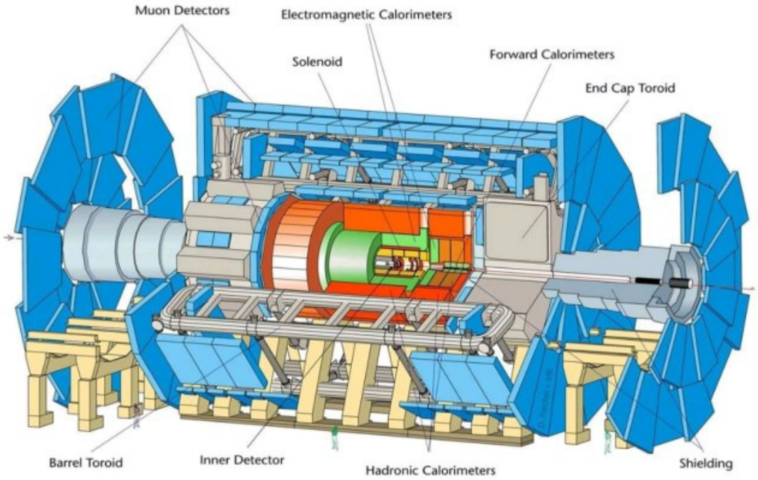


Figure 4.3: General overview of the ATLAS detector.

at the collisions points. Protons are accelerated at LHC by 16 superconducting radiofrequency cavities, 8 for each beam, place between the CMS and ALICE collision points. All these superconducting material operates at a temperature of 1.9 K, maintained with liquid helium.

4.2 The ATLAS detector

ATLAS (A Toroidal Lhc ApparatuS) [45] is a general purpose particle detector placed at the LHC. It has a cylindrical shape 44 m long, 25 m of diameter with a weight of approximately 7000 t (Figure 4.3). The detector is formed by several sub-detectors, each one of them with specific characteristics according to its functionality. The sub-detector closest to the interaction point is the inner detector, it tracks down charged particles in a strong magnetic field. Then is placed the electromagnetic and hadronic calorimeters where photons, electrons and hadrons are absorbed to measure their energy. Finally, the muon spectrometer wraps all previous sub-detectors in order to measure the momentum of the muons that cross the detector. ATLAS has a magnetic system which immerses the inner detector and the muon spectrometer in a 2 T solenoidal and in a 0.5-1 T toroidal magnetic fields, respectively. Figure 4.4 shows a cross sectional of the detector and illustrates how the different particles travel through it.

4.2.1 Coordinate system

The ATLAS coordinate system is such that the interaction point is defined as the origin. The beam direction defines the z-axis and, transverse to it, the x-y plane. The positive z-axis is defined by the anti-clockwise beam direction,

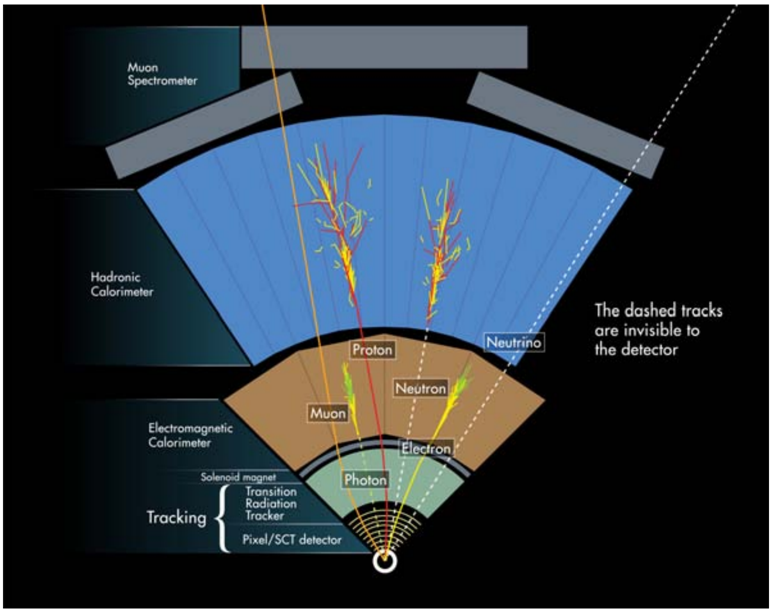


Figure 4.4: ATLAS cross sectional schematic showing particles travelling from the interaction point through the various sub-detectors.

the positive x-axis is defined as pointing from the center of the detector to the center of the LHC ring, and the positive y-axis points upwards. In addition, to the use of cylindrical coordinates, ϕ, θ, the experiment uses the rapidity, define as

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) . \tag{4.1}$$

The use of the rapidity is motivated by the fact that rapidity difference of two particles in the final state Δy is invariant against longitudinal boosts along the z-direction. For massless particles, the rapidity is equivalent to the pseudorapidity (η) define as

$$\eta = - \ln \tan \left(\frac{\theta}{2} \right) . \tag{4.2}$$

4.2.2 Inner detector

The Inner Detector (ID) [49] is the sub-detector closest to the interaction point. It is contained within a cylinder of 6 m length and 2 m diameter and it is the combination of three sub-detectors which permits to cover the range $|\eta| \leq 2.5$ (Figure 4.5). Its primary task is the precise reconstruction of the trajectories (tracks) of charged particles produced in the collisions. The ID is immersed in a 2 T solenoidal magnetic field. By combining the track and the magnetic field information it is possible to infer the p/q curvature

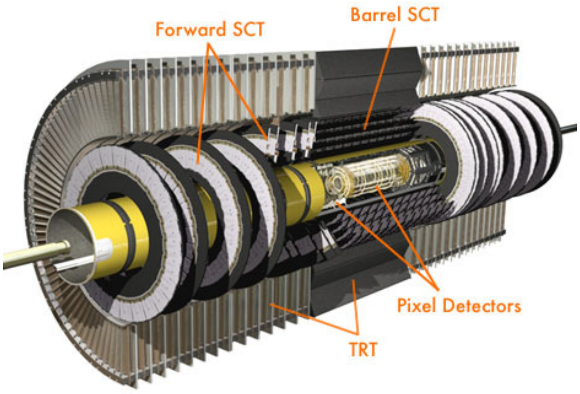


Figure 4.5: ATLAS Inner Detector layout.

and the point of closest approach of the particle trajectory to the beam line (the impact parameter) of the charge particles. Its expected momentum and transverse impact parameter resolutions are $\sigma_{p_T}/p_T = 0.05\%\ p_T \oplus 1\%$ and $10\,\mu$m, respectively. As the density of particles is higher closer to the interaction point, three sub-system with different technologies are used to reduce/avoid overlaps or ambiguities in the track reconstruction:

- The pixel detector: is the closest to the beam line, provides high-precision measurements around the interaction point, and allows secondary vertex identification from particle decays. It is formed by three barrel layers and three forward and backward end-cap disks which sum $\sim$80.4 million pixel channels with a size of $50 \times 140\,\mu\text{m}^2$.

- The Semi Conducting Tracker (SCT): is composed by silicon strip detectors distributed in four cylinders in the barrel region and nine disks in each end-cap region, providing $\sim$6.3 millions channels. A strip sensor alone can only provide a position measurement in one direction. Therefore, these sensors are attached back-to-back with a small angle of 40 mrad between them to solve any ambiguities and provide a space point.

- Transition Radiation Tracker (TRT): consists of thin drift tubes with a diameter of 4 mm. The tubes are distributed parallel to the beam axis in the barrel region and radially in the end-caps disks, giving 351 k read-out channels. This design guarantees that particles traversing the TRT will cross 35-40 straws in the region $|\eta| < 2$. The TRT provides discrimination between particles and contributes to the electron identification.

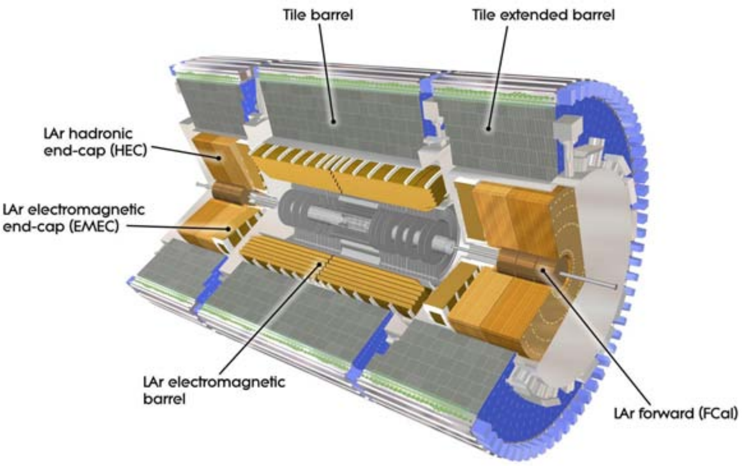


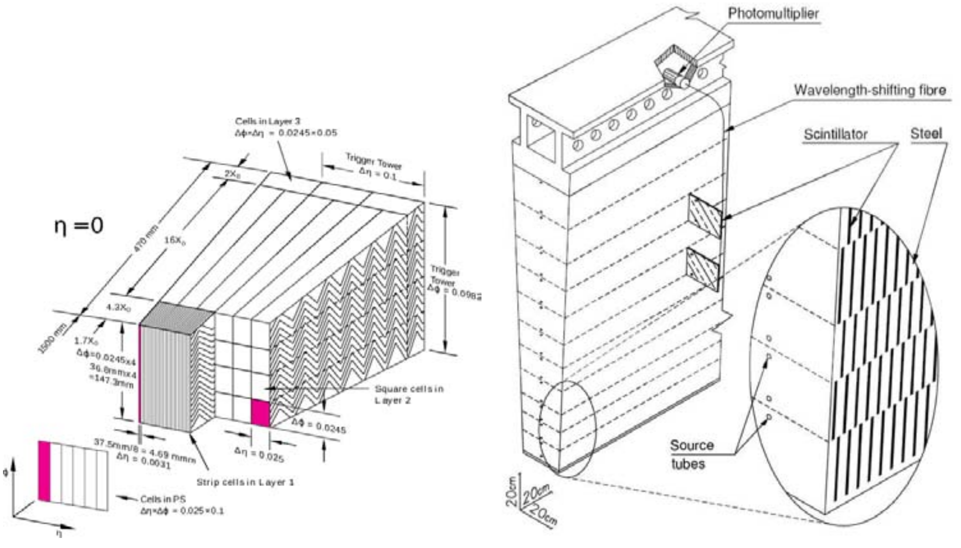
Figure 4.6: The ATLAS electromagnetic, hadronic and forward calorimeters layout.

4.2.3 Calorimeters

The calorimeters measure the position and energy deposited by electromagnetic (EM) and hadron showers. The methodology of the different technologies is to have alternating layers of absorbing and sampling material. The calorimeters are ϕ-symmetric covering a range up to $|\eta| < 4.9$. They are designed to capture the whole shower produced by an incident particle so its total energy can be inferred from the sum of the energy deposited in each sampling layer. The calorimeters also measure any imbalance in the total transverse energy to infer any weakly interacting neutral particle which carries away momentum unseen. Full containment of the showers is also need it to avoid punch-through of particles into the muon system. Figure 4.6 shows a layout of all the calorimeters installed in the ATLAS detector, as describe in the following Sections.

4.2.3.1 Electromagnetic calorimeter

ATLAS electromagnetic calorimeter (ECAL) [50] uses lead as absorber and Liquid Argon (LAr) as sampling material. It is designed to measure the position and energy of electrons and photons. High energy electrons that cross the calorimeter volume produce additional photons through the process known as Bremsstrahlung radiation. High energy photons instead, produce electron-positron pairs via pair production. This leads to a cascade which ionizes the LAr. The charge from ionization is collected in the readout electrodes made of copper and kapton. Collected charge is proportional to deposited energy. The ECAL energy resolution can be parameterized by



(a) (b)

Figure 4.7: Diagrams showing (a) the segmentation in three layers of a electromagnetic calorimeter barrel module. The cell and trigger tower granularity are shown as well as the radiation length of the different layers. (b) Schematic of a module of the hadronic barrel calorimeter TileCal.

$\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$.

The electromagnetic calorimeter is divided into the barrel (EMB), which covers a range up to $|\eta| < 1.5$, and the end-caps (EMEC) which covers the range $1.4 < |\eta| < 3.2$ (Figure 4.6). To increase the coverage to the region $3.1 < |\eta| < 4.9$, the first layer of the forward calorimeter (FCal) is optimized to make EM measurements (see Section 4.2.3.3). The EMB and EMEC are segmented in four sections in depth: the presampler and 3 additional layers. The presampler, installed in the region $|\eta| < 1.8$, is used to correct for the energy lost in the material before the calorimeter. The first layer contributes to the particle identification and precise position measurement. The second layer is where most of the shower is contained. Finally, the third layer is used to correct for the tail of very highly energetic EM showers. The granularity and the radiation length (X_0) of the cells depends on the layer and their η position, having a typical cell size of $\Delta\eta \times \Delta\phi = 0.025 \times 0.025$ and a radiation length higher than $22\,X_0$. The fast read-out system installed allows to use ECAL signals to trigger interesting events in the $|\eta| < 2.5$ range. An illustration of a EM barrel module is shown in Figure 4.7(a).

4.2.3.2 Hadronic calorimeter

The hadronic calorimeter (HCAL) is composed by various technologies. In the central region, the barrel calorimeter is made of scintillating tiles as active material and lead as absorber. This part is called TileCal calorimeter [51] and covers the region $|\eta| < 1.7$. Covering region $1.5 < |\eta| < 3.2$ there is the hadronic end-cap calorimeter (HEC) [50]. The HEC uses copper as passive material and LAr as active material, chosen for the radiation hardness needed in the very forward region. The HCAL energy resolution can be parameterized by $\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$.

TileCal is divided in a barrel ($|\eta| < 1$) and two extended barrels ($0.8 < |\eta| < 1.7$), and further divided into three layers. In the TileCal, particles interact with the lead tiles provoking the showers that illuminate the scintillating tiles. The light is collected by optic fiber and finally guide to a photomultiplier (see Figure 4.7b). Each photomultiplier reads a bunch of fiber corresponding to a cell of $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ in the first two layer and $\Delta\eta \times \Delta\phi = 0.2 \times 0.1$ in the third one. The interaction length (λ) varies depending on the layer being 1.4, 4.1 and 1.8λ in the barrel and 1.5, 2.6, 3.3λ in the extended barrel. The granularity in the HEC varies with η ranging between $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ and 0.2×0.2 and it has a interaction length of about 10λ.

4.2.3.3 Forward calorimeter

The Forward Calorimeter (FCal) [50] covers the region $3.1 < |\eta| < 4.9$, improving the calorimetry coverage to very forward regions. Each end-cap consists in three modules of about 10λ, where LAr is used as active material. The first module is optimized to measure EM measurements using copper as absorber. The other two modules use tungsten as absorber. The FCal energy resolution can be parameterized by $\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$.

4.2.4 Muon spectrometer

The muon spectrometer (MS) is the outermost part of ATLAS. Since muons mostly interact with matter like minimal ionizing particles, they traverse the inner detector and the calorimeters relatively undisturbed and a track is observed in the MS. To resolve the muon momentum with large accuracy, the system is immersed in a magnetic field which is produced by one barrel and two end-caps air-core toroids, providing on average 0.5 T and 1.0 T respectively. The MS provides momentum measurement with a resolution better than 3 % over a wide p_{T} range and up to 10 % for $p_{\mathrm{T}} \approx 1$ TeV.

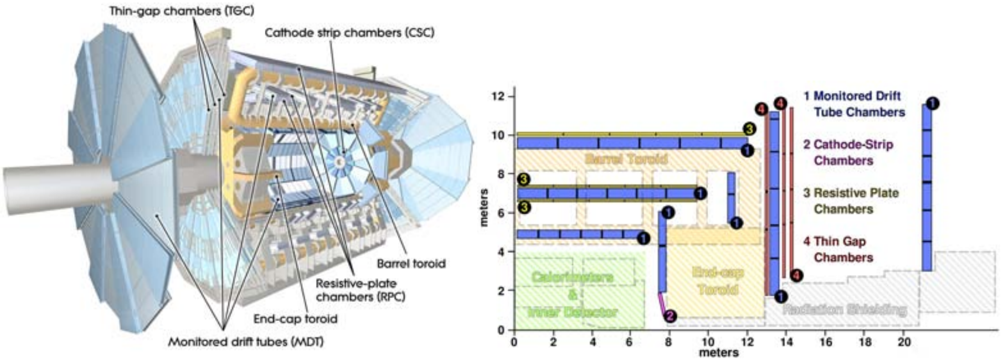
Four different gaseous detector technologies are used for high-precision tracking measurements and trigger. Figure 4.8 shows how the technologies are distributed in ATLAS.

- Monitored Drift Tubes (MDT): cover the $|\eta| < 2.7$ range and provide a precision measurement of the track coordinates.

- Cathode Strip Chambers (CSC): multiwire proportional chambers with cathodes segmented into strips. It provides precision measurements in the forward region $2.0 < |\eta| < 2.7$. Since the modules are located closer to the collisions point, they have higher rate capability, and time resolution compared to the MDTs.

- Resistive Plate Chambers (RPC): consist of parallel electrode-plates which cover the barrel region $|\eta| < 1.05$. They have a fast time respond to be used by the trigger system and enough spacial resolution to deliver extra tracking information.

- Thin-Gap Chambers (TGC): operate on the same principle than the RPC but in the end-cap region $1.05 < |\eta| < 2.7$, providing trigger only up to $|\eta| < 2.4$. They have a better granularity giving better rate capability.



(a) (b)

Figure 4.8: (a) Overview of ATLAS muon spectrometer. (b) Cross-sectional view of the detector showing the distribution of the different technologies used in the MS.

4.2.5 Minimum Bias Trigger Scintillators

The Minimum Bias Trigger Scintillators (MBTS) are segmented scintillator paddles located close to the beam-pipe, which trigger on the energy deposited by charged particles traversing the scintillator. During early running when the luminosity was very low, the random trigger installed was inefficient since the probability of an interaction during a bunch crossing

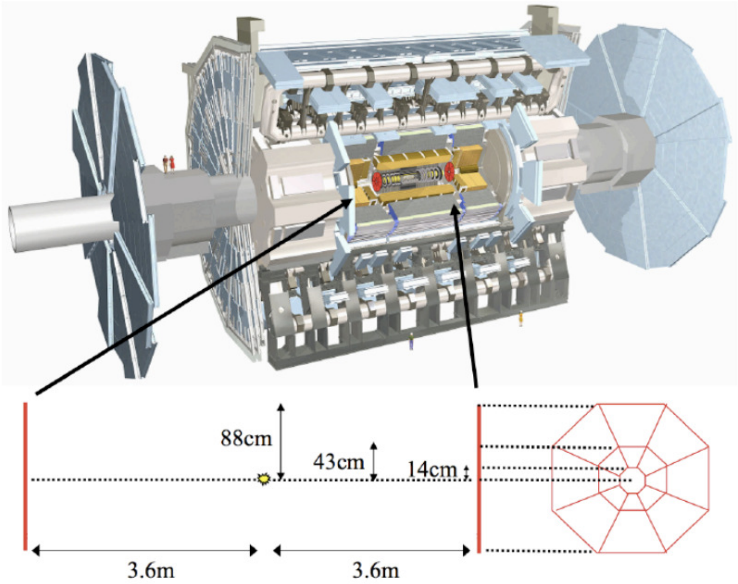


Figure 4.9: Overview and geometry of ATLAS MBTS.

was less than 1 % and MBTS was crucial for recording first physic collision events in ATLAS.

The MBTS detector consists of thirty-two scintillator paddles, each of which is 2 cm thick. The paddles are arranged into two disks and installed on the inner face of the end-cap calorimeters. Each disk has an inner and outer ring covering $2.82 < |\eta| < 3.84$ and $2.09 < |\eta| < 2.82$ respectively. The light emitted by each scintillator is transmitted through a wavelength-shifting optical fiber to a photomultiplier located in the TileCal hadronic calorimeter.

4.3 Luminosity

In ATLAS, the luminosity is determined indirectly using several detectors. These detectors measure the *pp* interaction rate per bunch crossing μ_{vis}, which is related to the luminosity with the expression [52]:

$$\mathcal{L} = \frac{\mu_{vis} n_b f_r}{\sigma_{vis}}, \tag{4.3}$$

where n_b is the number of bunches cross, f_r is the bunches revolution frequency (11245.5 Hz) and $\sigma_{vis} = \epsilon \sigma_{ine}$ is the total inelastic *pp* cross section within the detector acceptance ϵ. The n_b and f_r are known machine parameters, μ_{vis} is directly measured during collisions thus the only parameter which needs to be calibrated is σ_{vis}. The calibration is performed using scans of the beam separation, known as van der Meer scans [53]. An alternative expressing to inferred the luminosity from measured accelerator parameters during the scans can be used:

$$\mathcal{L} = \frac{n_b f_r n_1 n_2}{2\pi \Sigma_x \Sigma_y}, \tag{4.4}$$

where n_1 and n_2 are the number of protons in the colliding bunches and Σ_x and Σ_y give the extensions of the bunches perpendicular to the beam direction. The scan moves the beams in the horizontal and vertical plane to measure their profiles Σ_x and Σ_y. Simultaneously the μ_{vis} is measured and maxima in its profile are associated to zero beam separation. All the information can be combined, using Eq. 4.3 and 4.4, to estimate σ_{vis}:

$$\sigma_{vis} = \frac{\mu_{vis}^{max} 2\pi \Sigma_x \Sigma_y}{n_1 n_2}. \tag{4.5}$$

Instantaneous luminosity is integrated over time to account for the amount of collisions generated. Figure 4.10 presents the integrated luminosity delivered over time, recorded and certified to be good quality data, for *pp* collisions at 7 TeV and 8 TeV centre-of-mass energy in 2011 and 2012. The luminosity delivered accounts for collisions from the start of stable beams until the beam is dumped by LHC. The difference with the ATLAS integrated luminosity also accounts for the time to ramp-up HVs. The Data Quality (DQ) information finally decides on the fraction of data good for physics analysis. This leads to a total of 20.3 fb^{-1} at 8 TeV and 4.57 fb^{-1} at 7 TeV data.

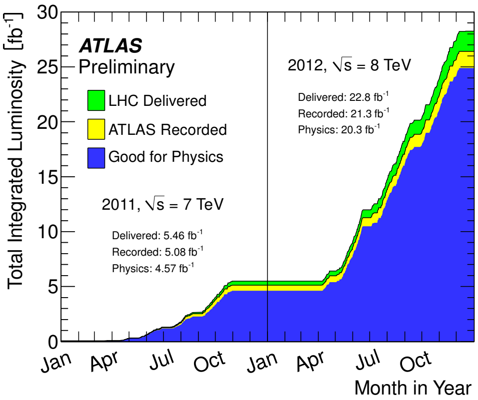


Figure 4.10: Cumulative luminosity versus time delivered to, recorded by ATLAS, and certified to be good quality data during stable beams and for *pp* collisions at 7 and 8 TeV centre-of-mass energy in 2011 and 2012.

While operating the ATLAS detector, a unit called lumiblock is introduced to handle event samples for practical matters. It can be considered as a quantum of data which roughly corresponds to $10 - 150\,\mathrm{nb}^{-1}$. When the machine runs smoothly, the operation time can be also expressed in number of lumiblocks taken. The correspondence of the lumiblock in seconds vary among the data periods since it depends on the beams configuration and the instantaneous luminosity.

4.4 Trigger system

In hadron colliders is impossible to store all the data generated by the huge amount of event collisions produced. At nominal luminosity the event rate reached in ATLAS is 1 GHz (10^9 events/s). Since the detector store rate is orders of magnitude smaller (∼400 Hz), a trigger system is needed to quickly decide whether a given event contains an interesting signature and should be stored or not. The trigger system is divided in three levels which progressively filter the events and reduces its rate to a storable level.

The first selection is made by the Level-1 (L1) trigger. The L1 is a hardware based trigger since a very fast response is needed it. It uses the muon detector and the calorimeters to measure signatures of high-p_{T} muons, energy deposits in the calorimeters and $E_{\mathrm{T}}^{\mathrm{miss}}$. The L1 also identifies Regions of Interests (ROI) corresponding to directions in the $\eta - \phi$ plane, where interesting features in the detector are observed, usually associated with high-p_{T} objects. The ROI information is stored in front-end pipelines pending for the L1 decision. Considering the pipeline buffer, the expected time to make a decision on the event, also known as latency, is $2.5\,\mu s$. The typical L1 input and output rates in 2012 were 20 MHz and 65 kHz, respectively.

Next step in the chain is the Level-2 (L2) trigger, a software base trigger with a longer processing time which allows to increase the complexity of the selection algorithm. This trigger is seeded with the ROI identified by L1. The L2 has a larger latency compare to L1, 100 ms, which makes possible to use data with full granularity and tracking information. The L2 output rate is 6.5 kHz

The final stage of the trigger system is the Event Filter (EF). The 1 s of processing time allowed to the EF makes possible to fully reconstruct the event using algorithms as those for the offline analysis. This reduces the event rate to a storable number of about 400 Hz, which corresponds to about 1 Gb/s data rate.

4.5 ATLAS detector simulation

The ATLAS detector simulation [54] models the detector material, its geometry and response to particles crossing based on the GEANT4 [55] program. The read-out electronic signals produced by the interaction with the different subsytems is also simulated. At this point data and simulated events use the

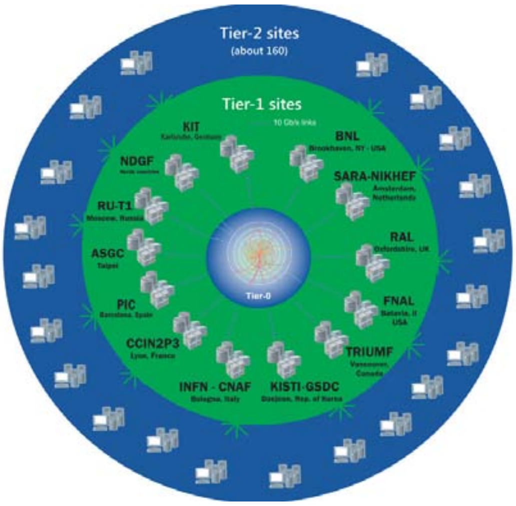


Figure 4.11: Diagram of the Worldwide LHC Computing Grid (WLCG) project with the hierarchy and connections of the different Tier sites.

same algorithms to reconstruct the physic objects such as electrons, muons or jets. The detector simulation is constantly updated with the information from the real detector conditions, e.g. dead channels or mis-alignments.

4.6 Computing facilities and data storage

Due to the huge amount of data generated by the ATLAS detector, estimated in few Petabyte per year, and the large number of physicists involved in their analysis, a Worldwide LHC Computing Grid (WLCG) [56] was designed. The LCG consists on a pyramid structure with a main computer center at CERN (the Tier-0 site), thirteen large computer centres (Tier-1 sites), and more than 160 Tier-2 and -3 sites in different institutions around the globe. The pyramid is designed to decrease the event size and detail level while increasing analysis specific information every time the datasets are transferred to a next Tier level.

Chapter 5

Physic objects reconstruction and identification

This Chapter is focused on the description of the reconstruction and the identification of the physics objects used in the $VH(H \to b\bar{b})$ analysis. This includes: the reconstruction and identification of charged particles (tracks); interaction vertices; electrons and muons; and jets and b-tagged jets. Finally, the reconstruction of the missing transverse momentum is discussed.

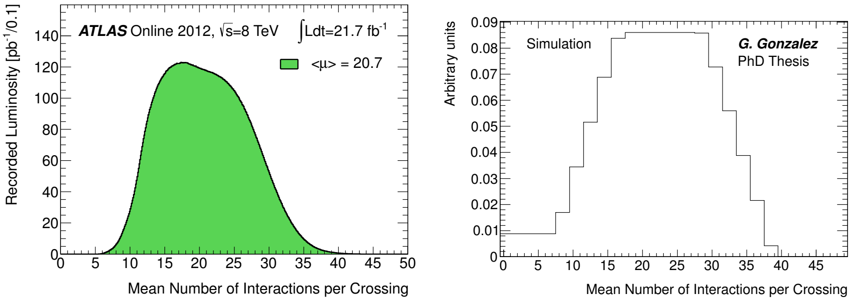
5.1 Track and vertex reconstruction

In ATLAS, tracks are reconstructed using the hit information from charged particles generated while passing through the ID. The reconstruction starts by transforming the hits information from pixel and the SCT into three dimensions space-points. At the same time, the TRT hit information is transformed into drift circles which correspond to the closest distance to a straw wire. The construction of track candidates starts by forming seeds from space-points from the three pixel layers and the first SCT layer. These seeds are then extended in search for more hits through the remaining layers of the SCT. Candidates are then fitted and fakes and ambiguities are removed [57]. Finally, the remaining tracks are extended to add the information from the TRT drift circles. This track-finding algorithm is called inside-out tracking. In addition, an outside-in track-finding algorithm is also performed. This second algorithm starts from the TRT seed and extends the tracks back to the inner SCT and pixels layers. It is used to better reconstruct tracks coming from secondary vertices from long-lived particles.

The detected tracks are then used to reconstruct the vertices (from the interaction points) and secondary vertices from particle decays. The procedure to reconstruct primary vertices (PVs) [58] starts by selecting tracks originated in the interaction region. A vertex seed is found based on a global maximum in the z distribution of the tracks. An algorithm which uses as input the seed and the tracks around it computes the vertex position [59]. This algorithm is rerun using as seeds close tracks that could not be associated with the already located vertices. This procedure is repeated until no unassociated tracks are left in the event or no additional vertex can be found. Vertices are required to have at least three associated tracks with $p_{\mathrm{T}} > 400\,\mathrm{MeV}$. From all PVs found, the primary interaction in the event is associated with the one with highest sum of the p_{T}^2 of the associated tracks.

5.1.1 Pile-up vertices correction

As mentioned in Section 2.3.4, pile-up is originated from multiple proton-proton interactions which happen in the same bunch crossing. The number of pile-up events turns to be proportional to the number of reconstructed primary vertices. The pile-up is simulated adding minimum bias events on top of the hard scattering event. The amount of pile-up varies depending on the beam configurations. Therefore, during the generation of the MC simulated samples, the value is approximated and a correction is applied afterward by weighting the number of vertices in simulation (Figure 5.1b) to match the data profiles (Figure 5.1a).



(a) (b)

Figure 5.1: a) Luminosity-weighted distribution of the mean number of interactions per crossing in the $\sqrt{s} = 8\,\mathrm{TeV}$ data set. Same distribution in arbitrary units is presented in b) for a typical MC sample.

5.2 Electrons

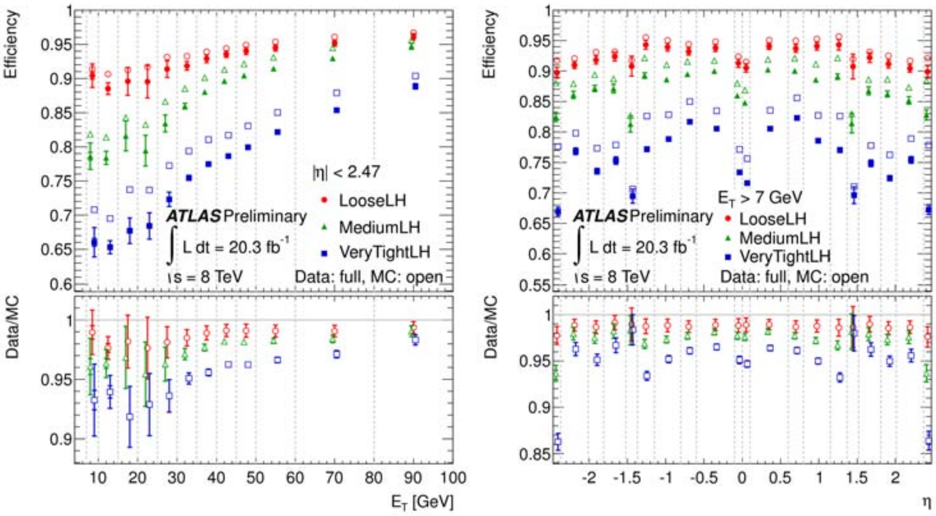
Electrons are reconstructed associating clusters of energy deposited in the electromagnetic (EM) calorimeter to a track. The energy clusters are built using a sliding-window clustering algorithm [60, 61]. First, the $\eta - \phi$ space of the calorimeter is divided into a grid of $\Delta\eta \times \Delta\phi$ elements with size 0.025×0.025. Inside each of these grid elements, the energy of all cells in all longitudinal layers is summed to form a tower of energy. A window of fixed size 3×5 is then used to scan across each element of the tower grid. From the scan, clusters with a local maxima above a 2.5 GeV threshold are selected. This limits the rate of fake clusters due to noise. Clusters matched to a well reconstructed ID track associated to a primary vertex are classified as electrons. Then the cluster energy is rebuilt using a calorimeter area of 3×7 and 5×5 in the Barrel Electromagnetic Calorimeter (EMB) and End-caps (EMEC), respectively. These areas are optimized for EM showers produced by electrons in the different detector regions, while suppressing the pile-up and noise contributions.

A high energy *pp* collision environment is characterized by the high rate of particles generated in each bunch crossing. In order to suppress the background produced by these particles and to make precise measurements, an object isolation is after required. In the case of electrons, track and calorimeter based isolations [62] are applied to suppress the jets of hadrons faking electrons. The energy associated to the tracks or the calorimeter cells around the electron should be low compared to the electron energy.

The electron identification criteria relies on a likelihood [62] selection. This likelihood-based identification combines shower-shape information, track quality criteria, the matching quality between the track and its associated energy cluster in the calorimeter, TRT information, and a criterion to help identifying electrons originated from photon conversions. Identified electrons are classified in three different quality categories: *loose*, *medium* and *tight*. The loose category fulfil a "loose likelihood" identification criteria. Medium electrons add to this criteria a cut on the transverse energy (E_{T}) to select only high energetic electrons. Finally, the "very tight likelihood" criteria, in addition to a track and calorimeter isolations, are required for the tight electrons.

To measure the identification efficiency of loose, medium and tight electrons the tag-and-probe method [62] is used in both data and MC simulation. The method requires a pure and unbiased sample of electrons, and $Z \rightarrow ee$ and $J/\psi \rightarrow ee$ event decays are used. One of the two electrons is selected with very tight requirements ("tag"), while the second electron candidate is selected applying rather loose conditions ("probe"). The efficiency ϵ is estimated by the fraction of probe electrons passing the tested criteria.

Data-to-MC scale factors (SFs) are applied to the simulated samples to correct for the differences in efficiencies observed in the simulation when compared to data. These corrections are parametrized as a function of E_{T} and η. Figure 5.2 shows the electron efficiencies in data and MC samples. The ratio plots correspond to SFs that vary between 0.9 and 1 depending on the selection quality, E_{T}, and η of the electron.



(a) (b)

Figure 5.2: Identification efficiency in data and MC as a function of (a) E_{T} and (b) η for loose, medium and tight selections. The lower panel shows the data-to-MC efficiency ratios. The uncertainties are statistical (inner error bars) and statistical+systematic (outer error bars). The dashed lines indicate the bins in which the efficiencies are calculated (taken from Ref.[62]).

5.3 Muons

Muons are identified and reconstructed [63] using the information from the muon spectrometer (MS), the ID and, to a lesser extent, the calorimeters. The reconstruction depends on the available information from the different detectors, leading to different types of muons:

- Stand-alone (SA) muon: is reconstructed using only the track MS information of muons that went through at least two layers of the chambers. To determine the parameters at the interaction point, the track is extrapolated to the closest point to the beam line taking into account the estimated energy loss in the calorimeters. In general, SA muons are used to extend the acceptance to the range $2.5 < |\eta| < 2.7$, where there is no coverage from the ID.

- Combined (CB) muon: is reconstructed by matching SA muon and ID track. Its momentum is defined as a weighted combination of the MS and the ID p_{T} measurements. The reconstruction of combined muons is limited to the ID coverage $|\eta| < 2.5$. These are the main type of reconstructed muons.

- Segment-tagged (ST) muon: is reconstructed from an extrapolated ID track to the MS, associating the inner track with at least one track segment reconstructed in a MDT or CSC chamber. ST muons increase the acceptance for low p_{T} muons, or muons which fall in regions with reduced MS acceptance.

- Calorimeter-tagged (CaloTag) muon: is reconstructed by associating an ID track with an energy deposit in the calorimeter, compatible with a minimum ionizing particle. Although this type of muon has a low purity, it recovers acceptance in the region $|\eta| < 0.1$ where there is limited coverage by the muon chambers.

Muons could also require to be isolated using track and calorimeter based isolations to further suppress the backgrounds and to avoid misidentifications.

The identification criteria classifies the muons by increasing background rejection in *loose*, *medium* and *tight*. The loose muons gather all types previously presented: SA, CB, ST and CaloTag. They are required to have an isolated track, with the exception of the SA type, where no track information is available. The medium muons are formed by CB and ST muons. They also required a track isolation in addition to a cut on the E_{T} to select only high energetic muons. Finally, tight muons are identified by tightening the medium isolation requirements.

The efficiency in the muon reconstruction and identification [63] is estimated using the same procedure as for the electrons. A tag-and-probed method is used in a pure and unbiased sample of muons both in data and MC simulated events. In this case, $Z \to \mu\mu$, $J/\psi \to \mu\mu$ and $\Upsilon \to \mu\mu$ events are used. As it was the case for electrons, data-to-MC SFs are applied to the MC samples to correct for the differences in efficiencies respect to data. Figure 5.3 shows the efficiency for the different muons types in MC and data events as a function of η. The corresponding data-to-MC correction is presented in the ratio plot, and varies between 1.02 and 0.98. The combination of all the muon reconstruction types (for CB, ST, and CaloTag muons) give a uniform muon reconstruction efficiency over almost all the detector regions.

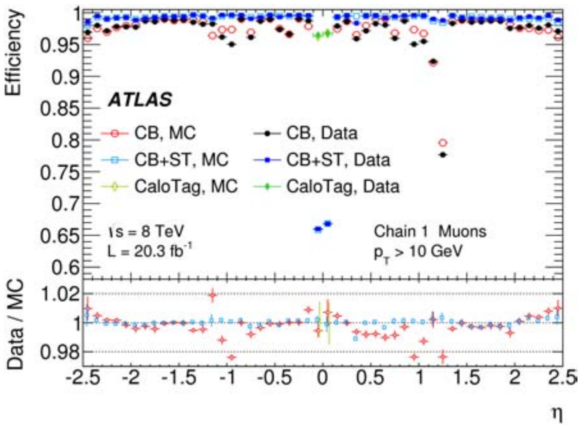


Figure 5.3: Muon reconstruction efficiency as a function of η measured in $Z \to \mu\mu$ events for muons with $p_T > 10\,\text{GeV}$ and different muon reconstruction types. CaloTag muons are only shown in the region $|\eta| < 0.1$. The error bars on the efficiencies indicate the statistical uncertainty. The panel at the bottom shows the ratio between the measured and predicted efficiencies. The error bars on the ratios are the combination of statistical and systematic uncertainties [63].

5.4 Jets

As discussed in Chapter 1, jets are used to identify collimated flows of particles from the presence of partons in the initial state. In the analysis, one can distinguish *truth* jets, *track* jets or *calorimeter* jets. Truth jets are defined in MC simulation from stable particles with a lifetime $\tau > 10^{-11}$ s. Track jets are reconstructed from charged particle tracks associated with a primary vertex. This Section is focused on the reconstruction and calibration of calorimeter jets, built from energy clusters in the calorimeter.

5.4.1 Topological clusters

Particles that interact with the calorimeter deposit their energy in different cells. To reconstruct the particle energies, cells are gathered and their energies summed. In particular, this reconstruction is performed using a 3-D topological clustering algorithm [60].

The topological cluster algorithm groups the cells that have significant energies compared to the expected noise. The noise is estimated for every cell and its dominant sources are the read-out electronics and the energy deposited by pile-up. The algorithm starts by selecting seed cells with a high signal to noise ratio, $|E|/\sigma > 4$. The next step checks the signal to noise ratio in the seed's neighboring cells, if $|E|/\sigma > 2$ they are added to the cluster seed. This procedure is repeated with their neighboring cells. The cluster stops growing when a neighboring ratio condition is not fulfiled. To finish, all the perimeter cells, regardless their signal-to-noise ratio, are added to the cluster. This ensures that tails of showers are not discarded. An illustrative diagram of the procedure is presented in Figure 5.4. Overlapping of particle showers are resolved by searching for local maxima inside a cluster and splitting it around those maxima.

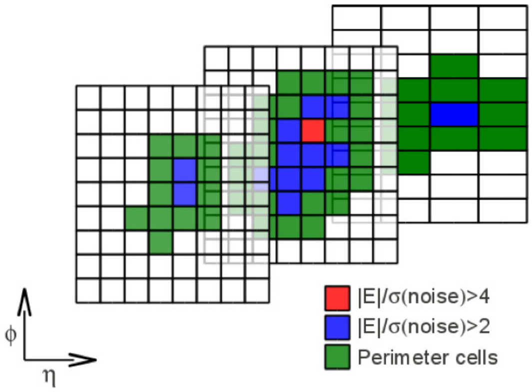


Figure 5.4: Illustrative $\eta - \phi$ grid which represents the three barrel hadronic layers. The picture shows the topo-cluster formation starting from a high signal to noise ratio cell (red), expanding in the 3-dimensional space to the medium significance cell (blue), and closing the topo-cluster by adding all perimeter cells (green).

5.4.2 Jet algorithm

In this analysis, the iterative anti-k_T algorithm [64] is used to reconstruct jets in the calorimeter, using topo-clusters as input. The algorithm starts by introducing the "distance" d_{ij} between two topo-clusters i and j and the “distance” between the topo-cluster i and the beam line d_{iB}. These variables are defined as

$$\begin{aligned} d_{ij} &= \min\left(\frac{1}{k_{\mathrm{T},i}^2}, \frac{1}{k_{\mathrm{T},j}^2}\right)\frac{\Delta_{ij}^2}{R^2}, \\ d_{iB} &= \frac{1}{k_{\mathrm{T},i}^2}, \end{aligned} \tag{5.1}$$

where k_T is the transverse momentum of the topo-clusters, Δ_{ij} is the distance in the rapidity-azimuthal plane and R is a free parameter associated with the jet size.

The clustering procedure identifies the smallest distance, if it is d_{ij} it recombines the topo-clusters i and j into a new pseudo-cluster. If instead it is d_{iB}, it calls i a proto-jet and stop using the cluster in future iterations. The procedure continues until no topo-clusters are left. To select jets associated with hard interactions, a minimum transverse momentum (p_T) is required to the proto-jets. Figure 5.5 illustrates the shape of the jets obtained by running the k_T algorithm on a set of hard and soft particles. As it is observed, algorithm provides well-defined conical shapes to the jets, allowing a robust pile-up correction.

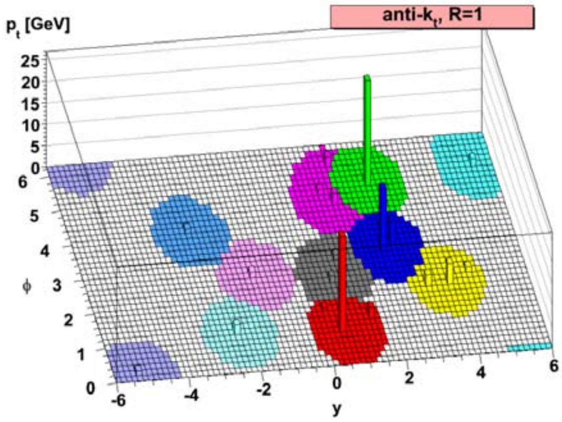


Figure 5.5: Illustrative plot showing the reconstructed jets using the k_T algorithm [64].

5.4.3 Jet energy calibration

Jets are formed from the calorimetric energy deposits reconstructed at the EM energy scale. This scale properly accounts for the EM components of the jet shower. A calibration, called local cell signal weighting (LCW), is performed to correctly reconstruct the calorimeter response to hadrons. This calibration uses the density of energy and the longitudinal shower depth to classify the topo-cluster as EM or hadronic.

An overview of the ATLAS jet reconstruction and calibration scheme is shown in Figure 5.6. The calibration of jets at the EM or LCW-scale is divided into four steps: energy correction to account for pile-up, position correction, energy and η calibration based on MC simulation and a residual *in-situ* calibration based on MC-data comparisons [65].

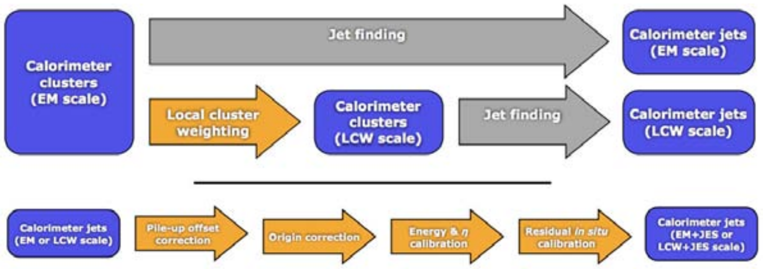


Figure 5.6: Overview of the ATLAS jet reconstruction and calibration scheme [65].

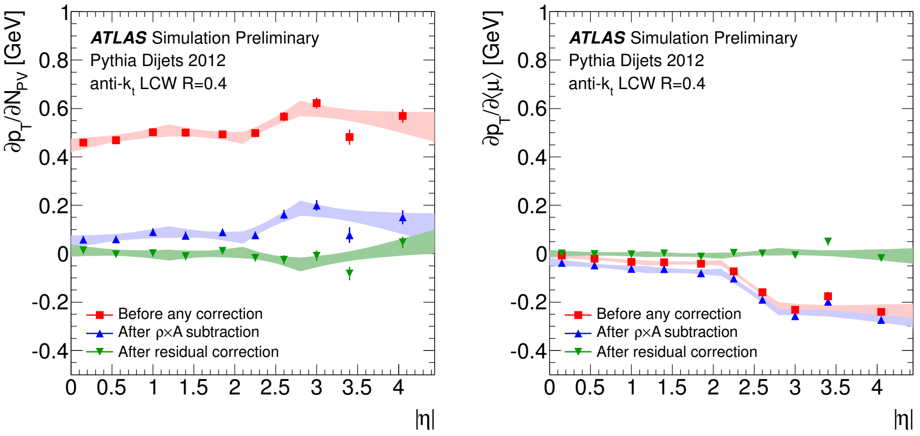
5.4.3.1 Jet pile-up correction

The pile-up activity produces an offset in the jet energy measurement. This offset is associated with two effects: the *in-time* and *out-of-time* pile-up. The in-time pile-up effect corresponds to the extra signals measure in the calorimeters due to the additional interactions produced in the same bunch crossing. The out-of-time pile-up instead, is related with further signal modulations produce by the interactions in the surrounding bunch crossings.

Two corrections are applied to mitigate the pile-up effects [66, 67]. The first one estimates the average energy deposited by the pile-up ρ, and subtracts it depending on the jet shape area A:

$$p_{\mathrm{T}}^{jet,corr} = p_{\mathrm{T}}^{jet} - \rho \cdot A. \tag{5.2}$$

This correction mitigates the in-time pile-up effects, but ρ does not properly describe the out-of-time effects, specially in the forward region of the detector. This motivates the introduction of a residual correction that depends on the jet p_{T}, the number of vertices in the event N_{PV} and (since it accounts for out of time pile-up effects) the average number of interactions per bunch crossing $\langle\mu\rangle$. Figure 5.7 shows the pile-up effects on the jet reconstructed p_{T} before and after applying the corrections.



(a) (b)

Figure 5.7: Dependence of the reconstructed jet p_{T} on in-time pile-up (a) and out-of-time pile-up (b), at various correction stages: before any correction, after the $\rho \cdot A$ subtraction, and after the residual correction [67].

5.4.3.2 Jet origin correction

Initially, the jet direction is estimated with the energy-weighted position of the topo-cluster cells pointing to the center of the detector. A correction makes the jet to point to the primary vertex without changing its energy. This correction improves the angular resolution and results in a small improvement ($< 1\,\%$) in the jet p_{T} response.

5.4.3.3 Jet energy scale

The jet energy scale (JES) calibration brings the jet energies measured at detector level back to the particle-level (truth jets) [68]. The aim is to correct for detector effects such as: partially measured energies specially in hadronic showers; energy leakages from showers not contained in the calorimeters or in the jet cones, or energy deposition in non-sensitive regions of the detector.

The calibration is based on MC simulation. It is extracted from matched pairs of isolated jets at calorimeter and truth levels, where the jets are close to each other within a cone of radius 0.3. Isolated jets are defined such that there is no additional jet with $p_{\mathrm{T}} > 7\,\mathrm{GeV}$ within $\Delta R = 2.5R$, where R is the jet algorithm size parameter.

Using these pairs of matched calorimeter-truth jets, the calibration starts by computing the energy response function

$$\mathcal{R}_{EM/LCW} = \frac{E^{jet}_{EM/LCW}}{E^{jet}_{truth}} \tag{5.3}$$

for each pair in the MC sample. Results are binned in the truth jet energy E^{jet}_{truth} and the calorimeter jet detector pseudorapidity η_{det}, defined as the jet η before the origin correction. Then, a Gaussian fit is performed to estimate the averaged jet response ($\langle\mathcal{R}_{EM/LCW}\rangle$) for each ($E^{jet}_{truth}$,$\eta_{det}$) bin. In the same bins, the average jet calorimeter energy ($\langle E^{jet}_{EM/LCW}\rangle$) is also derived as the mean of the $E^{jet}_{EM/LCW}$ distribution. Figure 5.8 shows the average response of simulated jets formed at the EM scale as a function of η_{det} for various truth-jet energies.

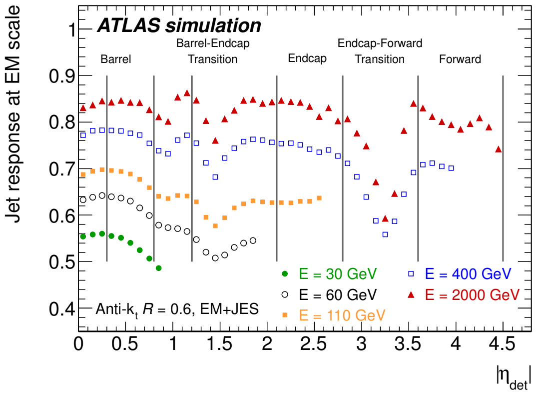


Figure 5.8: Average response of simulated jets formed from topo-clusters at the EM scale ($\langle\mathcal{R}_{EM}\rangle$) as a function of the uncorrected jet pseudorapidity η_{det}. The response is shown separately for various truth-jet energies. Also indicated are the different calorimeter regions [68].

The jet calibration function ($\mathcal{F}_{calib}$) depends on the $E^{jet}_{EM/LCW}$ and it is obtained from the fit of the $\langle\mathcal{R}_{EM/LCW}\rangle$ and $\langle E^{jet}_{EM/LCW}\rangle$ values for each E^{jet}_{truth} bin. Figure 5.9 presents the average jet energy scale correction ($1/\mathcal{F}_{calib}$) as function of the jet p_{T}. The values of the function decrease with p_{T}, ranging in the central region of the detector from 2.1 at low p_{T} to 1.2 at high.

The final jet energy scale correction that relates the measured calorimeter energy back to the particle-level is defined as

$$E^{jet}_{EM/LCW+JES} = \frac{E^{jet}_{EM/LCW}}{\mathcal{F}_{calib}(E^{jet}_{EM/LCW}, \eta_{det})} \tag{5.4}$$

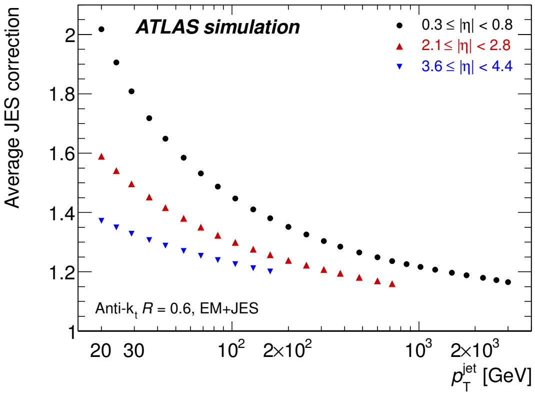


Figure 5.9: Average jet energy scale correction $(1/\mathcal{F}_{calib})$ as a function of the calibrated jet transverse momentum for three representative η_{jet}-intervals obtained from the nominal Monte Carlo simulation sample [68].

5.4.3.4 In-situ calibration

After the EM/LCW+JES calibration is applied to the p_{T} of the jets, residual differences between data and MC simulation in the jet p_{T} need to be addressed. This is important since the MC simulation is used to bring the measurements at the detector level back to the truth particle level. In case of difference, a residual correction is applied to the data such that comes closer to the simulation.

Samples of γ+jet, Z+jet events are used to compare the performance of the jet calibration in data and simulation where the photon or the Z boson act as reference. For very high-p_{T} jets, this is complemented with a multijet sample in which a high-p_{T} jet recoils a system of low-p_{T} multijets. In each case, the calibration is applied as a jet-by-jet correction defined as

$$\mathcal{C}_{in-situ}({p_{\mathrm{T}}}^{jet},\eta)=\frac{\langle p_{\mathrm{T}}^{jet}/p_{\mathrm{T}}^{ref}\rangle_{MC}}{\langle p_{\mathrm{T}}^{jet}/p_{\mathrm{T}}^{ref}\rangle_{data}}. \tag{5.5}$$

For the central part of the detector $|\eta| < 1.2$, three analyses are performed to estimate the in-situ correction. The analysis where Z boson is used as the reference objects, allows to calibrate jets with $15 < p_{\mathrm{T}} < 200$ GeV. Instead, when γ's are used, the calibration covers the range $20 < p_{\mathrm{T}} < 800$ GeV. Finally, for jets with $p_{\mathrm{T}} > 800$ GeV, the calibration is performed using a system of a low-p_{T} jets recoiling against a high-p_{T} jet, where the low-p_{T} jet is calibrated using the γ-jet or Z-jet in-situ techniques. Figure 5.10 shows the in-situ correction estimated using the three analyses. The correction ranges between -4 % and -2 % as increases the p_{T}.

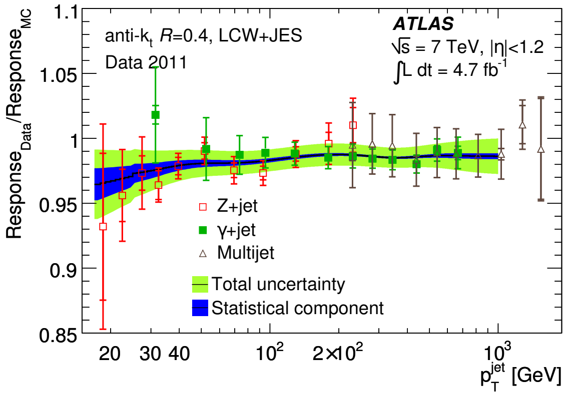


Figure 5.10: Ratio of the average jet response $\langle p_{\mathrm{T}}^{jet}/p_{\mathrm{T}}^{ref}\rangle$ measured in data to that measured in MC simulations for jets within $|\eta| < 1.2$ as a function of the transverse jet momentum p_{T}. The data-to-MC jet response ratios are shown separately for the three in situ techniques used in the combined calibration: direct balance in Z-jet events, photon-jet events, and multijet p_{T} balance in inclusive jet events [65].

5.4.4 Jet vertex fraction

The jet vertex fraction (JVF) is a variable used to reduce the contribution of jets from pile-up. It is calculated as the sum of the p_{T} of tracks matched to the jet that originate from the primary vertex divided by the sum of p_{T} of all tracks matched to the jet. The JVF is only required for jets with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ and $|\eta| < 2.4$, since low p_{T} jet measurements are more affected by pile-up. Due to detector acceptance, only jet associated tracks within $|\eta| < 2.4$ are considered. Figure 5.11 illustrates the JVF distribution for jets produced in a hard-scatter or a pile-up interaction in the simulation.

5.5 Objects overlap removal

By construction, leptons and jets are reconstructed independently. As consequence, close objects could lead to a double counting of the energy deposited in the calorimeters. Another consequence, is that the same object can be reconstructed by different algorithms, e.g. electrons and jets are both reconstructed as jets by the jet algorithms. Therefore, a hierarchy, presented in Table 5.1, is established to resolve overlaps.

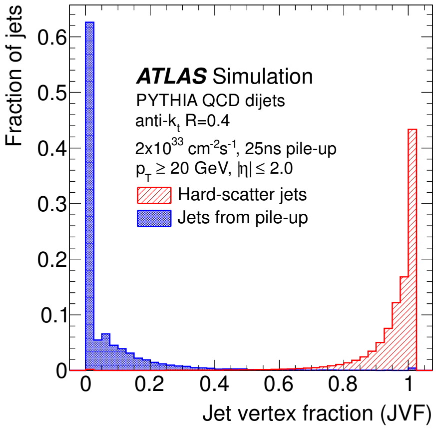


Figure 5.11: The JVF distribution for simulated jets associated to a hard-scattering or a pile-up interaction.

Case	Order	Keep μ	Keep e	Keep jet
$\Delta R(jet, e) < 0.4$	1	–	Yes	No
$\Delta R(jet, \mu) < 0.4$	2	$N_{trk} \leq 3$	–	$N_{trk} \geq 4$
$\Delta R(\mu, e) < 0.2$	3	if not Calo μ	if Calo μ	–

Table 5.1: Overlap hierarchy and the order in which the overlap is solved where N_{trk} corresponds the jet track multiplicity.

5.6 b-jets

The bottom quark hadronizes in a B hadron when produced in a particle collision. These hadrons have a lifetime of the order of 10^{-12} s. At the energies they are normally produced at LHC, they can travel few millimeter before decaying. This relatively long lifetime allows the reconstruction of a displaced secondary vertex which respect to the primary collision point. Algorithms exploit this property to identify the jets generated by B hadrons, referred as to b-jets.

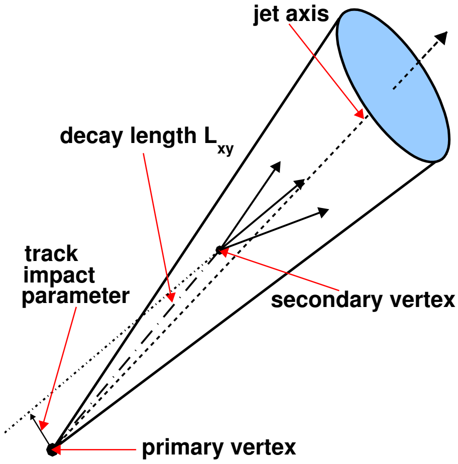


Figure 5.12: Schematic view of a B hadron decay resulting in a secondary vertex. The vertex is significantly displaced with respect to the primary vertex. The track impact parameter, which is the distance of closest approach between the extrapolation of the track and the primary vertex, is shown in addition for one of the secondary tracks [69].

5.6.1 b-tagging algorithm

The b-jets are identified using an algorithm called b-tagging. In general, it defines a weight corresponding to the probability for a jet to be generated from a B hadron. The algorithm used in this analysis is called MV1c [70]. Itself combines, using a neural network, the information from various algorithms based on the impact parameter of the jet tracks, the presence of a secondary vertex and the topology of *bottom* and *charmed* hadron decays. Figure 5.12 shows a schematic view of a B hadron decay and highlights some of the variables used by the b-tagging algorithms.

The output of the MV1c algorithm is a weight (w) that ranges between

0 and 1. The closer the associated w value of a jet is to 1, the more likely to be a b-jet. The algorithm performance is characterized by the efficiency of tagging a b-jet (ϵ_b) versus the probability of mis-identifying (mis-tag) a jet as b-jet when it is generated by a c-quark, a τ-jet or a light-flavor parton (u, d, s quarks or a gluon).

Analyses normally set a w threshold value, or working point, that defines if a jet is considered to be a b-jet. Typically, a working point close to 1 gives high purity b-jet samples but with a low b-tagging efficiency, while a working point close to 0 gives low purity b-jet samples with a high b-tagging efficiency. Therefore, when choosing the working point a balance between the purity and the efficiency is needed.

5.6.2 b-tagging calibration

The MV1c algorithm needs to be calibrated such that efficiencies in simulation mimic the data. In MC generated events, the efficiencies are estimated independently for the different jet flavors. The flavor is determined by searching for different particles within a cone of 0.4 around the jet axis just after the parton in the final state has hadronized. The jets are labeled as b-jet if a B hadron is found, c-jet if a *charmed* hadron and not a B hadron is found, or τ-jet if a τ lepton is found. If none of the presented requirements are fulfiled the jet is labeled as light-jet.

The following Sections explain the different methods used to estimate the tagging efficiencies of b-jets [71], c-jets and light-jets [70], both in data and simulation. Data-to-MC SF are estimated as the ratio between data and simulation efficiencies and are applied on MC to account for the differences with data. These SFs are binned in jet p_{T} and b-tagging weight. For light-jets, efficiencies are also binned in $|\eta|$ since the statistics allows to do so. To some extent, the SFs include effects and physics modeling and therefore, they are only strictly valid for the generator used to derive them. The differences observed when the efficiencies are measured with different MC generators are taken into account by additional MC-to-MC SFs.

5.6.2.1 b-jet efficiency

The b-jet tagging efficiency ϵ_b is estimated in MC and data $t\bar{t}$ events. To obtain the highest possible b-jet purity samples, an analysis is performed on $t\bar{t}$ candidates selected with two opposite charged leptons in the final state (dilepton top channel).

There are several methods to estimate the b-jet efficiency. In the case of this analysis, where events with two or three jets in the final state are selected, a likelihood fit-based approach is used to easily take into account the correlations between the jet flavors. This method constructs likelihood functions using the probabilities of b-tagging certain flavor jets given their *b*-tagging weights and p_{T}. Likelihood is fitted to data and the probability density function of tagging a *b*-jet is obtained. Efficiencies are estimated integrating the probability density function over the *b*-tagging weight variable.

5.6.2.2 c-jet efficiency

The efficiency with which the *b*-tagging algorithm tags *c*-jets is referred to as the *c*-jet tagging efficiency ϵ_c. It is calibrated using a jet sample enriched in *c*-jets containing a D^{*+} meson identified in the $D^0(\rightarrow K^-\pi^+)\pi^+$ decay mode. The efficiency is measured by comparing the yield of mesons before and after the tagging requirement.

5.6.2.3 light-jet efficiency

The light-jet efficiency ϵ_l is defined as the fraction of jets originating from light flavor which are tagged by a *b*-tagging algorithm. Light-flavor jets are mis-tagged as *b*-jets mainly because of the finite resolution of the inner detector and the presence of tracks from long-lived particles or from the interactions with the dead material in the detector. Prompt tracks which are apparently displaced due to the finite resolution of the tracker, will often appear to be originate from a point behind the primary vertex with respect to the jet axis. The negative tag method exploits this feature of the light-jets, by negative tagging in impact parameter or in decay length. The method translates this negative tag rate into a measurement of the mis-tag rate.

5.6.3 pseudo-continuous b-tagging

In this analysis, several working points are used in a procedure known as *pseudo-continuous tagging*. This procedure allows not only to define if a jet is considered a *b*-jet but also further classified them according to its purity. The categories defined are *loose* (with an efficiency of 80 %), *medium* (70 %) and *tight* (50 %). Table 5.2 summarizes the information for the different working points used in the analysis. Figure 5.13 shows an illustrative example of MV1c weight distributions for the different jet flavors and the different analysis working points. The advantages of using a *pseudo-continuous tagging* is that low purity with high statistics, and high purity with low statistics *b*-jet samples can be defined in the same analysis.

5.6.4 b-jet energy and dijet mass resolution

As it will be described in Chapter 6, the most powerful variable to search for the process VH ($H \rightarrow b\bar{b}$) is the reconstructed mass of the two *b*-jets. Therefore and in order to maximize the discriminant power and the discovery

Name	w value	ϵ_b (%)	c-jet RF	τ-jet RF	l-jet RF
80 'Loose' or L	0.4050	79.85	3.04	6.40	29.12
70 'Medium' or M	0.7028	70.00	5.34	14.90	135.76
50 'Tight' or T	0.9237	49.99	26.22	120.33	1388.28

Table 5.2: The MV1c output w value and the efficiencies associated with the working points estimated from simulated events. RF stands for Rejection Factor which is the reciprocal of the efficiency.

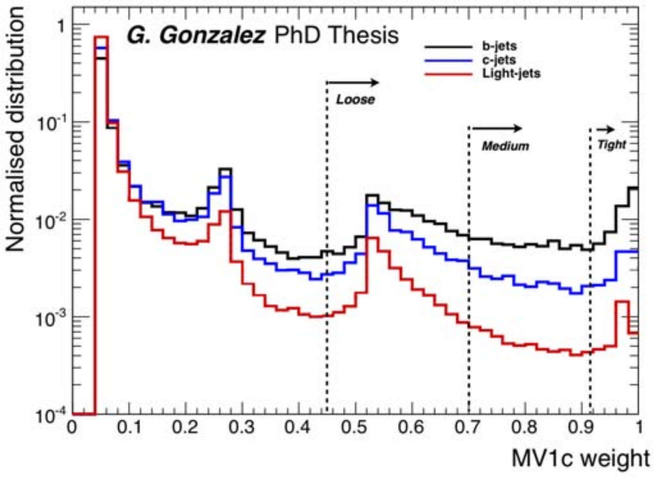


Figure 5.13: Illustrative example of the MV1c weight distributions for different jet flavors in simulation. Dash lines correspond to the analysis working points loose with an efficiency of 80 %, medium with 70 % and tight 50 %.

potential, a jet-by-jet calibration and two corrections are applied to the dijet invariant mass to improve the jet energy resolution.

The extra jet calibration, known as global sequential calibration (GSC) [68], is applied on top of the LCW+JES one. This calibration parametrized the jet response (Eq. 5.3) as a function of variables (x_i) related with the longitudinal and transverse jet energy deposition. For every variable, a multiplicative correction to the jet energy measurement is derived by inverting the calibrated jet response as a function of the variable

$$\mathcal{C}(x) = \frac{\mathcal{R}^{-1}(x)}{\langle\mathcal{R}^{-1}(x)\rangle}, \tag{5.6}$$

where $\langle\mathcal{R}^{-1}(x)\rangle$ corresponds to the average inverse jet response. This procedure requires that the correction for a given variable x_i (C_i) is calculated using jets to which the correction for the previous variable x_{i-1} (C_{i-1}) has already been applied. The jet p_{T} after the correction i is defined as

$$p_{\mathrm{T}}^{i} = C_i(x_i) \times p_{\mathrm{T}}^{i-1} = C_i(x_i) \times C_{i-1}(x_{i-1}) \times p_{\mathrm{T}}^{i-2} = ... \tag{5.7}$$

By applying these corrections to the jet p_{T}, the dependencies between the jet response and the variables are removed, translating in a reduction of the spread of the reconstructed jet p_{T} without changing its average energy.

The first of the two extra corrections applied to the dijet mass is called muon-in-jet correction. This correction is motivated by the presence of a muon and a neutrino when a b-quark decays semileptonically. The presence of neutrinos degrades the energy resolution of the jets since they do not deposit any energy in the calorimeters. Since on average the muon and neutrino kinematics are not very different, the p_{T} of the muon is added to the jet p_{T} when a reconstructed muon is found inside the jet. This compensates the lost of the neutrino and improves the m_{bb} resolution by 12 % which respect to the one given by LCW+JES+GSC jets.

A last correction, called reconstruction correction, is applied to compensate for the bias introduced by the underlying jet p_{T} spectrum in the average jet response function. It is derived from simulated $ZH \to llb\bar{b}$ events by comparing the truth and the reconstructed jet p_{T}. This correction is parametrized as a function of the reconstructed p_{T} and it is defined as

$$\mathcal{C}_{p_{\mathrm{T}}-reco}(p_{\mathrm{T}}^{reco}) = \langle\frac{p_{\mathrm{T}}^{truth}}{p_{\mathrm{T}}^{reco}}\rangle. \tag{5.8}$$

Applying this corrections to GSC calibrated jets, make them vary their p_{T} around 5 % for $p_{\mathrm{T}} > 60\,\mathrm{GeV}$ and increasing up to 15 % for lower energies. This correction adds an extra 2 % improvement to the m_{bb} resolution.

Figure 5.14 illustrates the impact of the different corrections in the dijet mass resolution for one of the analysis distributions. The dijet mass resolution for the signal is improved by 14 % after all these corrections.

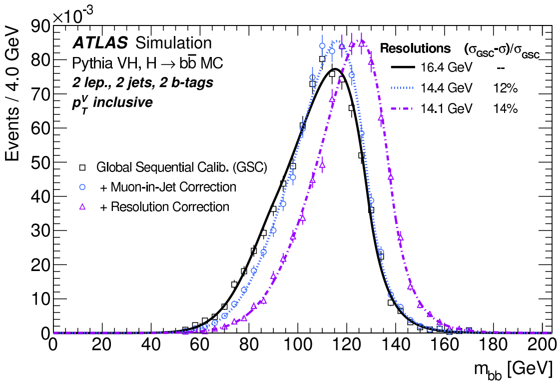


Figure 5.14: Dijet mass distribution for the decay products of a Higgs boson with $m_H = 125\,\text{GeV}$ in the 2-lepton analysis selection (see Chapter 6). In solid line the distribution using b-jets with the global sequential calibration (GSC), dash line if jets have also the muon-in-jet correction and dash-dotted line if the resolution correction is also applied.

5.7 Missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$

The missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$ is defined as the momentum imbalance in the plane transverse to the beam axis. The reconstruction of the $E_{\mathrm{T}}^{\mathrm{miss}}$ [72, 73] uses the energy deposits in the calorimeters and the momentum measured in the muon spectrometer. The calorimeter part is estimated associating the energy cell to a reconstructed and calibrated object in the following order: electrons, photons, hadronically decaying τ-leptons and jets. The different calorimeter terms used in the estimation of the $E_{\mathrm{T}}^{\mathrm{miss}}$ are:

- $E_{\mathrm{x(y)}}^{\mathrm{miss,e}}$: is reconstructed from cells in clusters associated to *medium* electrons with $p_{\mathrm{T}} > 10\,\text{GeV}$.

- $E_{\mathrm{x(y)}}^{\mathrm{miss},\gamma}$: corresponds to cluster cells associated to photons passing a *tight* identification criteria and with $p_{\mathrm{T}} > 10\,\text{GeV}$ [61].

- $E_{\mathrm{x(y)}}^{\mathrm{miss},\tau}$: is reconstructed from cell cluster calibrated at the LCW-scale associated with hadronically decaying τ-leptons passing a *tight* identification criteria and with $p_{\mathrm{T}} > 10\,\text{GeV}$ [74].

- $E_{\mathrm{x(y)}}^{\mathrm{miss,jets}}$: is reconstructed from LCW calibrated cells using the anti-k_T algorithm with an $R = 0.6$. Jets are required to have $p_{\mathrm{T}} > 20\,\text{GeV}$.

- $E_{\mathrm{x(y)}}^{\mathrm{miss,SoftTerm}}$: contains the LCW calibrated cluster energies not associated with any object, the contributions from jets with $7 < p_{\mathrm{T}} < 20\,\text{GeV}$ and reconstructed tracks with $p_{\mathrm{T}} > 400\,\text{MeV}$. Track momenta is added to account for low-p_{T} particles that do not reach the calorimeter or do not have enough energy to seed a topocluster.

- $E_{\mathrm{x(y)}}^{\mathrm{miss},\mu}$: corresponds to the p_{T} lost by muons in the calorimeter. As will be explained below, this contribution is sometimes not taken into account in favor of precise muon momentum measurements performed with the muon spectrometer.

In the case of the muons, two different procedures are applied to properly deal with its energy deposited in the calorimeter. If the muon is isolated, its momentum is taken from the measurement of the muon spectrometer and the inner detector (see Section 5.3), not considering the energy it deposits in the calorimeter. On the other hand, when the muon is non-isolated the energy it deposits in the calorimeter is considered, and the p_{T} measured by the MS is added. Finally, in the region $2.5 < |\eta| < 2.7$, where there is no coverage of the ID, the p_{T} measured by the MS is used for both isolated and non-isolated muons.

The $E_{\mathrm{T}}^{\mathrm{miss}}$ is defined as the negative vector sum of the physics objects in the event, thus its x and y components can be expressed as

$$E_{\mathrm{x(y)}}^{\mathrm{miss}} = E_{\mathrm{x(y)}}^{\mathrm{miss,e}} + E_{\mathrm{x(y)}}^{\mathrm{miss},\gamma} + E_{\mathrm{x(y)}}^{\mathrm{miss},\tau} + E_{\mathrm{x(y)}}^{\mathrm{miss,jets}} + E_{\mathrm{x(y)}}^{\mathrm{miss,SoftTerm}} + E_{\mathrm{x(y)}}^{\mathrm{miss},\mu}. \tag{5.9}$$

Using its x and y components, the $E_{\mathrm{T}}^{\mathrm{miss}}$ and its azimuthal coordinate (ϕ^{miss}) can be written as

$$\begin{aligned} E_{\mathrm{T}}^{\mathrm{miss}} &= \sqrt{(E_{\mathrm{x}}^{\mathrm{miss}})^2 + (E_{\mathrm{y}}^{\mathrm{miss}})^2} \\ \phi^{\mathrm{miss}} &= \arctan{(E_{\mathrm{x}}^{\mathrm{miss}}, E_{\mathrm{y}}^{\mathrm{miss}})}. \end{aligned} \tag{5.10}$$

Chapter 6

Search for the Standard Model Higgs boson in the VH production channel $ZH \to \nu\nu bb$

The highest Higgs boson production rate is given by the gluon-gluon fusion initiated process $ggF \to H \to b\bar{b}$, but the huge $b\bar{b}$ background produced at LHC makes virtually impossible the search and study of the Higgs boson via this process. Therefore, the most promising way to study this decay mode is to search for the Higgs production in association with a vector boson $VH(H \to b\bar{b})$.

The search is divided according to the vector boson decay modes in three analysis channels. These channels target the processes $ZH \to \nu\nu b\bar{b}$, $WH \to \nu l b\bar{b}$ and $ZH \to llb\bar{b}$ and present different background compositions. Due to the number of leptons (electrons or muons) in their final states they are referred to as 0-lepton, 1-lepton and 2-lepton channel analyses, respectively. A simultaneous fit to all the channels permits the constrain of the different background processes taking into account properly the correlations.

The VH search is performed using two independent analysis procedures. The first one selects events by applying cuts on individual kinematic variables and uses the invariant mass of two b-jets as the discriminant variable to search for the Higgs. In this document, this procedure is referred to as "dijet mass analysis". The second relies on multivariate techniques. This procedure uses an algorithm to determine a probability of an event to be originating from a signal or background process.

This Chapter is devoted to a detailed description of the 0-lepton dijet mass analysis. As explained, the VH search is performed combining it with the information from the 1- and 2-leptons channels. The general characteristics of the 1- and 2-leptons analyses are found in Appendix A. Along with the dijet mass analysis results, this thesis also presents the multivariate results, hence a brief description of this technique is given in Appendix B.

6.1 Data samples and trigger selection

The analysis uses all the good data for physics analyses recorded by the ATLAS experiment during Run I. This corresponds to 4.6 fb^{-1} and 20.3 fb^{-1} at the center-of-mass of $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$, respectively [75]. Figure 4.10 shows the integrated luminosity as a function of time during the years 2011 and 2012.

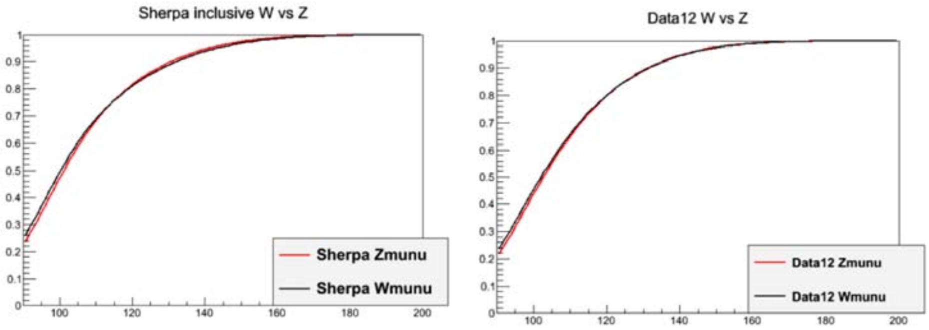
This data is selected in the 0-lepton analysis using three different $E_{\text{T}}^{\text{miss}}$ triggers (see Table 6.1). All of them have a threshold of $E_{\text{T}}^{\text{miss}} > 80\,\text{GeV}$ in the final level of the trigger chain while having different requirements in the intermediate chain levels L1 and L2 (a detailed description of the Trigger chain is found in Section 4.4). The reason to use three triggers is to maximize the luminosity used in the analysis. There was a brief data taken period where the first bunch crossing of particles could not be used by the $E_{\text{T}}^{\text{miss}}$ trigger with the loosest requirements (*trigger 1* in Table 6.1). A second trigger (*trigger 2*) was used in this period to minimize the lost of high $E_{\text{T}}^{\text{miss}}$ data events. This problem was overcame by incorporating a noise suppression tool in the forward region of the calorimeter (*trigger 3*).

	Selection	L1	L2	EF	Luminosity (fb^{-1})
trigger 1	$E_{\text{T}}^{\text{miss}} < 160$	40	45	80	1.92
trigger 2	$E_{\text{T}}^{\text{miss}} > 160$	50	55	80	2.13
trigger 3(*)	-	40	45	80	18.13

Table 6.1: The different triggers used in the analysis, the $E_{\text{T}}^{\text{miss}}$ requirements, the threshold values and associated luminosities. All threshold and selection values are presented in GeV. (*) Trigger with the noise suppression in the forward region of the calorimeter.

6.1.1 Trigger efficiency

To study the efficiency of the $E_{\text{T}}^{\text{miss}}$ triggers, unbiased samples of $Z \to \mu\mu$ and $W \to \mu\nu$ events are used. Both simulated and data events samples are used to model the so-called turn-on efficiency curve as a function of the offline $E_{\text{T}}^{\text{miss}}$. Figure 6.1 shows the *trigger 3* turn-on curves using SHERPA MC and data samples. This trigger is found to be around 80 % efficient at $E_{\text{T}}^{\text{miss}} = 120\,\text{GeV}$, becoming fully efficient at $E_{\text{T}}^{\text{miss}} = 150\,\text{GeV}$. The analysis uses the trigger down to $E_{\text{T}}^{\text{miss}} = 100\,\text{GeV}$, and relies on the determination of the turn-on curves.



(a) (b)

Figure 6.1: Trigger efficiency turn-on curves estimated using $W \to \mu\nu$ and $Z \to \mu\mu$ (a) SHERPA MC and (b) data samples.

The trigger efficiency curve is slightly different in data and MC simulation. Therefore, a scale factor (SF) correction is applied to the MC samples. This SF is defined as the ratio between both efficiencies, Eff^{Data}/Eff^{MC}. Figure 6.2 shows the correction results as a function of the $E_{\mathrm{T}}^{\mathrm{miss}}$ for the *trigger 3*. In the region $E_{\mathrm{T}}^{\mathrm{miss}} > 100\,\mathrm{GeV}$, where the trigger is used, the SF ranges between ~0.94 and 1.

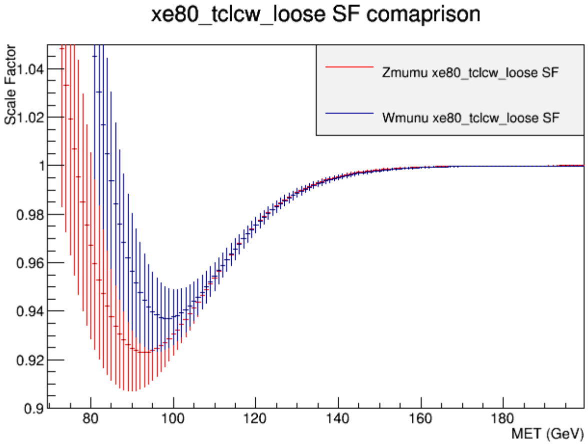


Figure 6.2: Calculated scale factor for *trigger 3* (technically named EF_xe80_tclcw_loose) from $W \to \mu\nu$ and $Z \to \mu\mu$ samples.

Correlations between the trigger efficiency as a function of the $E_{\mathrm{T}}^{\mathrm{miss}}$ and other relevant variables for the analysis are studied. An example of these variables are the p_{T} of the jets, the $\Delta\phi$ between jets, or the dijet invariant mass m_{jj}. This study shows a dependency with the sum of the p_{T} of the jets. As it is explained in Section 6.5, this dependency is solved by removing the bias kinematic region.

6.2 Object selection

In this Section, the details on the selection of the physic objects, as described in Chapter 5, are given.

6.2.1 Leptons

As mention in Chapter 5, leptons (electrons and muons) are classified in the VH analysis in three different categories according to their purity: loose, medium and tight. The 0-lepton analysis applies a lepton veto by requiring no loose leptons in the events. Loose electrons are considered with $E_{\mathrm{T}} > 7\,\mathrm{GeV}$ and $|\eta| < 2.5$. In the case of the muons, there are three types of loose muons defined according to the reconstruction procedure that is used. This allows a uniform muon reconstruction efficiency across rapidity. Type (1) muons are reconstructed in both the inner detector and the muon spectrometer in the region $|\eta| < 2.7$ and with $E_{\mathrm{T}} > 7\,\mathrm{GeV}$. Type (2) corresponds to muons with $E_{\mathrm{T}} > 20\,\mathrm{GeV}$ detected in the calorimeter with an associated track in the region $|\eta| < 0.1$, to compensate the muon spectrometer coverage there. Finally, type (3) are the muons detected by the muon spectrometer, outside the inner detector coverage, in the range $2.5 < |\eta| < 2.7$ and with $E_{\mathrm{T}} > 7\,\mathrm{GeV}$. Type (1) and (2) are required to have a track impact parameter which respect to the primary vertex smaller than 0.1 mm in the transverse plane and 10 mm along the z-axis.

The leptons generated from the vector boson decays are expected to be isolated. Therefore, a track based isolation is required to the leptons. This requirement uses the variable iso_{track}, defined as the scalar sum of the transverse momenta of the tracks within a cone of radius 0.2 centered on the lepton track over the lepton p_{T}. All leptons with an associated track are required to have $iso_{track} < 0.1$. Table 6.2 summarizes all lepton requirements.

6.2.2 Jets

Jets are reconstructed from 3D topological clusters using the anti-k_T algorithm with a radius parameter $R = 0.4$. Detailed explanation of all jet reconstruction and calibration procedure is found in Section 5.4.

The analysis uses two orthogonal jet categories, the *central* and the *forward* jets. The criteria to select them is presented in Table 6.3. The *central* jets are reconstructed in the region $|\eta| < 2.5$ and are required to have a $p_{\mathrm{T}} > 20\,\mathrm{GeV}$. The *forward* jets are selected with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ in the region

Flavor Type	Electrons	Muons (1)	 (2)	 (3)
$\|\eta\|$	< 2.5	< 2.7	< 0.1	in [2.5, 2.7]
E_{T} (GeV)	> 7	> 7	> 20	> 7
$\|d_0\|$ (mm)	–	< 0.1	< 0.1	–
$\|z_0\|$ (mm)	–	< 10	< 10	–
iso_{track}	< 0.1	< 0.1	< 0.1	–

Table 6.2: Loose lepton requirements.

$2.5 < |\eta| < 4.5$. The *central* jets are used to reconstruct potential Higgs candidates and its multiplicity helps in defining the analysis categories. Events with *forward* jets are discarded to reduce the $t\bar{t}$ background. In order to suppress fake jets coming from pile-up, JVF> 0.5 (Section 5.4.4) is requested for the *central* jets.

Type	Central Jet	Forward Jet
p_{T} (GeV)	>20	>30
$\|\eta\|$	< 2.5	in [2.5, 4.5]
|JVF|	> 0.5	–

Table 6.3: Selection used to define the sets of signal and forward jets.

6.2.3 b-jets

The *b*-jets and the algorithm to define them are key parts of the analysis given the Higgs signature. The identification of a jet being generated by a *b*-quark is carried out using the *b*-tagging algorithm called MV1c. The *pseudo-continuous tagging* used in the analysis allows to classify the *b*-jets according to their purity in *loose*, *medium* and *tight*. A detailed description of the *b*-jets identification and calibration can be found in Section 5.6.

6.2.3.1 Parametrized b-tagging

The generation of a large sample of MC events with heavy-flavor jets is demanding. In addition, the MV1c *b*-jet mis-tag is small and when this algorithm is applied to the light-jet enriched samples, few events fulfil the *b*-jet requirements. Therefore, instead of directly tagging the *c*- and *light*-labeled jets with the MV1c algorithm, a parametrization of their probability of being mis-tagged as *b*-jets is derived as a function of p_{T} and η. This probability is then applied as a weight to the events. This procedure applied to some light-jet enriched samples to enlarge their statistics after requiring events with b-jets is called *parametrized b-tagging.*

6.2.4 $E_{\mathrm{T}}^{\mathrm{miss}}$ and track-based missing transverse momenta

The missing transverse momentum is measured from the negative vector sum of the following objects: electrons, photons, τ's, jets, muons, and energy in calorimeter clusters which have not been associated with a reconstructed object. The detailed description of its reconstruction can be found in Section 5.7. The $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution is a very important quantity in the 0-lepton channel since is the one used to estimate the Z boson transverse momentum p_T^Z.

A track-based missing transverse momentum ($p_{\mathrm{T}}^{\mathrm{miss}}$) is also calculated based on the vector sum of the transverse momenta of the tracks associated to the primary vertex. The tracks used fulfil the following requirements: $p_{\mathrm{T}} > 0.5\,\mathrm{GeV}$, $|\eta| < 2.4$, pixel hits > 0, SCT hits > 5, $d_0 < 1.5\,\mathrm{mm}$ and $z_0 \sin\theta < 1.5\,\mathrm{mm}$. As will be presented in Section 6.5, a cut on this variable is used to suppress the multijet background production.

6.3 Event Categorization

The composition of backgrounds varies in the different VH channels, being the most important W+jets, Z+jets and $t\bar{t}$. The goal of the categorization is to classify the events in regions to study and constrain these backgrounds ("control regions") and isolate the high sensitive $S/\sqrt{B}$ regions ("signal regions"). As it is presented in this Section, the construction of the control and signal regions translates into a complex event categorization.

The lepton multiplicity defines the three orthogonal analyses that compose the VH search: no leptons that targets the signal1 $ZH \to \nu\nu bb$, one lepton for $WH \to l\nu bb$ and two leptons for $ZH \to llbb$. In addition, central jets are used to select the events compatible with a Higgs decay. According to their jet multiplicities the events are categorized in 2- or 3- jets regions which helps to isolate regions with a large signal to background ratio. Table 6.4 summarizes these requirements.

Analysis	Leptons			Jets	
	VH Loose	VH Medium	VH Tight	Central	Forward
0-lepton	0	–	–	2-3	0
1-lepton	0	0	1		
2-lepton	1	1	–		

Table 6.4: Lepton and jet selection for the three VH analyses.

To construct the control regions, events are also categorized according to their b-jet multiplicities in 0-, 1- and 2-tag exclusive regions. Table 6.5 shows for the three lepton channels the main background composition of these b-jet categories. It can be observed that the main backgrounds of 1- and 2-lepton control regions are W+jets and Z+jets, respectively. This makes these categories perfect to study and constrain them. Contrarily, the mixture of backgrounds in the 0-lepton channel makes more difficult to obtain such good constrains in its control regions. Therefore, one of the main motivation of using such a complex categorization is that the W+jets and Z+jets constrains obtained from 1- and 2-leptons are then applied to the 0-lepton channel.

	0-lepton	1-lepton	2-lepton
0-tag	Z+jets W+jets	W+jets	Z+jets
1-tag	Z+jets W+jets $t\bar{t}$	W+jets	Z+jets
2-tag	Z+jets W+jets $t\bar{t}$	W+jets $t\bar{t}$	Z+jets

Table 6.5: Main background composition of the different lepton channels for the different b-jet multiplicity categories.

The use of the *pseudo-continuous tagging* allows a further selection in the events with two b-tag jets in three exclusive categories. This selection is motivated by the higher sensitivity of the tighter b-tag regions. The different 2-tag categories are: two exclusive loose b-tags, two exclusive medium b-tags and two tight b-tags which are expressed as ‘LL’, ‘MM’ and ‘TT’ respectively. Figure 6.3 summarized the different b-jet categories using a graphical representation.

A final event categorization is done according to different intervals of the transverse momentum of the recoiling vector boson (p_{T}^{V}), which in the 0-lepton channel corresponds to the $E_{\mathrm{T}}^{\mathrm{miss}}$. This is motivated by the higher sensitivity presented at high p_{T}^{V}. The dijet mass 0-lepton uses four p_{T}^{V} bins: $100 < p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$, $120 < p_{\mathrm{T}}^{V} < 160\,\mathrm{GeV}$, $160 < p_{\mathrm{T}}^{V} < 200\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{V} > 200\,\mathrm{GeV}$.

Summarizing, the 0-lepton dijet analysis is divided according to the jet multiplicity in 2- and 3-jets and to the b-jet multiplicity in 1- and 2-tag regions. The 2-tag regions are further split depending on the b-tag efficiency in 2-loose, 2-medium and 2-tight. Finally, all the regions are binned in four p_{T}^{V} regions. This results in a total of 24 regions for the 0-lepton channel.

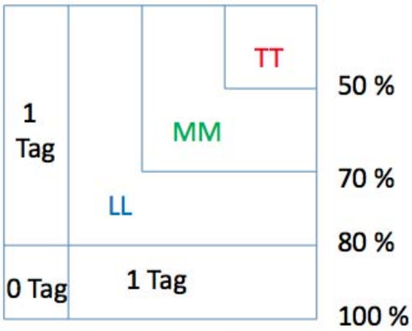


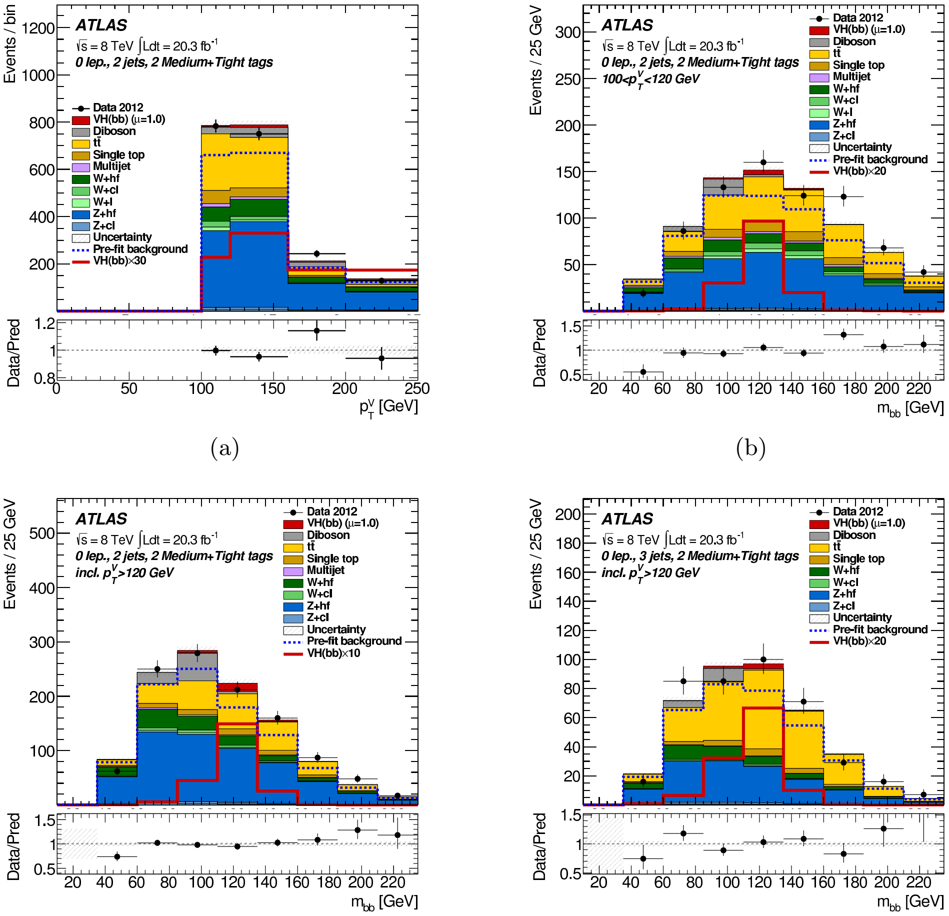
Figure 6.3: A graphical representation of the exclusive b-tagging categories used. 50 %, 70 %, 80 % and 100 % correspond to the different efficiency values.

6.4 Backgrounds composition

Some Standard Model processes can be mis-identified as the Higgs in association with a vector boson representing backgrounds to the analysis. As mentioned before, the main backgrounds present in this analysis are the production of a vector boson in association with jets (V+jets) and the top pair production ($t\bar{t}$). Other small contributions include, the diboson production (WW, ZZ and WZ), single top production, including the three different production modes (s-channel, t-channel and Wt-channel) and the multijet production.

In the case of the V+jet backgrounds, the events are categorized according to their jet flavors as it is explained in Section 5.6.2 in: Vbb, Vbc, Vbl, Vcc, Vcl and Vll. The combination of Vbb, Vbc, Vbl and Vcc is denoted as vector boson plus heavy flavors $V + hf$.

The background processes composition varies with the p_{T}^{V} region, with the jet multiplicity, and with the b-tagging category considered. In general, the 2-jet, tight b-tagging and high p_{T}^{V} categories have higher sensitivity $S/\sqrt{B}$ compared to 3-jet, loose b-tagging and low p_{T}^{V} categories. Figure 6.4 shows the p_{T}^{V} and dijet mass distributions for some selected regions or combination of them to illustrate the background composition of the analysis. As anticipated, the main backgrounds are Zbb, $t\bar{t}$ and, less significant Wbb. The Zbb contribution arises from $Z \rightarrow \nu\nu$ events in addition to the extra radiated b-jets which mimic the signal signature $ZH \rightarrow \nu\nu bb$. In the case of the Wbb in $W \rightarrow l\nu$+jets processes, their contribution to the total background is less than the Zbb one, since the lepton should be first produced outside of the detector acceptance or not identified to mimic the signal signature. Furthermore, the $t\bar{t}$ contribution arises from the top-quark decays via the process $t \rightarrow Wb \rightarrow l\nu b$. When both leptons are not identified, events could mimic the signal as well. The relative $t\bar{t}$ contribution changes, being larger in the low p_{T}^{V} and in the 3-jet categories.



(a) (b)

(c) (d)

Figure 6.4: Dijet mass and p_{T}^{V} distributions in data (points with error bars) and simulation (histograms) with Medium (MM) and Tight (TT) categories combined. The p_{T}^{V} distribution for (a) 2-jet region and the dijet mass distribution for the (b) 2-jet region in the $100 < p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$ interval, the (c) 2-jet region in the three intervals with $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ combined and the (d) 3-jet region in the three intervals with $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ combined.

6.5 Kinematic selection

To suppress the backgrounds and increase the sensitivity $S/\sqrt{B}$ of the analysis, an event kinematic selection is applied. The full list of event selection criteria is given in Table 6.6. Common to all regions is a requirement on the leading jet to have $p_{\mathrm{T}}^{j1} > 45\,\mathrm{GeV}$. Cuts on the $\Delta R(jet_1, jet_2)$ are optimized to profit at high p_{T}^{V} from the collimation between the signal jets and to reduce the V+jets background by applying the minimum cut. $\Delta\phi(V,H)$ variable account for the separation between the vector boson and the Higgs candidate (H), the latter reconstructed from vectorial sum of the two leading p_{T} jets. As mentioned before, correlation studies between relevant variables and the trigger efficiency show a dependence with the sum of the p_{T} of the jets ($\sum p_{\mathrm{T}}^{j_i}$). Therefore, to have a full efficient trigger, $\sum p_{\mathrm{T}}^{j_i} > 120(150)\,\mathrm{GeV}$ is required in events with 2-jets (3-jets) and $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$. The rest of the cuts presented in the table are done to suppress the multijet contribution. For events with $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ a $p_{\mathrm{T}}^{\mathrm{miss}} > 30\,\mathrm{GeV}$, a $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}})$ and a $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jets)$ cuts are sufficient to suppress most of it. Instead for $p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$ regions, multijet events present different kinematics and those cuts suppress the signal. Therefore, the requirements are made on the $\Delta\phi(jet_1, jet_2)$, the $E_{\mathrm{T}}^{\mathrm{miss}}\mathrm{Sig}$, defined as $E_{\mathrm{T}}^{\mathrm{miss}}/\sqrt{p_{\mathrm{T}j_1} + p_{\mathrm{T}j_2}}$, and the output variable of a likelihood-based technique. This technique uses four variables (explained below) to build a multijet background discriminator, giving as output values between 0 and 1, where multijets events are expected to give values close to 1. The variables used to build this likelihood are $\Delta\phi(jet_1, jet_2)$, $\Delta\phi(V,H)$, the p_{T} of the sum of the two leading jets, $p_{\mathrm{T}}^{j_1+j_2}$, and the ratio of this variable and the scalar sum p_{T} of the jets, $p_{\mathrm{T}}^{j_1+j_2}/(p_{\mathrm{T}}^{j_1} + p_{\mathrm{T}}^{j_2})$.

Variable	Dijet Mass Analysis			
p_{T}^{V} (GeV)	100-120	120-160	160-200	> 200
$\Delta R(jet_1, jet_2)$	0.7-3.0	0.7-2.3	0.7-1.8	< 1.4
$\Delta\phi(V,H)$	>2.2	> 2.8		
$\sum p_{\mathrm{T}}^{j_i}$ $N_{jet} = 2$		> 120		
$\sum p_{\mathrm{T}}^{j_i}$ $N_{jet} = 3$		> 150		
$p_{\mathrm{T}}^{\mathrm{miss}}$ (GeV)		> 30		
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}})$		$< \pi/2$		
$\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jets)$		> 1.5		
$\Delta\phi(jet_1, jet_2)$	<2.7			
$E_{\mathrm{T}}^{\mathrm{miss}}\mathrm{Sig}$	>7.0			
Likelihood	>0.5			
$p_{\mathrm{T}}^{j_1}$(GeV)	> 45			

Table 6.6: Event kinematic selection applied in the 0-lepton analysis.

6.6 Background Monte Carlo samples

Background event samples are generated using MC techniques, as explained in Chapter 2, where detailed information of the simulation of the events and the programs used is presented. A complete list of the generators utilized in the different processes can be found in Table 6.7. As an exception, multijet samples are generated using data-driven techniques as will be explained in Section 6.7.4.

To simulate all W+jets and Z+jets at Leading-Order (LO) in QCD, the SHERPA generator with CT10 parton distribution functions (PDFs) [76] is used. The other main background, $t\bar{t}$, is generated using POWHEG with CT10 PDFs, interfaced with PYTHIA6 with CTEQ6L1 PDFs [77] and the Perugia2011C tune [78, 79]. The tune indicates the parametrization used in the simulation of the parton shower, the hadronization and the underlying event. Theoretical cross sections at Next-to-Next-to-Leading-Order (NNLO) for both V+jets [80] and $t\bar{t}$ [81] are used to initially normalize the samples and to optimize the selection, although most of the final normalizations are then constrained from data in the analysis.

The single-top-quark background is formed by the s-channel and t-channel exchange processes and Wt production. The s-channel and Wt processes are simulated using POWHEG, while t-channel uses ACERMC and all are interfaced with PYTHIA6 with CTEQ6L1 PDFs and the Perugia2011C tune. Diboson processes are generated using POWHEG with CT10 PDFs and are interfaced with PYTHIA8 with the AU2 tune [78, 79]. The single-top-quark and diboson cross sections used to normalize the samples are taken from Ref. [82, 83, 84] and Ref. [85], respectively.

6.7 Background modeling

This Section presents the background modeling studies performed in order to validate the different MC simulations of the main background processes. Corrections to improve the description of data distributions are extracted when is needed. Modeling uncertainties associated to each background process are presented in Section 6.8.

6.7.1 Vector Boson+jets

The modeling studies of the W+jets and Z+jets backgrounds are performed using 1- and 2-lepton control regions, respectively. As explained before, these regions allow a more precise constrains of these backgrounds that the one given by the 0-lepton.

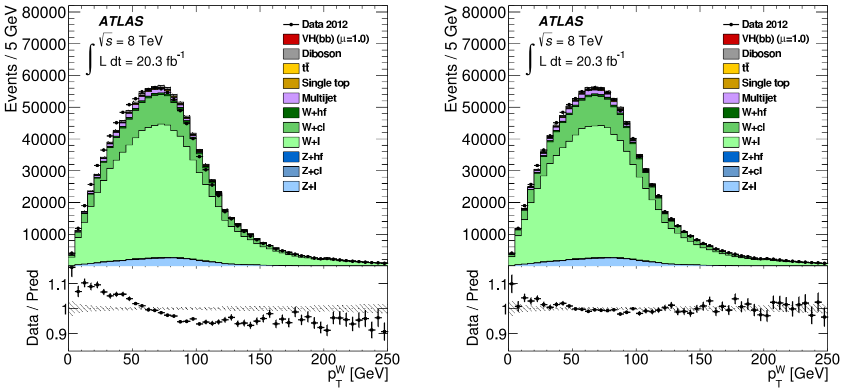
Several studies show a mis-modeling of the W boson p_{T} (p_{T}^{W}) in the SHERPA MC with respect to data for W+jets as can be seen in Figure 6.5(a). It is found that this mis-modeling is strongly correlated with the $\Delta\phi(jet_1, jet_2)$

Process	Generator
Vector boson + jets	
$Z \rightarrow \nu\nu$	SHERPA
$W \rightarrow l\nu$	SHERPA
$Z/\gamma* \rightarrow ll$	SHERPA
Top-quark	
$t\bar{t}$	POWHEG+PYTHIA
s-channel	POWHEG+PYTHIA
Wt–channel	POWHEG+PYTHIA
t-channel	AcerMC+PYTHIA
Di-boson	
WW	POWHEG+PYTHIA8
WZ	POWHEG+PYTHIA8
ZZ	POWHEG+PYTHIA8

Table 6.7: Monte Carlo generators used to simulate the most relevant background processes in the analysis.

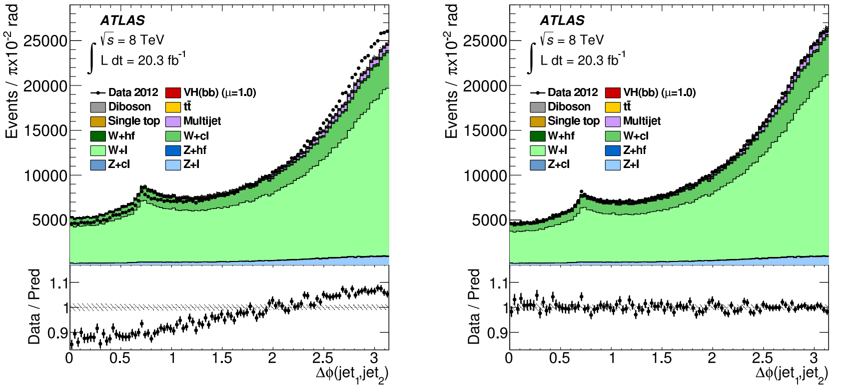
distribution (Figure 6.6). Its origin is related with the usage of a LO order generator and a parton shower that gives a poor description of the kinematics in events with multiple jets in their final states. The solution applied to solve the mis-modeling is a re-weighting correction. This is extracted from a parameterized fit to the ratio of data to MC in the $\Delta\phi(jet_1, jet_2)$ distribution in 0-tag events. The kinematics of the events are different depending on the jet multiplicity and the p_{T}^{V}, thus four different re-weighting functions are derived for 2- and 3-jet, and for $p_{\mathrm{T}}^{W} > 120\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{W} < 120\,\mathrm{GeV}$ regions. The contribution of heavy components in the regions used to extract the correction is negligible, therefore the re-weighting is only applied to Wl and Wcl components in the MCs in all regions and channels. This correction typically ranges between 0.9 and 1.1, depending on the $\Delta\phi(jet_1, jet_2)$ value. Figure 6.5 shows the p_{T}^{W} distribution before and after applying the correction. The impact of the re-weighting in the 2-jet category is to increase the p_{T}^{W} normalization by 7 % in the $p_{\mathrm{T}}^{W} < 120\,\mathrm{GeV}$ regions and reduce it by 5.8 % for $p_{\mathrm{T}}^{W} > 120\,\mathrm{GeV}$. In the 3-jet category, the corresponding increase and decrease are 5.8 % and 2 %, respectively.

A mis-modeling is also observed in the Z boson p_{T} (p_{T}^{Z}) distributions. The origin is the same as in the p_{T}^{W}, but the $\Delta\phi(jet_1, jet_2)$ correction is not enough to mitigate it and, on top of it, an extra correction is applied to the p_{T}^{Z} distributions. As in the previous case, the $\Delta\phi(jet_1, jet_2)$ correction is derived from 0-tag 2-lepton distributions and applied to Zl events. The second correction is applied to Zc and Zb events and it is derived from the p_{T}^{Z} 2-tag 2-lepton distribution. In this case, the 100–150 GeV dijet mass



(a) (b)

Figure 6.5: p_{T}^{W} distribution for the 2-jet 0-tag control region of the 1-lepton, (a) before and (b) after the $\Delta\phi(jet_1, jet_2)$ re-weighting.



(a) (b)

Figure 6.6: $\Delta\phi(jet_1, jet_2)$ distribution for the 2-jet 0-tag control region of the 1-lepton with all p_{T}^{W} regions combined , (a) before and (b) after re-weighting.

interval is excluded to avoid any bias in the region where the Higgs signal is expected.

In order to construct the Vl and Vc backgrounds with b-jets out of the inclusive sample, the *parametrized b-tagging* is applied (see Section 6.2.3.1). The strong dependence in Vcc events between the tagging efficiency and the ΔR (the angular separation between the tagged jet and the closest other jet) produces a mis-modeling in the ΔR distribution once the *parametrized b-tagging* is used. A correction on the $\Delta R < 1$ is applied to Vcc events to mitigate it.

6.7.2 $t\bar{t}$

The $t\bar{t}$ background is a process extensively studied in ATLAS. Some of these studies have shown a disagreement which is consistent with the top p_{T} being too hard in simulation [86]. The disagreement arises from the the difficulties to simulate the kinematics of $t\bar{t}$ events with additional jets in their final state, specially the high jet p_{T} spectrum. Analysis results present the same trend as the one shown by the ATLAS 7 TeV unfolded measurement in Figure 6.7. Therefore the distribution is used to derive a re-weighting correction that is applied to the average top p_{T} distributions. The correction applied corresponds to the ratio between POWHEG+HERWIG and data which corrects the top p_{T} distribution a few percent.

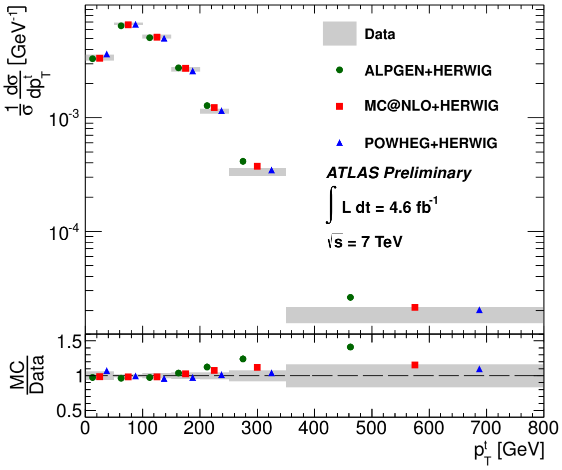


Figure 6.7: The ATLAS 7 TeV $t\bar{t}$ p_T^{top} unfolded result taken from Ref. [87] used to re-weight the ${p_T}^{top}$ distribution in the analysis.

6.7.3 Diboson

The diboson background includes the processes which originate two vector bosons in their final state (WW, WZ and ZZ). From these processes the one which most contributes in to the 0-lepton background is the $ZZ \to \nu\nu b\bar{b}$. The cross section of $V(Z \to b\bar{b})$ is about 5 times the Higgs one, making the diboson background a perfect candidate to validate the analysis against a known process. In order to construct the WW background with b-jets out of the inclusive sample, the *parametrized b-tagging* is applied.

6.7.4 Multijet

The multijet background (MJ) mostly arises from pure QCD processes. Most of the 0-lepton MJ contribution comes from 'fake' $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement provoked by large fluctuations in the jet energy measurement in the calorimeters. Although, the probability of this to happen is low, the huge cross sections of the processes produce a sizable amount of events passing the selection. Simulating the events is not practical thus a data-driven ABCD method is used to estimated the contribution in the different analysis regions.

The ABCD method divides the data in four orthogonal regions, the signal region used in the analysis and three where the MJ is the dominant background. Cuts in $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}})$ and $min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jet)]$ distributions are used to define the regions. As an exception, in the $100 < p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$ regions the $min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jet)]$ distribution is substituted for the Likelihood variable used to suppress the MJ describe in Section 6.5. Correlations of these variables have been probed to be weak. Kinematics of events with no 'fake' $E_{\mathrm{T}}^{\mathrm{miss}}$ are expected to have the directions of $E_{\mathrm{T}}^{\mathrm{miss}}$ and $p_{\mathrm{T}}^{\mathrm{miss}}$ similar. On the other hand, events with 'fake' $E_{\mathrm{T}}^{\mathrm{miss}}$ are expected to have the direction of the $E_{\mathrm{T}}^{\mathrm{miss}}$ closer to the direction of the badly measured jet. Therefore, the regions are defined as follows:

- A: $min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jet)] > 1.5$, $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}}) < \pi/2$
- B: $min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jet)] < 0.4$, $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}}) < \pi/2$
- C: $min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jet)] > 1.5$, $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}}) > \pi/2$
- D: $min[\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jet)] < 0.4$, $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, p_{\mathrm{T}}^{\mathrm{miss}}) > \pi/2$

The contribution of MJ events in the signal region A is then computed using the relation expressed in Eq. 6.1 which assumes that the fraction of events between region A and C is conserved in B and D. The number of events suffer from low statistics in all regions after the 2-tag requirement. To increase them and allow a robust estimation, the b-tagging requirements are dropped from B, C and D and ΔR cuts (see Section 6.5) from B and D regions. To properly estimate the multijet contribution, the effect of the b-tagging requirement is recovered by applying a normalization factor R.

This factor is defined as the ratio of events before after requiring 2-tag in the region D.

$$N_{MJ}(A) = N(C) \times \frac{N(B)}{N(D)} \times R \tag{6.1}$$

The ABCD method estimates a MJ contribution that ranges between 0.5 % and 2 % of the total background depending on the analysis region. By varying the cuts to define the regions and the normalization factor R a systematic uncertainty of 100 % is assigned to the estimation.

6.8 Uncertainties on the modeling of the backgrounds

This Section presents all the systematics uncertainties related with the modeling of the background processes. These uncertainties are associated with both the corrections applied to some processes described in Section 6.7, and the modeling uncertainties related with the MC simulation of events. Uncertainties quoted in this Section correspond to the 1σ prior assigned before the fit to the data.

The different studies to estimate the uncertainties are performed when possible in data control regions where the relative contribution of the process under study is large. This is the case for Z+jets, W+light jets and $t\bar{t}$ backgrounds. When no control regions with enough purity of the process under study are found, uncertainty estimations are performed using different MC generator programs. This is the case for single top, W+heavy jets and diboson processes.

When estimating the source associated with the generator programs used, the tests that can be performed are: the usage of different PDFs sets, variations in the renormalization (μ_R) and factorization (μ_F) scales or the usage a different program to simulate the hard process, the parton shower or the underlying event.

There are several types of systematics. The most generic accounts for possible variations in the overall normalization of a distribution. The uncertainties that accounts for possible migration of events across bins in a distribution are classified as shape uncertainties. This uncertainty is typically estimated for the discriminant variable of the analysis m_{bb} and some p_{T}^{V} distributions. Finally, an uncertainty accounts for migration of events across analysis categories or V+jets contributions. E.g. a 3-to-2-jet ratio uncertainty takes into account the possible migration of 3-jet events to the 2-jet region and vice versa. Another example is the uncertainty that accounts for migration of Wbl events to a Wbb distribution and vice versa.

All background modeling systematics that are going to be discussed below are summarized in Table 6.8.

Z+jets	
Zl normalization, 3/2-jet ratio	5%
Zcl 3/2-jet ratio	26%
Z+hf 3/2-jet ratio	20%
Z+hf/Zbb ratio	12%
$\Delta\phi$, p_{T}^{V}, m_{bb}	S
W+jets	
Wl normalization, 3/2-jet ratio	10%
Wcl, W+hf 3/2-jet ratio	10%
Wbl/Wbb ratio	35%
Wbc/Wbb, Wcc/Wbb ratio	12%
$\Delta\phi$, p_{T}^{V}, m_{bb}	S
$t\bar{t}$	
3/2-jet ratio	20%
High/low-p_{T}^{V} ratio	7.5%
Top-quark p_{T}, m_{bb}	S
Single top	
Cross section	4% (s-,t-channel), 7% (Wt)
Acceptance (generator)	3%–52%
m_{bb}, $p_{\mathrm{T}}^{b_2}$	S
Diboson	
Cross section and acceptance (scale)	3%–29%
Cross section and acceptance (PDF)	2%–4%
m_{bb}	S
Multijet	
Normalization	100%

Table 6.8: Summary of the systematic uncertainties on the background modeling. An “S” symbol is used when only a m_{bb} shape uncertainty is assessed. The % corresponds to the 1σ prior estimated for the systematic as input to the fit.

6.8.1 V+jets modeling systematics

As previously explained, a $\Delta\phi(jet_1, jet_2)$ and p_{T}^{Z} corrections are applied to the V+jets backgrounds. In the case of the W+jets contributions, the correction is applied only to Wl and Wcl components where half of it is assigned as systematic. Although, no correction is applied on Wcc and Wbb backgrounds due to the small relative contributions compared to the light-flavor components in the regions used to the extract the re-weighted. Although, to account for any residual effect, a systematic equal to the full correction to the light components is assigned to these heavy flavor contributions. The Z+jets contributions are instead re-weighted with two corrections. A systematic uncertainty equal to half of the $\Delta\phi(jet_1, jet_2)$ correction is applied to Zl components and equal to the full correction if the components are Zc or Zb. To account for the p_{T}^{Z} corrections another set of systematics equal to half of the correction are assigned to Zl and Zc or Zb.

The normalization and the 3-to-2-jet ratio for the Wl background are taken directly from simulation, both with a 10 % uncertainty. This is based on the agreement observed between data and simulation in the 0-tag regions. Same 10 % systematic is assigned to the 3-to-2-jet ratio in Wcl background. In the case of Zl, the normalization and the 2-to-3-jet ratio uncertainties are estimated from data in 0-tag regions, both with an uncertainty of 5 %. The 2-to-3-jet ratios uncertainties for the Zcl, Zhf and Whf are estimated comparing events generated with SHERPA with samples generated with POWHEG+PYTHIA8, ALPGEN+HERWIG and MC@NLO+HERWIG or ALPGEN. Uncertainties are estimated to be 26 % for Zcl, 20 % for Zhf and 10 % for Whf.

The parton shower and hadronization modeling could have an impact in the relative flavor composition of the samples. Comparison between different generators are also used to estimated the systematic uncertainties associated with the effect. The uncertainties assigned to the Whf samples are 35 % for *bl*/*bb* and 12 % for each of bc/bb and cc/bb. In the case of Zhf samples the error are 12 % for each of *bl*/*bb*, *cc*/*bb* and *bc*/*bb*, with *bl*/*bb*.

In W+jets the shape uncertainties assigned to the m_{bb} and p_{T}^{V} distributions are estimated by comparison of MC generators. This comparison estimates that the m_{bb} can increase at 50 GeV up to 23 % and decrease at 200 GeV down to 28 %. This m_{bb} shape variation impacts the p_{T}^{W} shape and uncertainties are also estimated for this distribution, 9 % increase at $p_{\mathrm{T}}^{W} = 50\,\mathrm{GeV}$ and a 23 % decrease at $p_{\mathrm{T}}^{W} = 200\,\mathrm{GeV}$. Also m_{bb} shape systematics are estimated in Z+jets by comparing data and simulation. The numbers are 3 % increase at 50 GeV and a decrease of 5 % at 200 GeV.

6.8.2 $t\bar{t}$ and single-top modeling systematics

The $t\bar{t}$ process presents mis-modelings corrected by re-weighting the average top p_{T}. Half of this correction is assigned as a systematic uncertainty to $t\bar{t}$ events.

The $t\bar{t}$ is normalized from a fit to data, although the 3-to-2-jet ratio is obtained from the MC, and an uncertainty of 20 % is assigned. From comparison of generator predictions, an extra 7.5 % error is also applied to the $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ regions. These comparisons are carried out between the nominal generator POWHEG+PYTHIA with MC@NLO+HERWIG, POWHEG+HERWIG, ACERMC+PYTHIA and ALPGEN+PYTHIA. In general, higher discrepancies are observed with ALPGEN and are those the one used to estimate the systematic uncertainties.

Shape uncertainties on the m_{bb} distribution are also estimated by comparing the predictions from the MC generators. In the $t\bar{t}$, the shape increases m_{bb} by 3 % at 50 GeV while decreases by 1 % at 200 GeV in the 2-jet regions. In 3-jet events, the result is the same but with opposite signs.

The single top components s-channel, t-channel and Wt-channel are determined with a cross section production uncertainty of 4 %, 4 % and 7 %, respectively [88]. Studies using different event generators show discrepancies in the acceptance of the events. These discrepancies are taken as normalization systematics and they can be as large as 52 % for 2-jet events in the t-channel, of the order of 5 % for Wt-channel or 20 % for the s-channel.

For single top, the m_{bb} shape systematic is estimated by comparing the predictions from different generators. The only non-negligible contribution arises for Wt-channel events with $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$. The uncertainty is estimated to increase the rate by 20 % at 50 GeV and decrease it by 40 % at 200 GeV. For 3-jet it was estimated that the corresponding shifts are 25 % and 20 %. Finally, a third uncertainty is needed for the p_{T} distribution of the second-leading jet in $p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$ and 2-jet regions.

6.8.3 Diboson modeling systematics

The diboson normalization is estimated from MC predictions. Variations in the scales μ_R and μ_F by factors of 2 or 0.5 lead to normalization uncertainties which range between 3 and 29 % depending on the process and the analysis region. Another set of normalization systematics which range from 2 to 4 % account for the errors in the PDFs.

Finally, a m_{bb} shape systematic is assigned comparing the results of the Z line shape in VZ production given by the difference between the nominal generator POWHEG+PYTHIA8 and HERWIG. The difference between the shapes is of 20 % for $m_{bb} = 125\,\mathrm{GeV}$.

6.9 Experimental uncertainties

The relevant experimental systematic uncertainties that are taken into account come from: the luminosity estimation, the trigger selection, the physics object reconstruction, identification, momentum and energy resolution and calibration. From all of them, the ones with higher impact in the analysis are those associated with the b-tagging efficiency and the jet energy scale (JES), both explained in detail in this Section.

The uncertainty on the $\sqrt{s} = 8\,\text{TeV}$ luminosity is 2.8 % [52] and it is applied to the simulated signal and backgrounds (except from multijet estimated from data). An additional 4 % error is also applied to these samples to account for the pile-up correction.

Two sources of uncertainty associated with $E_{\mathrm{T}}^{\mathrm{miss}}$ trigger selection are applied. These uncertainties arises from the method to estimate the trigger efficiency (Section 6.1), one to account for the statistical uncertainty of the method and a second for the differences observed in the two event classes used, $W \to \mu\nu$ and $Z \to \mu\mu$. Their combination gives a relative uncertainty of the order of 3 % of the total background in the $100 < p_{\mathrm{T}}^{V} < 120\,\text{GeV}$ region and below 1 % for $p_{\mathrm{T}}^{V} > 120\,\text{GeV}$.

The impact of systematic uncertainties for electron and muon reconstruction, identification and isolation efficiencies are very small, less than 1 %. The uncertainty associated with the lepton veto efficiency applied in the 0-lepton analysis is negligible.

The jet energy scale uncertainty [65] gather different sources of systematics coming from the jet calibration. These are related with the pile-up dependent corrections, the flavor composition of jets in different event classes and the in-situ corrections. The sources are separated into uncorrelated components. Typically, the systematic uncertainty on the JES ranges from 3 % to 1 % for a 20 GeV and 1 TeV central jet respectively. The calibration of b-jets is affected by an additional 1-2 %. The jet energy resolution (JER) uncertainty [89] is also considered and its contribution is separated for jets and b-jets. The systematic uncertainty on the JER ranges from about 10 to 20 %, depending on the η range, for jets with $p_{\mathrm{T}} = 20\,\text{GeV}$ to less than 5 % for jets with $p_{\mathrm{T}} > 200\,\text{GeV}$. To account for the m_{bb} resolution improvements, explained in Section 5.6.4, a 0.4 % uncertainty is added to the b-jets JER contribution.

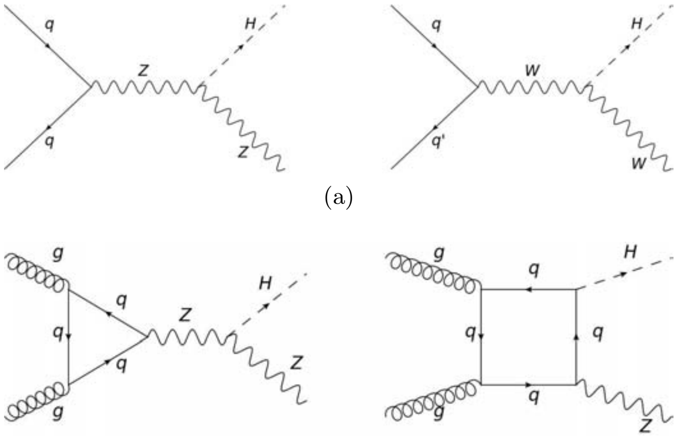
All the uncertainties presented and related with energy and momentum calibration of the physics objects are propagated to the $E_{\mathrm{T}}^{\mathrm{miss}}$. Additional uncertainties on $E_{\mathrm{T}}^{\mathrm{miss}}$ are estimated from the impact of the variations of the scale (8 %) and resolution (2.5 %) of the energy in calorimeter clusters which have not been associated with a reconstructed object.

As explained in Section 6.2.3, b-tagging efficiencies for b-, c- and light-jets are estimated and corrected to match the data estimation. This MC-to-data corrections are close to unity, with uncertainties that range from 2 % to 3 %

over most of the jet p_{T} range, reaching 5 % for $p_{\mathrm{T}} = 20\,\mathrm{GeV}$ and 8 % above 200 GeV. In addition, half of the MC-to-MC SF (see Section 5.6.2) is applied as an uncertainty. Extra uncertainties, which depend on p_{T} and the MV1c weight, are decomposed into uncorrelated components: 10 components to account for b-jets, 15 for c-jets and 10 for light-jets. These uncertainties account for the MC modeling of the $t\bar{t}$ and background processes utilized for the b-tagging calibration process, together with detector related ones, such as lepton identification efficiencies, JES calibration, etc.

6.10 VH signal characterization

The Standard Model VH production is dominated by the quark-quark-initiated (qq) process (see Figure 6.8a). Additionally, a non-negligible contribution arises from higher orders gluon-gluon-initiated (gg) ZH processes (see Figure 6.8b). The 0-lepton analysis tries to target $ZH \rightarrow \nu\nu b\bar{b}$ but a few percentage of the total signal contribution arises from $WH \rightarrow \nu l b\bar{b}$ process due to a production of the lepton outside of the detector acceptance or the presence of a lepton not identified. The acceptance times efficiency of the event selection in the 2-tag signal regions is presented in Table 6.9 for a Higgs of 125 GeV. The contribution from Higgs final state decays different than $b\bar{b}$ is estimated to be less than 1 %.



(a)

(b)

Figure 6.8: Leading Order Feynman diagrams for (a) quark-quark-initiated VH production and gluon-gluon-initiated ZH production.

$m_H = 125\,\mathrm{GeV}$ at $\sqrt{s} = 8\,\mathrm{TeV}$			
Signal	Generator	$\sigma \times BR$ (fb)	Acceptance·ϵ (%)
$q\bar{q} \rightarrow WH \rightarrow \nu lbb$	PYTHIA8	131.7 fb	0.3
$q\bar{q} \rightarrow ZH \rightarrow \nu\nu bb$	PYTHIA8	44.2 fb	3.8
$gg \rightarrow ZH \rightarrow \nu\nu bb$	POWHEG+PYTHIA8	3.8 fb	5.0

Table 6.9: The 0-lepton analysis signal processes, generators used, cross section times branching ratios and acceptances times efficiencies. The acceptance is calculated as the fraction of events remaining in the combined 2-tag signal regions after the full event selection.

6.10.1 Signal Monte Carlo samples and cross section computation

The qq-initiated (W/Z)H events are simulated using PYTHIA8 with the CTEQ6L1 PDFs and the AU2 tune for the parton shower, hadronization, and underlying event. While the QED final-state radiation is simulated using the PHOTOS [90] program. In addition, gg-initiated ZH and crosschecks to qq-initiated processes are simulated using POWHEG and it is computed at NLO using the Multi-scale improved (MiNLO) [91] to generate $HW/HZ+$ 0- and 1-jet events. In this case the generator uses the CT10 PDFs and is interfaced with PYTHIA8 with the AU2 tune.

Several signal samples with different Higgs boson masses are simulated within the range 100 to 150 GeV in steps of 5 GeV. Cross sections for all signal masses and processes are taken from Ref. [92] and the branching ratios (BR) are calculated with HDECAY [93]. Table 6.9 presents the cross section times BRs for $m_H = 125\,\mathrm{GeV}$ processes at 8 TeV. The simulated W and Z decays include all charged lepton flavors thus the leptonic τ decay can also be selected in the analysis. For the Higgs boson, only the $b\bar{b}$ decay mode is used in the analysis.

Both WH and ZH qq-initiated production cross sections and associated uncertainties are computed at the NNLO in QCD [94, 95, 96] and with electroweak corrections at NLO [97]. An additional electroweak NLO normalization correction as a function of the p_{T}^{V} is applied [98] to better model the recoil between the vector boson and the Higgs. Figure 6.9 shows these corrections and the size of the associated uncertainty. The gg-initiated ZH are calculated at NLO [99] increasing its cross section at LO by 5 %.

6.10.2 Signal systematics

A complete study of the systematic uncertainties of the Higgs signal has been performed. Table 6.10 summarizes the uncertainties discussed in this

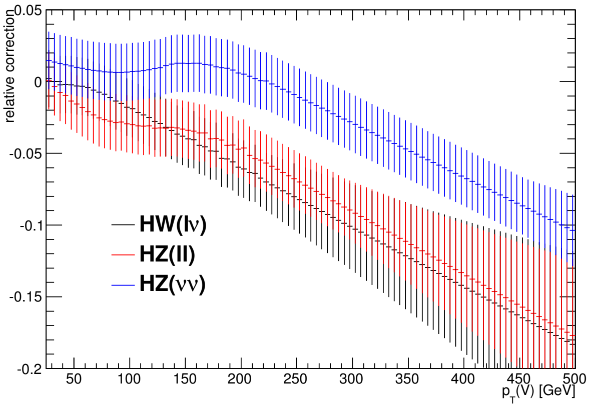


Figure 6.9: Relative NLO electroweak corrections to VH cross sections and the size of the associated uncertainties is shown as a function of p_{T}^{V}.

Section. As previously explained, uncertainties quoted in this Section correspond to the 1σ prior assigned before the fit to the data.

The uncertainties associated with the cross section computation include those related with the scales μ_R and μ_F, and the PDFs. In the case of qq-initiated processes, the uncertainties are estimated to be 1 % and 2.4 %, respectively [100]. For the gg-initiated process, the same errors are estimated to be 50 % and 17 %. An extra uncertainty to account for electroweak NLO normalization correction is applied. This uncertainty ranges approximately from 2 % at low p_{T}^{V} up to 2.5 % at high p_{T}^{V} (see Figure 6.9). The relative uncertainty on the 125 GeV Higgs boson branching ratio to $b\bar{b}$ is 3.3 % [93].

Several acceptance uncertainties are estimated comparing the total number of events that pass the event selection once the scales or the PDFs are changed. Variations in the scales μ_R and μ_F by factors of 2 or 0.5 lead to uncertainties of 3.0 %, 3.4 % and 1.5 % for $qq \to WH$, $qq \to ZH$ and $gg \to ZH$ respectively, for the 2- and 3-jet categories combined, and of 4.2 %, 3.6 % and 3.3 % for the 3-jet category.

Since the p_{T}^{V} shape distribution is seen to be affected as well by the variation of the scales, a shape uncertainty is estimated. This systematic produces an increase of the p_{T}^{V} of 1 % at 50 GeV and a decrease of 3 % at 200 GeV in the $qq \to (W/Z)H$ samples. The same estimations give 2 % and 8 % respectively for the $gg \to ZH$ samples. Finally, acceptance uncertainties are derived varying the PDFs ranging from 2 % in the 2-jet $gg \to ZH$ to 5 % in the 3-jet $qq \to ZH$.

The results given by the nominal MC generator used, PYTHIA, and HERWIG are compared to estimate the uncertainties on the parton shower modeling. This gives an acceptance error of 8 %, except in the 3-jet and $p_{\mathrm{T}}^{V} > 120$ GeV regions, where the error estimated is 13 %.

Signal uncertainties	
Cross section (scale)	1% (qq), 50% (gg)
Cross section (PDF)	2.4% (qq), 17% (gg)
Branching ratio	3.3 %
Acceptance (scale)	1.5%–3.3%
3-jet acceptance (scale)	3.3%–4.2%
p_{T}^{V} shape (scale)	S
Acceptance (PDF)	2%–5%
p_{T}^{V} shape (NLO EW correction)	S
Acceptance (parton shower)	8%–13%

Table 6.10: Summary of the systematic uncertainties on the signal modeling. An "S" symbol is used when only a m_{bb} shape uncertainty is assessed. The % corresponds to the 1σ prior estimated for the systematic as input to the fit.

Chapter 7

VH search statistical procedure

This Chapter describes the statistical procedure for the VH search in the dijet mass analysis. This combines into a complex fit the information provided by the 0-, 1- and 2-leptons channels described in Chapter 6. As a result of the fit, Higgs signal strength (μ), 95 % CL limits, and p_0 values are obtained for the $VH(H \to b\bar{b})$ process. All statistical tools used in the search are described in Chapter 3.

7.1 Global fit

The statistical procedure combines distributions from the three channels (0- ,1- or 2-lepton) in a global fit performed maximizing a likelihood. Two distributions are used to construct the likelihood: the invariant mass of the two b-jets m_{bb} and the b-tagging MV1c discriminant described in Section 5.6.1. As a reminder, the analysis is divided in three channels, up to five p_{T}^{V} intervals, two jet multiplicities (2 and 3-jet), two b-jet multiplicities (1- and 2-tag) and three b-tagging categories (LL, MM and TT) (see Section 6.3).

Distributions are chosen according to the analysis region: the MV1c distribution in 1-tag regions and m_{bb} distribution in the 2-tag one. Table 7.1 presents a summary of the distributions and the analysis regions used. As an exception, in the 0-lepton channel low p_{T}^{V} interval (100 -120GeV) only uses the 2-jet category since bad modeling is found in the 3-jet one. The use of the MV1c distributions is motivated by the high c-jet rejection provided by the b-tagging algorithm. This allows a better constrain of V+c-jets contributions. The MV1c p_{T}^{V} intervals are combined in the ranges with $p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$. In total, 81 m_{bb} distributions and 11 MV1c distributions are used.

Channel			0-lepton	1-lepton	2-lepton
2-/3-jet p_{T}^{V} intervals		1-tag	$MV1c$		
	LL MM TT	2-tag	m_{bb}		

Table 7.1: The distributions used in each region by the likelihood fit in the dijet mass analysis for the 8 TeV data. Four p_{T}^{V} intervals are used in the 0-lepton and five in 1- and 2-lepton channels. These distributions are input to the fit for the 2-jet and 3-jet categories separately.

7.1.1 Treatment of nuisance parameters

The systematics uncertainties on the signal and the background expectations are described by NPs. The analysis divides them in two types: floating parameters and parameters with priors. A floating parameter is generally associated with the cross section and acceptance where absolute ignorance of the rate is assumed and are completely determined from data. Flat pdfs are use to describe them. On the other hand, when a prior is used, it is assumed that there is a reasonable range within which the true value is believed to lie. When one of this NP could affect the normalization of the expected values, it is described with a log-normal prior. If the NP could affect the shape of the m_{bb} distribution it is describe with a Gaussian prior. The prior is added as a penalty term to the likelihood, $L(\mu, \theta)$. This term decreases as soon as θ is shifted away from its nominal value.

The fit is expected to accommodate the values and uncertainties of the NP in order to maximize the likelihood function value. Therefore, the NPs could be pulled away from its initial value and its uncertainty could be reduced with respect to its prior uncertainty. The fit could also developed correlations between initially uncorrelated NPs.

Statistical uncertainties in the MC are estimated to vary the expected significance of the analysis less than 1 %. Therefore, to decrease the number of NPs and computational time, these uncertainties are neglected and only the modeling and experimental uncertainties are taken into account. In total, the fit has to handle almost 170 NPs. Table 7.2 brings together all background and signal modeling systematics and their priors previously described in Sections 6.8 and 6.10.2, respectively. In addition, all experimental uncertainties applied to the analysis were presented in Section 6.9.

7.1.2 Background normalizations

The analysis is designed with a complex categorization. This provides distributions to the fit with the statistical power to constrain the backgrounds.

Signal	
Cross section (scale)	1% (qq), 50% (gg)
Cross section (PDF)	2.4% (qq), 17% (gg)
Branching ratio	3.3 %
Acceptance (scale)	1.5%–3.3%
3-jet acceptance (scale)	3.3%–4.2%
p_{T}^{V} shape (scale)	S
Acceptance (PDF)	2%–5%
p_{T}^{V} shape (NLO EW correction)	S
Acceptance (parton shower)	8%–13%
Z+jets	
Zl normalization, 3/2-jet ratio	5%
Zcl 3/2-jet ratio	26%
Z+hf 3/2-jet ratio	20%
Z+hf/Zbb ratio	12%
$\Delta\phi$, p_{T}^{V}, m_{bb}	S
W+jets	
Wl normalization, 3/2-jet ratio	10%
Wcl, W+hf 3/2-jet ratio	10%
Wbl/Wbb ratio	35%
Wbc/Wbb, Wcc/Wbb ratio	12%
$\Delta\phi$, p_{T}^{V}, m_{bb}	S
$t\bar{t}$	
3/2-jet ratio	20%
High/low-p_{T}^{V} ratio	7.5%
Top-quark p_{T}, m_{bb}	S
Single top	
Cross section	4% (s-,t-channel), 7% (Wt)
Acceptance (generator)	3%–52%
m_{bb}, $p_{\mathrm{T}}^{b_2}$	S
Diboson	
Cross section and acceptance (scale)	3%–29%
Cross section and acceptance (PDF)	2%–4%
m_{bb}	S
Multijet	
Normalization	100%

Table 7.2: Summary of the systematic uncertainties on the signal and background modeling. An “S” symbol is used when only a m_{bb} shape uncertainty is assessed. The % corresponds to the 1σ prior estimated for the systematic as input to the fit.

Sufficient statistical power is obtained to normalize the $t\bar{t}$, Wbb, Wcl, Zbb and Zcl background contributions with floating parameters (see the V+jets flavor composition explained in Section 6.4). The rest of the contributions (Vl, single-top and diboson processes) are constrained within their uncertainties using NP with priors.

The normalizations of the Wbb, Wcl, Zbb and Zcl backgrounds are correlated among channels. Therefore, the normalization of a given process is mainly driven by the distributions in the regions with the highest relative contribution of that process. In the case of the Wbb and Zbb, these regions are the ones with 2 b-jets (2-tag) in the 1-lepton and 2-lepton channels, respectively. The Wcl and Zcl normalizations are mainly driven by 1-tag distributions in the same channels.

The $t\bar{t}$ process is normalized independently in each lepton channel using distributions with 2 jets. The reason to uncorrelate it among channels is driven by the fact the 2-jet samples for each channel are very different. The 0-lepton channel is limited to $p_{\mathrm{T}}^{V} > 100\,\mathrm{GeV}$ and mostly accepts $t\bar{t}$ events coming from $(W \to l\nu)b$ (fully leptonic) decays with the two leptons undetected and $(W \to qq')b$ (semileptonical) decays with a missed lepton (usually a τ lepton) and a missed light-quark jet. The 1-lepton channel targets fully leptonic decays with one of the leptons undetected and $t\bar{t}$ events where one of the top quarks decays fully leptonic and the other semileptonically with a missed light-quark jet. Finally, the 2-lepton channel $t\bar{t}$ contribution is coming from both top quarks decaying fully leptonically.

7.1.3 Higgs signal extraction from data

The Higgs signal strength is obtained from a fit to data ($\hat{\mu}$). The $\hat{\mu}$ is estimated by maximizing the likelihood with respect to all parameters in the fit. Its uncertainty σ_μ is obtained by varying the function $2\ln\tilde{\lambda}(\mu)$ by one unit, and the corresponding μ values delimit the $\pm 1\sigma$ range.

7.1.4 Correlation of nuisance parameters

The number of NPs is larger than the sources of systematic uncertainties considered. This comes from the fact that the impact of a NP, a-priori correlated, can be different across background processes or regions with very different kinematics. The solution is to uncorrelate the NP to avoid undesirable constrains in our results. The correlations of the different NPs in the analysis have been carefully designed. An example of this is the uncertainty related with the $\Delta\phi$ correction applied to V+jets processes (see Section 6.7.1). Since the correction is estimated independently for events with 2- and 3-jets, the associated NP is uncorrelated between 2- and 3-jets distributions, although the source of the uncertainty is the same.

Another uncorrelation of the same systematic source arises from the $t\bar{t}$

contributions in the 2-tag 3-jet region for the different channels. In the 0- and 1-lepton channels it is likely that a jet from a $t \to b(W \to qq)$ decay is missed while in the 2-lepton channel it is likely that an ISR or FSR jet is selected. Therefore two NPs are used, one correlated among the 0- and 1-lepton analyses and a second one for the 2-lepton.

7.1.5 Technical implementation

The statistical fitting machinery is based on the RooStats framework [101, 102] which implements all the statistical procedures: the construction of the likelihood from the bins of the input distributions involving the number of data events and expected signal and background yields; the maximization of the likelihood, by properly treating the background normalization of the different samples and their systematics uncertainties; and the extraction of the signal strength, and computation of the discovery significance or upper limits on μ.

7.2 Fit model simplifications

The VH analysis constructs a complex likelihood function that needs to be maximized to search for the Higgs boson. It includes 92 m_{bb} and MV1c nominal distributions. In addition and in order to account to account for the systematics uncertainties, most of the 170 NPs are introduced as distributions that contain up-and-down ($\pm 1\sigma$) variations of the nominal. In conclusion, this fit machinery could end up handling around fifteen thousand distributions.

The maximization of this complex likelihood is time and resource consuming, and can lead to instabilities in the results obtained. Therefore, three methods are applied to simplify it without affecting the physics result. The first one reduces the number of bins per m_{bb} distribution. The second one rejects the systematic uncertainties with a negligible impact on the final results. Finally, large local fluctuations in the distributions used to asset systematics are smoothed out to avoid instabilities.

7.2.1 The m_{bb} discriminant transformation

To reduce the number of bins from the m_{bb} distributions, a procedure to rebin them is performed. The tails of the m_{bb} distribution present high statistical uncertainties, while giving little information about the background composition or the background shape to the fit. This statistical uncertainties could be reduced by using wider bins, however this reduces the sensitivity in those regions eventually enrich in signal. A feasible solution to reduce the number of bins, increase the sensitivity, and decrease the statistical error in the distribution tails is the use of bins with a variable size.

Channel	2-jet	3-jet
0-lepton	$z_s = 6$ $z_b = 2$	$z_s = 4$ $z_b = 2$
1-lepton	$z_s = 6$ $z_b = 2$	$z_s = 4$ $z_b = 2$
2-lepton	$z_s = 4$ $z_b = 4$	$z_s = 2$ $z_b = 2$

Table 7.3: Optimized free parameters used in the transformation algorithm of the m_{bb} distribution.

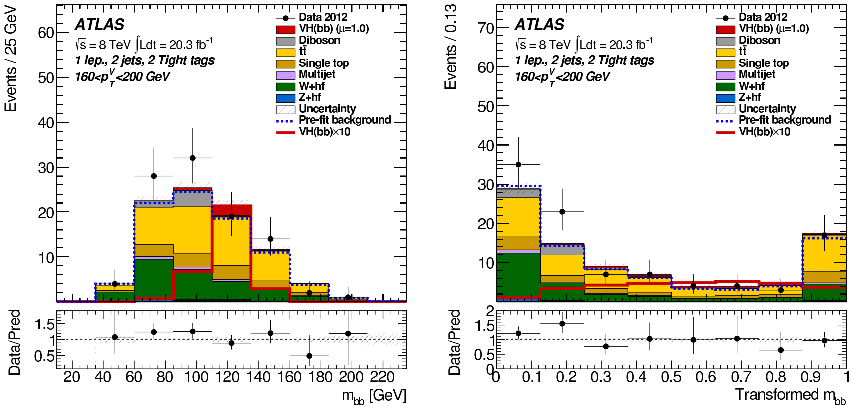
The procedure to rebin the m_{bb} distribution with variable bins size starts with fine-binned histograms. Bins are then merged moving from high to low m_{bb} values using the algorithm

$$Z = z_s \frac{n_s}{N_s} + z_b \frac{n_b}{N_b}, \tag{7.1}$$

where N_s and N_b correspond to the total number of signal and background events in the histogram; n_s and n_b are the total number of signal and background events in a certain bin interval; and z_s and z_b are free parameters of the algorithm. The procedure calculates Z in a certain interval of bins starting from the high m_{bb} bins and adding bins until the condition $Z > 1$ is fulfil. The interval is then remap in a new bin and the procedure is repeated without using the remapped bins until no bins are left. The z_s and z_b values are used to define the relative signal and background compositions in the remapped bins. The free parameters of the transformation algorithm are optimized for 2- and 3-jet regions and for each lepton channel separately and their values are summarized in Table 7.3. The effect of the transformation on the m_{bb} distributions can be seen in Figure 7.1 where an example of a 1-lepton region using a fix bin size and the result of the rebin algorithm are shown.

7.2.2 Pruning and smoothing of systematic uncertainties

To reduce the number of distributions used to construct the likelihood, systematic uncertainties with a negligible impact on the final results are neglected in the fit. A normalization uncertainty is removed if the variation which respect to the nominal is below 0.5 %, as such small variation is tested to not affect the result. In the case of shape uncertainties, the same criteria is applied if the variation is below 0.5 % in all bins. Shape uncertainties with same-sign variations are also pruned. This is performed to avoid instabilities in the fit and after carefully study the source. Finally, systematics variations which impact the total background prediction by less than 0.5 % in regions where the signal contribution is less than 2 % of the total background are also dropped. This pruning procedure reduces the nominal number of systematic variation distributions by a factor of two.



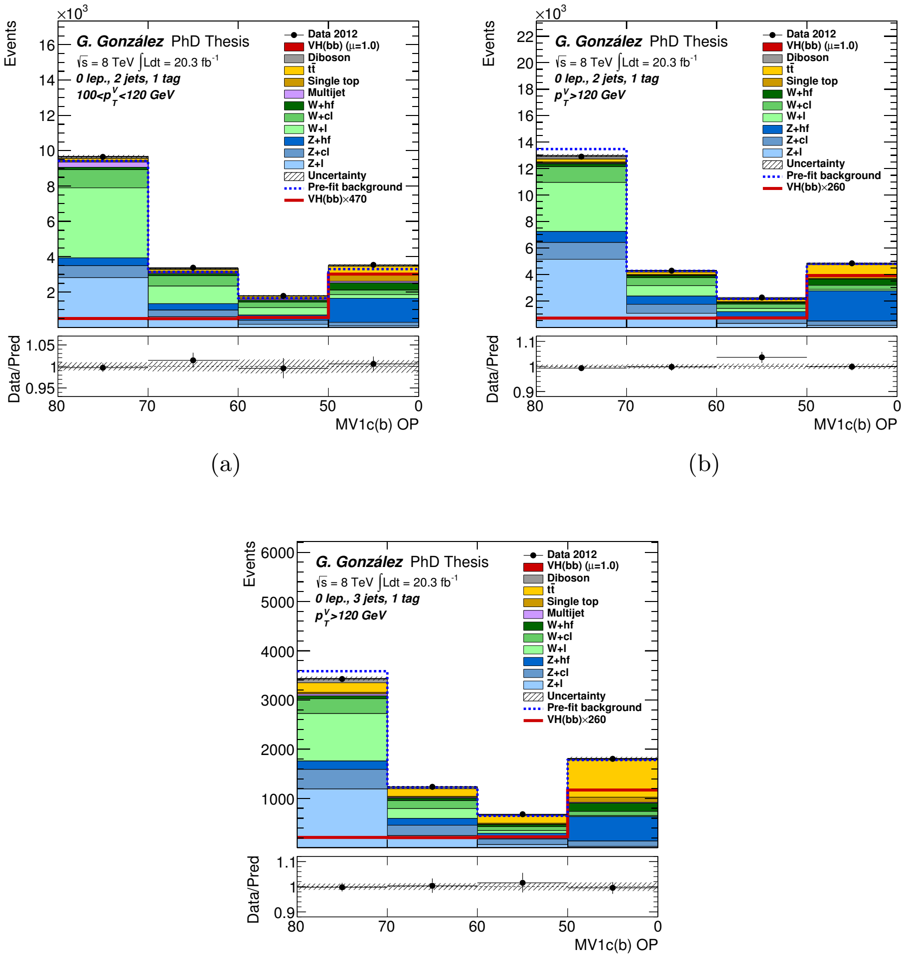
(a) (b)

Figure 7.1: The m_{bb} distributions for the 1-lepton channel and the 2-jet 2-tag TT category for the $160 < p_T^V < 200$ GeV region (a) before and (b) after applying the transformation of the histogram bins.

Some MC samples suffer from limited statistics and this can cause large local fluctuations in the distributions of the nominal and/or the systematic variations. In the cases where only the systematic variation is affected, e.g. the JES uncertainties, a smoothing algorithm is applied to the distributions to mitigate the fluctuations. This smoothing algorithm merges bins until the statistical uncertainty in each bin is less than 5 %.

7.3 Results of the global fit

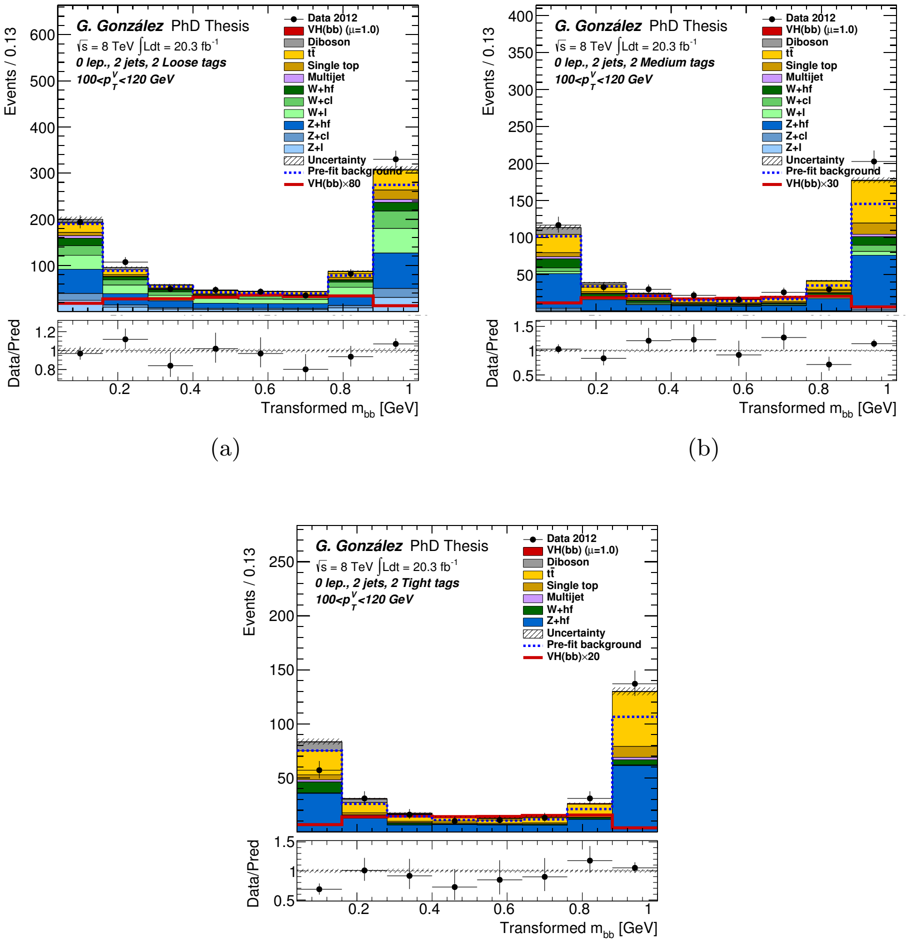
This Section presents the 0-lepton MV1c and m_{bb} distributions obtained from the combined fit of all VH analysis regions. The fit is performed using all the data recorded at 8 TeV during Run I. Figure 7.2 shows the MV1c distributions for the 1-tag region. The 2-jet category is binned in $p_T^V > 120$ GeV and $100 < p_T^V < 120$ GeV, while the 3-jet one has a single bin of $p_T^V > 120$ GeV. Figures 7.3 to 7.9 present the m_{bb} distributions. They are separated in the p_T^V bins $100 < p_T^V < 120$ GeV, $120 < p_T^V < 160$ GeV, $160 < p_T^V < 200$ GeV and $p_T^V > 200$ GeV. Figures alternate the 2- and 3-jet categories besides the bin $100 < p_T^V < 120$ GeV, where only the 2-jet category is used. Each Figure shows the different b-tagging categories: (a) the 2-tag 2-loose jets, (b) 2-tag 2-medium jets and (c) 2-tag 2-tight jets. All distributions show a good agreement between data and MC expectation.



(a) (b)

(c)

Figure 7.2: The output of the MV1c b-tagging algorithm is shown for the b-tag jet, where the left bin boundaries denote the operating points (MV1c(b) OP) of the MV1c b-tagging algorithm as defined in Section 5.6.1, corresponding to b-tagging efficiencies of 80 %, 70 %, 60 %, 50 %. In (a) 2-jets $100 < p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$, in (b) 2-jets $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ and in (c) 3-jets $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ regions. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

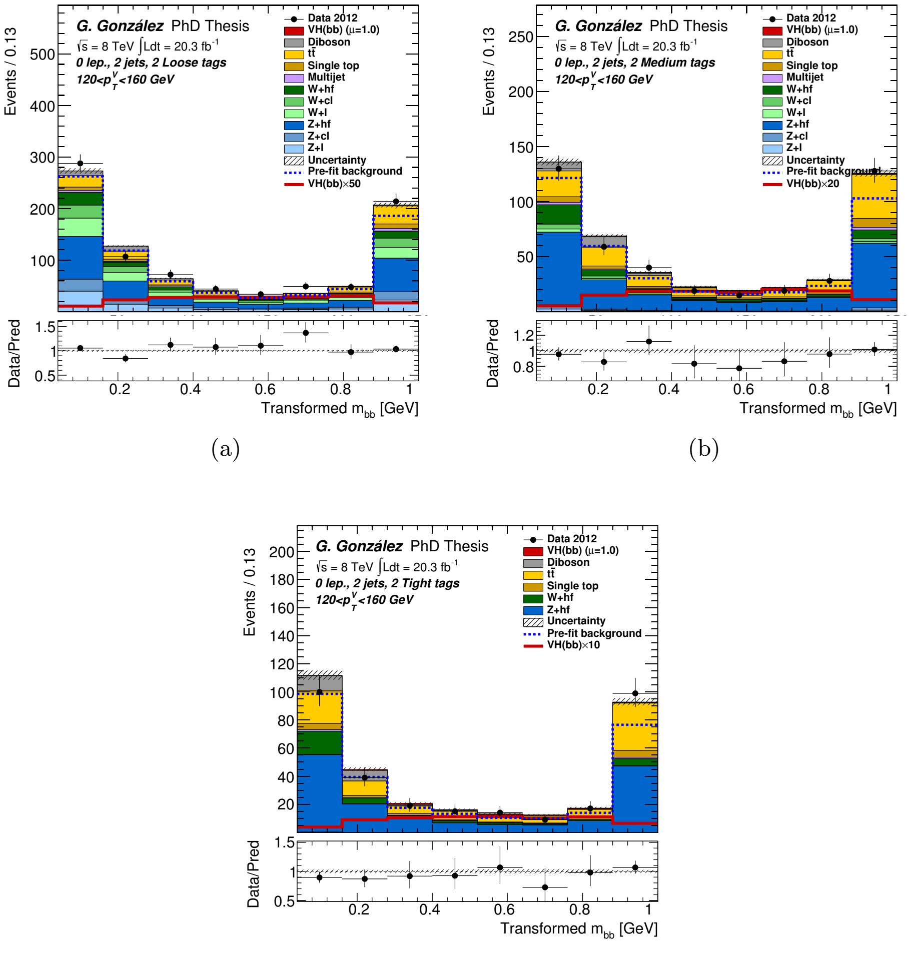


(a)

(b)

(c)

Figure 7.3: The m_{bb} transform distributions for following 2-jet $100 < p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$ signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

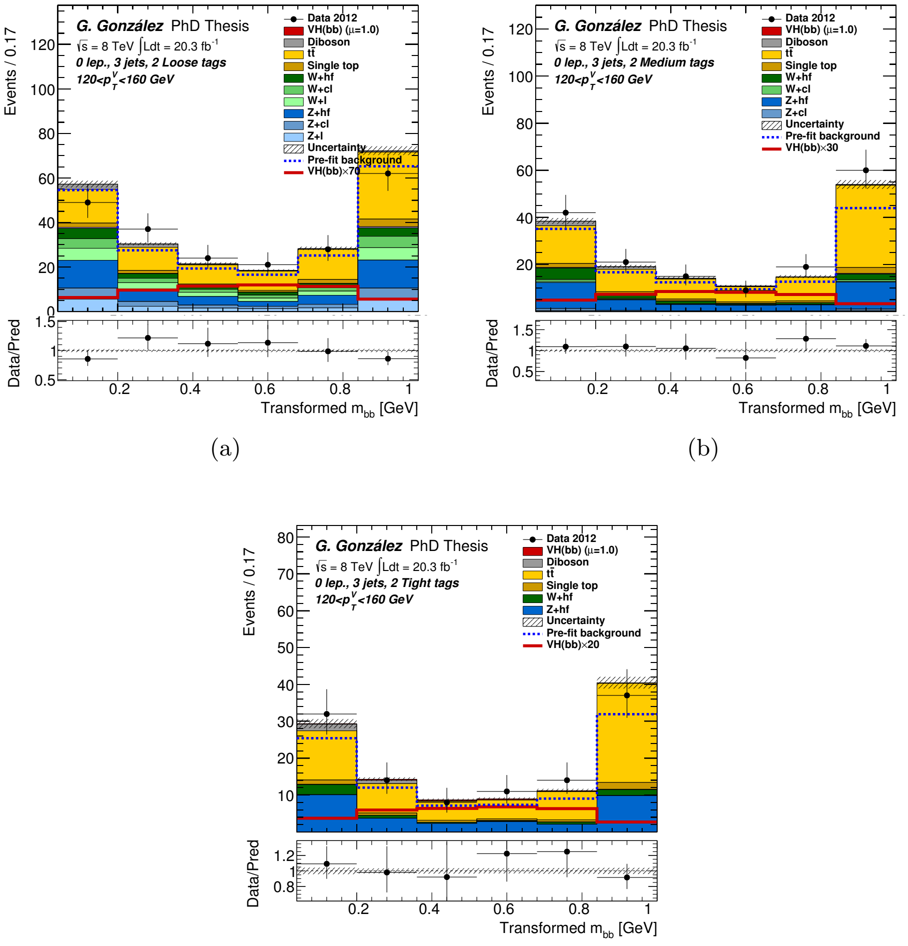


(a)

(b)

(c)

Figure 7.4: The m_{bb} transform distributions for following 2-jet $120 < p_{\mathrm{T}}^{V} < 160\,\mathrm{GeV}$ signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

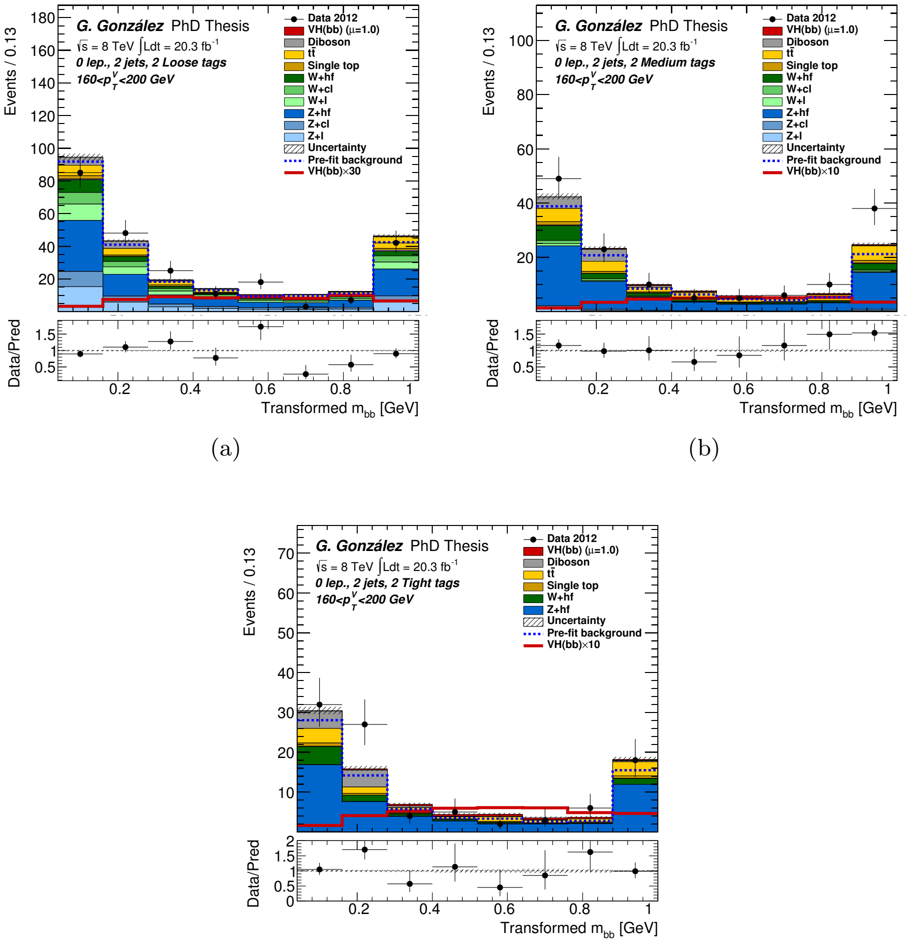


(a)

(b)

(c)

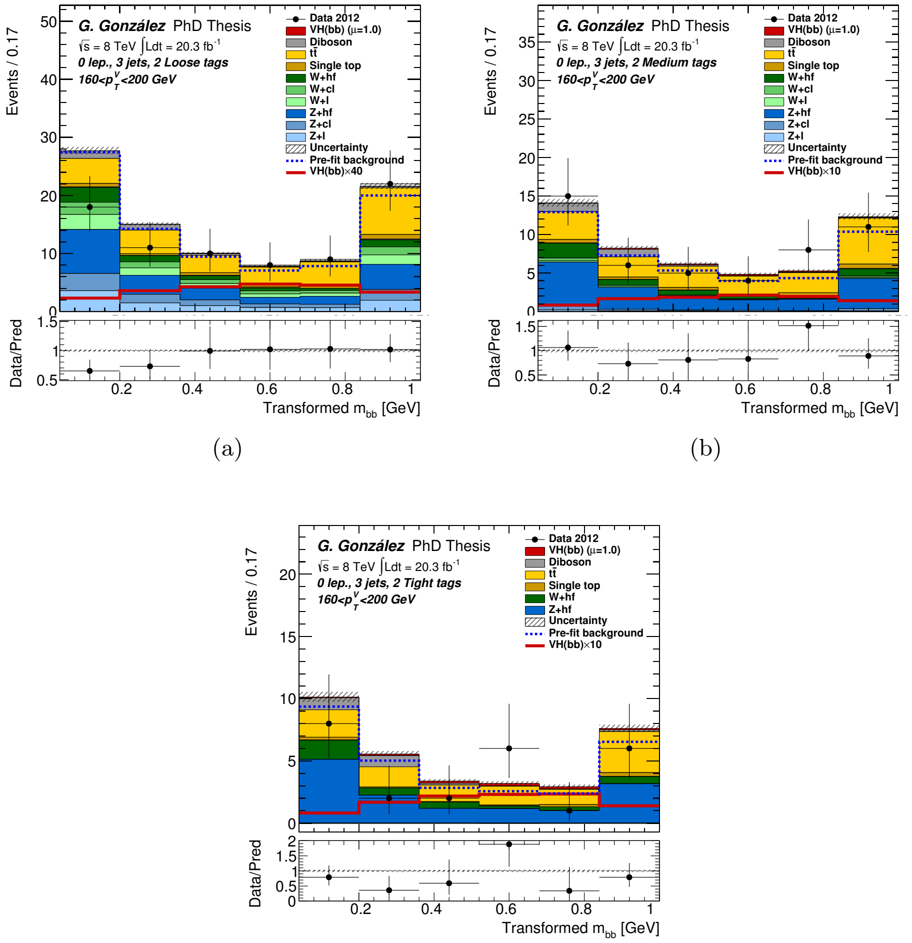
Figure 7.5: The m_{bb} transform distributions for following 3-jet $120 < p_{\mathrm{T}}^{V} < 160\,\mathrm{GeV}$ signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.



(a) (b)

(c)

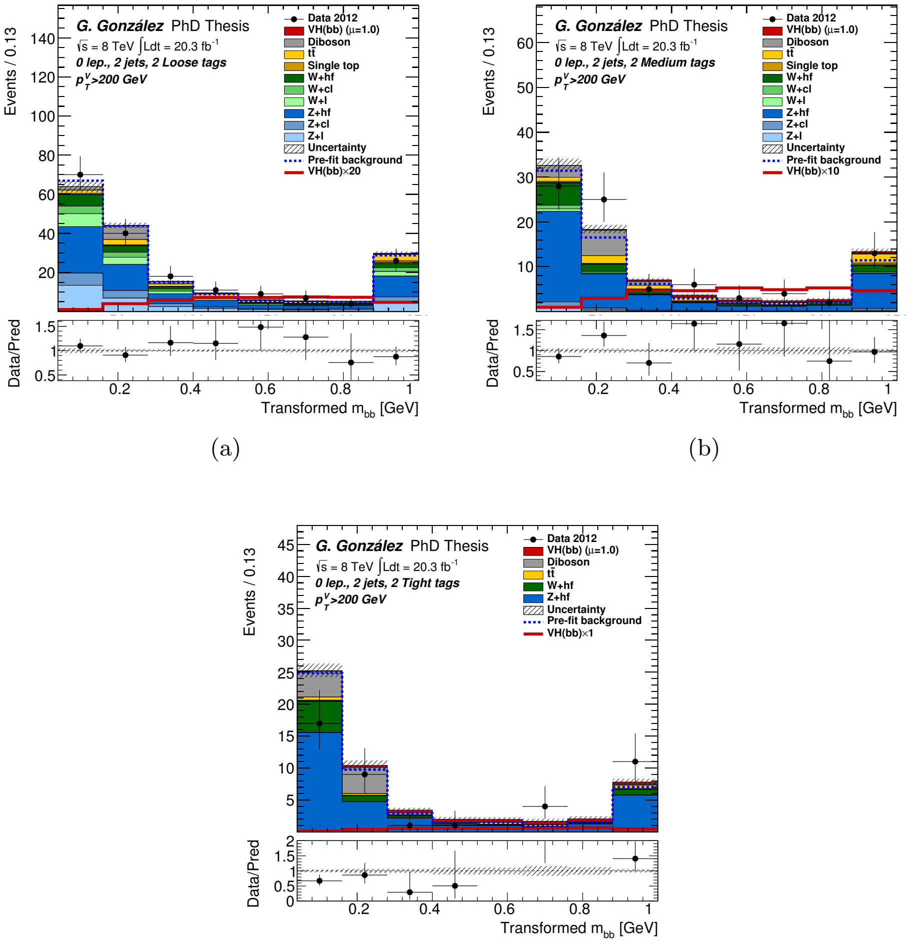
Figure 7.6: The m_{bb} transform distributions for following 2-jet $160 < p_{\mathrm{T}}^{V} < 200\,\mathrm{GeV}$ signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.



(a) (b)

(c)

Figure 7.7: The m_{bb} transform distributions for following 3-jet $160 < p_{\mathrm{T}}^{V} < 200\,\mathrm{GeV}$ signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

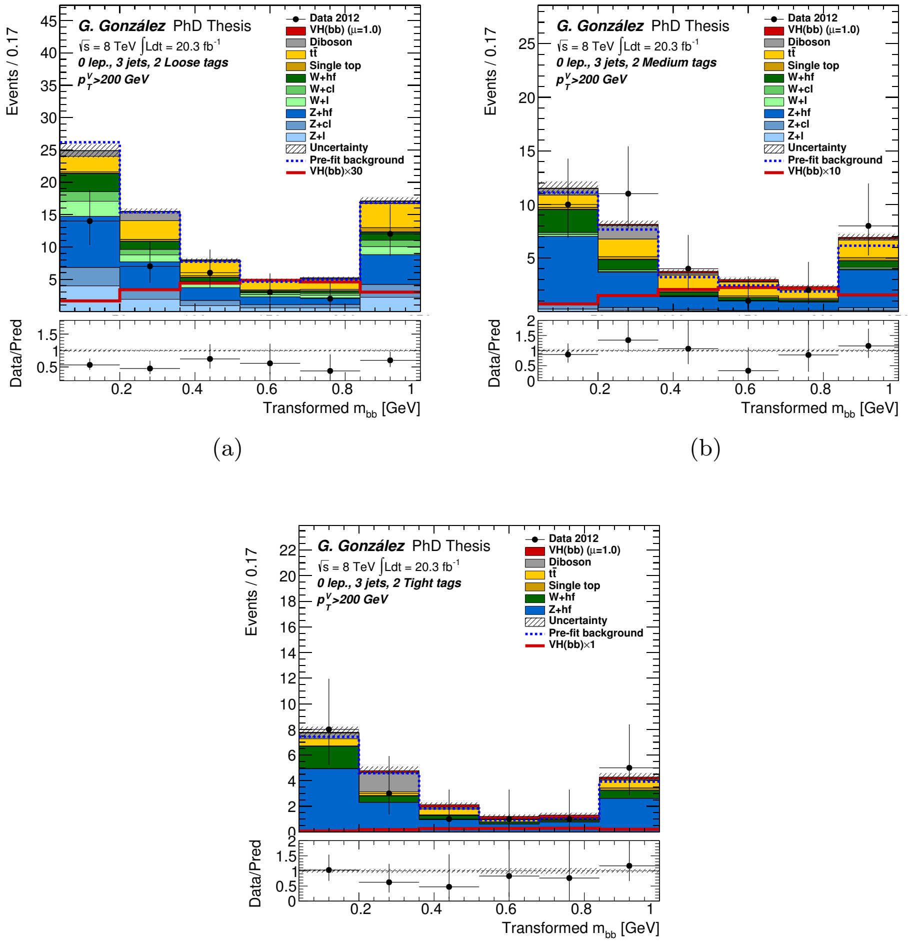


(a)

(b)

(c)

Figure 7.8: The m_{bb} transform distributions for following 2-jet $p_{\mathrm{T}}^{V} > 200\,\mathrm{GeV}$ signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.



(a) (b)

(c)

Figure 7.9: The m_{bb} transform distributions for following 3-jet $p_T^V > 200$ GeV signal regions: (a) 2-tag LL, (b) 2-tag MM and (c) 2-tag TT. The Higgs boson signal ($m_H = 125$ GeV) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

7.4 Fit crosschecks

The behaviour of the global fit is evaluated by several checks. First, pseudodata, derived from the nominal MC expectation, are used to understand possible arising features without introducing any bias from the observation of real data results. Once a robust fit model is constructed, checks are performed in data and results extracted.

One of the studies performed tests how much each NP is pulled away from its initial value, and the correlations developed between initially uncorrelated systematic uncertainties. Usually, features arise from propagated constrains between partially correlated kinematic regions. Generally, the solution adopted is to uncorrelate the NP among the regions. Examples were previously presented with the studies on the NPs associated with $\Delta\phi$ correction and the normalization of the 3-jet $t\bar{t}$ distributions in Section 7.1.4.

Another procedure performed to understand the impact of the NPs is the so-called NP ranking. This procedure identifies the NP with larger impact on the determination of the signal strength μ. After the maximum likelihood value is found, each NP is pulled and fixed to the $\pm 1\sigma$ values while the other parameters are allowed to vary to properly take into account the correlations between systematic uncertainties. Then the magnitude of the shift in the new fitted signal strength $\hat{\mu}$ is used to rank the impact of the NP in the final results.

Figure 7.10 shows the ranking plot obtained with the dijet mass analysis applied to the 8 TeV data set with the NPs that more impact the $\hat{\mu}$ value. As expected, most of the normalization NPs related with the backgrounds in the most sensitive regions of the analysis are in the list. These correspond to the 2-lepton $t\bar{t}$, Zbb and Wbb normalization NPs, as well as, the NPs associated with the migration of events among Zbb and Zbl distributions, or Wbb and Wbl. In the top of the list is the NP associated with the m_{bb} shape of the Wbb and Wcc distributions. As expected, among the different experimental uncertainties, the list includes NPs associated with b-tagging procedure and the b-jet energy resolution. Theoretical uncertainties associated with the Higgs signal modeling are also present in the list.

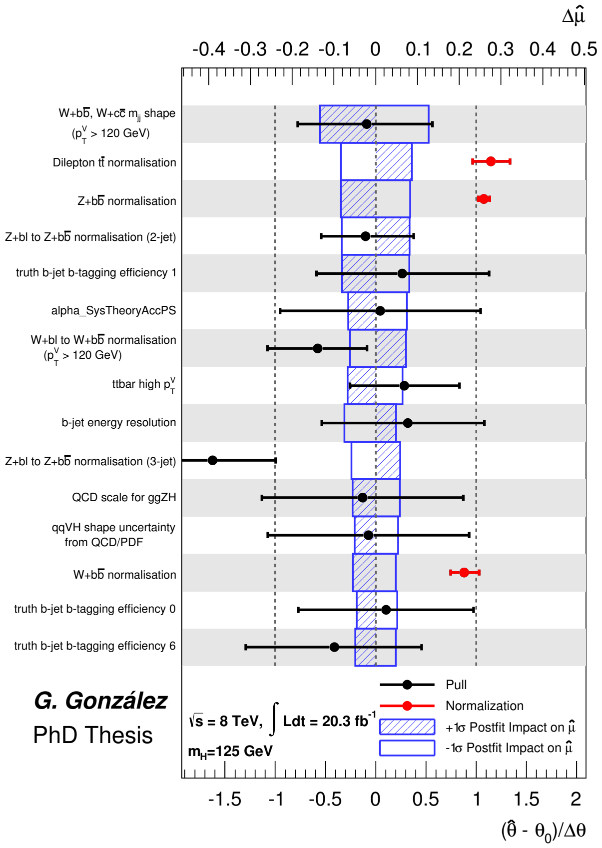


Figure 7.10: Impact of systematic uncertainties on the fitted signal strength parameter $\hat{\mu}$ applied to the 8 TeV data. The systematic uncertainties are listed in decreasing order of their impact on $\hat{\mu}$ on the y-axis. The boxes show the $\hat{\mu}$ variations, referring to the top x-axis, when fixing the corresponding individual NP θ to its post-fit value $\hat{\theta}$ modified upwards or downwards by its post-fit uncertainty, and repeating the fit as explained in the text. The hatched and open areas correspond to the upwards and downwards variations, respectively. The black filled circles, referring to the bottom x-axis, show the deviations of the fitted nuisance parameters $\hat{\theta}$ from their nominal values θ_0, expressed in terms of standard deviations with respect to their nominal uncertainties $\Delta\theta$. The associated error bars show the post-fit uncertainties of the nuisance parameters, relative to their nominal uncertainties. The red circles with their error bars, also referring to the bottom x-axis, show the fitted values and uncertainties of the normalization parameters that are freely floating in the fit. The normalization parameters have a pre-fit value of one. As explained in Section 6.9, the b-tagging uncertainties are decomposed into uncorrelated components; the labels 0, 1 and 6 refer to such components.

Chapter 8

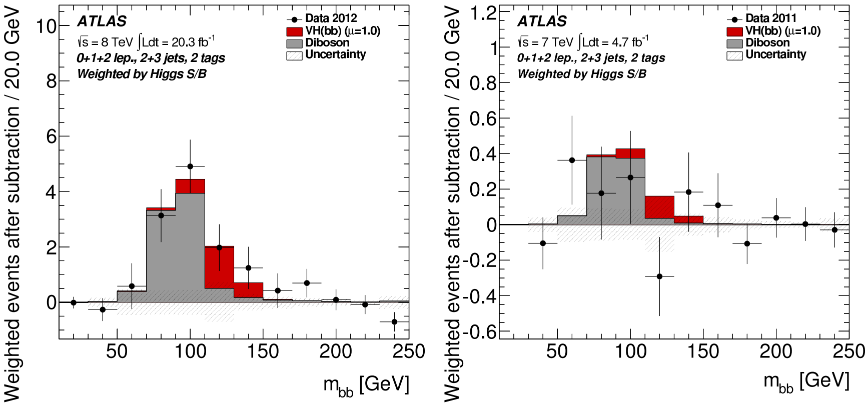
Results

This Chapter presents the results of the $VH(H \to b\bar{b})$ searches using the dijet mass and MVA analyses (see Appendix B). The complex fit procedure is first validated with the search of the diboson process VZ. As explained in Chapter 7, results are obtained from maximum-likelihood fits to data. The final discriminants use for extracting the results depend on the region, being different in the signal regions with 2-jets tagged as b-jets (2-tag) or in the control regions with only one b-jet tagged (1-tag). For the 2-tag regions the transformed m_{bb} and the multivariate algorithm output distributions are in the dijet mass and the MVA analyses, respectively. In the 1-tag regions the MV1c distributions are used in both analyses. Although this thesis is focused in the 8 TeV analyses, combined results after adding the 7 TeV dataset are also presented. Section 8.5 briefly discusses the 7 TeV analysis and results.

8.1 Diboson results

The diboson process where a Z boson decays into a pair of b-quarks in association with a W or another Z boson, has a very similar signature to that of the Higgs in the ZH channel. The main difference between the diboson and the Higgs events is that the first have a softer p_T^{bb} spectrum and that its m_{bb} distribution peaks around the Z mass. The diboson cross section is five times larger than the SM 125 GeV Higgs boson. This makes the search for the diboson production a perfect test to validate the analysis procedure and to check the fit performance.

The diboson fit, here referred as "VZ fit", uses a multiplicative scale factor μ_{VZ} to account for the normalization of its contribution with respect to the SM expectation. The small WW contribution is treated as a background and constrained within its uncertainty. A 125 GeV Higgs boson is also treated as background in this fit. Its cross section is set to the SM prediction value with an uncertainty of 50 %. Figure 8.1 shows the distribution of m_{bb} in data after subtraction of all backgrounds except for the diboson processes, as obtained with the Higgs dijet mass analysis for the 8 and 7 TeV datasets (see Section 8.5). The distribution associated with 8 TeV presents good agreement between number of data and expected events where the VZ process is clearly seen. Instead the 7 TeV shows a deficit of events in data compared to the background expectation, specially in the bin where the Higgs $m_H = 125\,\text{GeV}$ signal is expected.

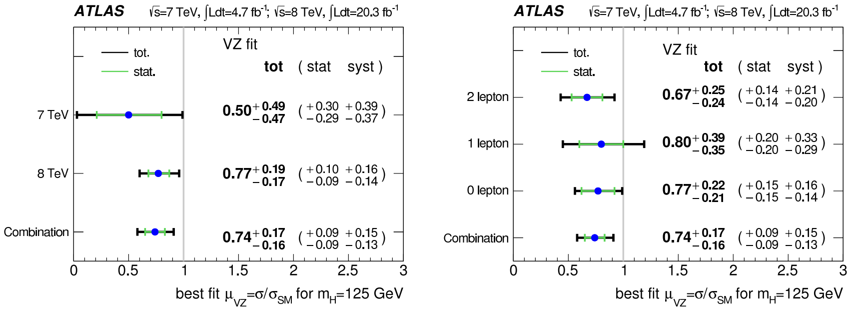


(a) (b)

Figure 8.1: The distribution of m_{bb} in data after subtraction of all backgrounds except for the diboson processes, as obtained with the Higgs dijet mass analysis for the (a) 8 TeV and (b) 7 TeV data. The contributions from all lepton channels, p_T^V intervals, number of jets and 2-tag b-tagging categories are summed weighted by their respective values of the ratio of expected Higgs boson signal to fitted background. The contribution of the associated WH and ZH production of a SM Higgs boson with $m_H = 125\,\text{GeV}$ is shown as expected for the SM cross section (indicated as $\mu = 1.0$). The size of the combined statistical and systematic uncertainty on the fitted background is indicated by the hatched band.

The diboson signal strength for the dijet mass analysis using the 8 TeV dataset is $\mu_{VZ} = 0.79 \pm 0.11(stat.) \pm 0.16(syst.)$, while the result using the MVA analysis in the same data is $\mu_{VZ} = 0.77 \pm 0.10(stat.) \pm 0.15(syst.)$. If only the 7 TeV dataset is used the signal strength obtain is $\mu_{VZ} = 0.50 \pm 0.30(stat.) \pm 0.38(syst.)$. It is important to note that the lower value of the 7 TeV compared with the 8 TeV, is the result of the deficit of events previously mentioned.

The MVA analysis discriminates better between processes leading to a very small diboson contribution in the most significant bins of its discriminant variable, making the analysis more sensitive compared with the dijet mass. Therefore, it is used to combine the result with the 7 TeV dataset giving a diboson signal strengths of $\mu_{VZ} = 0.74 \pm 0.09(stat.) \pm 0.14(syst.)$. The observed significance of the VZ signal is 4.9σ, to be compared to an expected significance of 6.3σ. Additional measurements where the signal strengths are floated independently for the 7 and 8 TeV datasets and for the three lepton channels are carried out as crosschecks. Figure 8.2 shows the results.



(a) (b)

Figure 8.2: The fitted values of the diboson signal strength μ_{VZ} for (a) the 7 TeV, 8 TeV and combined datasets, and (b) for the three lepton channels separately and combined, for the combined dataset. The MVA is used for the 8 TeV data. The individual μ_{VZ} values for the lepton channels are obtained from a simultaneous fit with the signal strength for each channel floating independently.

As a crosscheck as well, additional fits are performed with two signal strengths left freely floating for the VZ and Higgs processes in order to check any possible bias. The test uses the same final discriminants as for the Higgs boson analysis search on the 8 TeV dataset. The results of this test do not show any variation in the signal strength compare to the nominal procedure where the Higgs boson is normalized to its SM cross section value with an uncertainty of 50 %. Correlations between μ_{VZ} and μ_{VH}, obtained from the correlation matrices given by the fit, are 9 % in the dijet mass analysis and -3 % in the MVA. The reason for the low correlation between the signal strengths is the different shape of the p_{T}^{V} distributions for VZ and Higgs boson events. In addition and as previously mentioned, the MVA analysis discriminates better between processes.

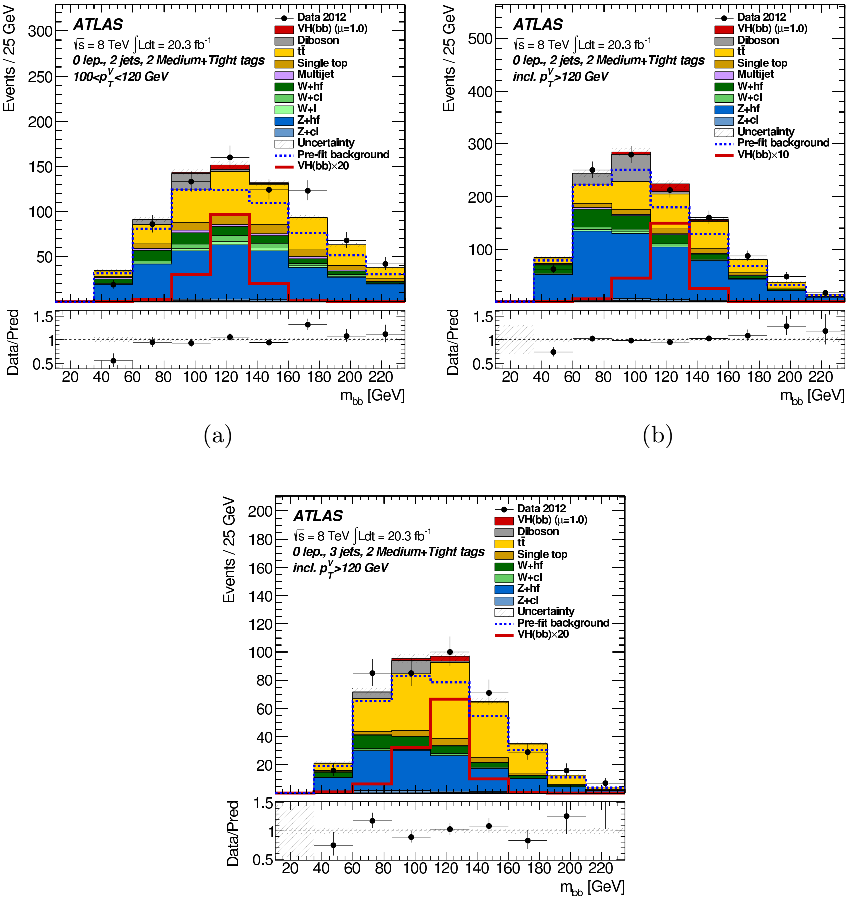
8.2 Dijet mass analysis results

Final results using the dijet mass analysis strategy on the 8 TeV data are presented in this Section. A full set of distributions for the 0-lepton m_{bb} and the MV1c discriminants, with the background normalizations and nuisance parameters adjusted from a fit to data, were already presented in Section 7.1. Figures 8.3 and 8.4 show a summary of m_{bb} distributions for the 0-, 1- and 2-lepton channels, where the b-jet categories with the highest purity, 2-Medium and 2-Tight, as well as, the p_{T}^{V} intervals with higher values are combined. In general, good agreement between data and expected number of events is presented.

The 95 % CL upper limits on the cross section times branching ratio for $pp \to (W/Z)(H \to b\bar{b})$ in the Higgs boson mass range 110–140 GeV is presented in Figure 8.5. The Figure shows how the analysis is most sensitive to the presence of a Higgs signal in the low mass range. The reason is that the branching ratio of the Higgs boson decaying into b-quarks decreases with its mass. The Figure also presents differences in all the mass range between the expected limit in the absence of a signal and the corresponding observed values. This indicates the presence of an excess of events above the expected background. In particular, the observed limit in the absence of signal obtained for $m_H = 125$ GeV is 2.2 times the SM value, to be compared to an expected limit of 1.02.

The probability p_0 associated with the background-only hypothesis at the Higgs boson mass of 125 GeV is 1.3 %. This corresponds to an excess observed at this mass with a significance of 2.2σ, to be compared to an expectation of 1.9σ. The p_0 values for the Higgs mass range 110–140 GeV obtained with the 8 TeV dataset is presented in Figure 8.6.

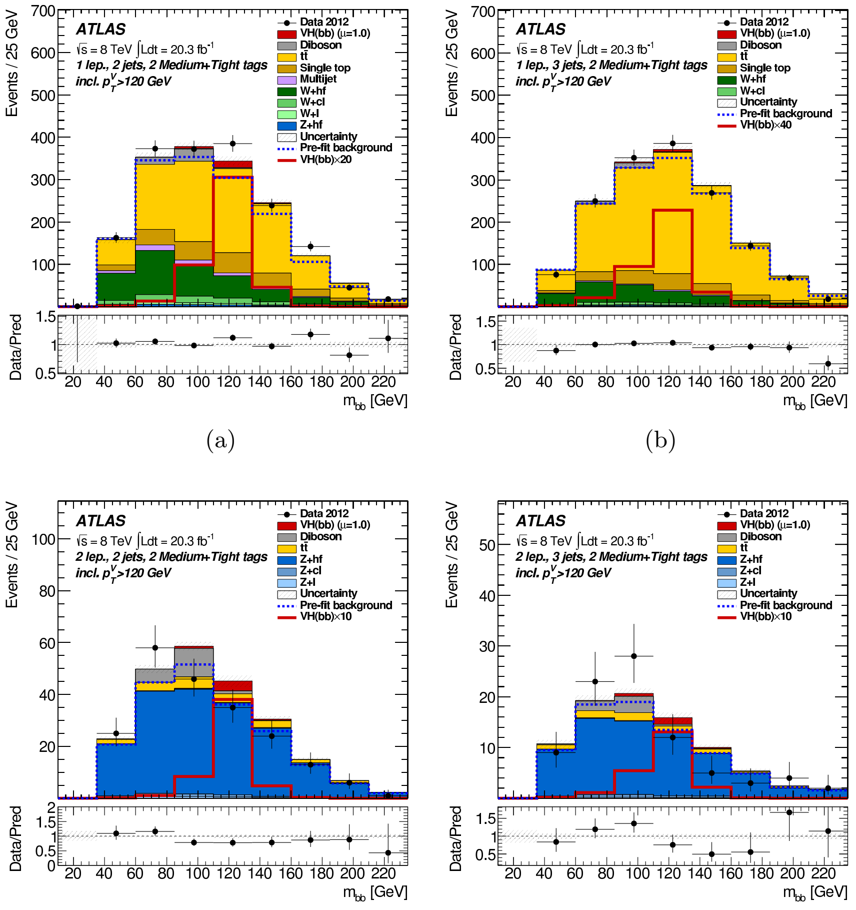
The fitted signal strength value for the dijet mass 8 TeV analysis for a 125 GeV Higgs boson mass is $\mu = 1.23 \pm 0.44(stat.) \pm 0.41(syst.)$. Additional measurements are also performed, where the signal strengths are floated independently for the WH and ZH production processes or the three lepton channels. Figures 8.7 and 8.8 show the fitted μ values for these additional measurements compared to the nominal dijet result. The compatibility between the signal strengths in the WH and ZH processes and the three lepton channels with the combined result are at the level of 8 % in both cases. This compatibility is calculated performing a χ^2 test, where fit results are compared. The high signal strength associated with the 1-lepton, around two times SM expected, is correlated with the large excess of events above the expected background observed in Figure 8.4a at $m_{bb} = 125$ GeV. Since most of the WH signal is obtained from the 1-lepton channel, similar values are expected between their associated signal strengths. Instead, the ZH signal is obtained from 0- and 2-leptons analyses. Non-trivial correlations among these analyses explain the differences between ZH and 0- and 2-lepton signal strengths which show a compatibility at the level of 15 %.



(a) (b)

(c)

Figure 8.3: The dijet-mass distribution observed in data (points with error bars) and expected (histograms) in the 0-lepton channel with the Medium and Tight b-tagging categories combined for (a) the 2-jet signal region in the $100 < p_T^V < 120$ GeV interval, (b) the 2-jet signal regions in the three intervals with $p_T^V > 120$ GeV combined, and (c) the 3-jet signal regions in the three intervals with $p_T^V > 120$ GeV combined. The background contributions after the global fit of the dijet-mass analysis are shown as filled histograms. The Higgs boson signal ($m_H = 125$ GeV) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The entries in overflow are included in the last bin. The size of the combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.



(a)

(b)

(c)

(d)

Figure 8.4: The dijet-mass distribution observed in data (points with error bars) and expected (histograms) with the Medium and Tight b-tagging categories combined and the three intervals with $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ combined for (a) the 2-jet signal regions of the 1-lepton channel, (b) the 3-jet signal regions of the 1-lepton channel, (c) the 2-jet signal regions of the 2-lepton channel, and (d) the 3-jet signal regions of the 2-lepton channel. The background contributions after the global fit of the dijet-mass analysis are shown as filled histograms. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The entries in overflow are included in the last bin. The size of the combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

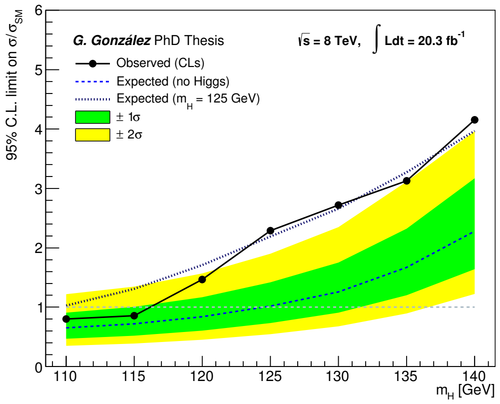


Figure 8.5: Observed (solid) and expected 95 % CL cross section upper limits, normalized to the SM Higgs boson production cross section, as a function of m_H for all channels using the 8 TeV dijet mass analysis dataset. The expected upper limit is given for the background-only hypothesis (dashed) and with the injection of a SM Higgs boson signal at a mass of 125 GeV (dotted). The dark and light shaded bands represent the 1σ and 2σ ranges of the expectation in the absence of a signal.

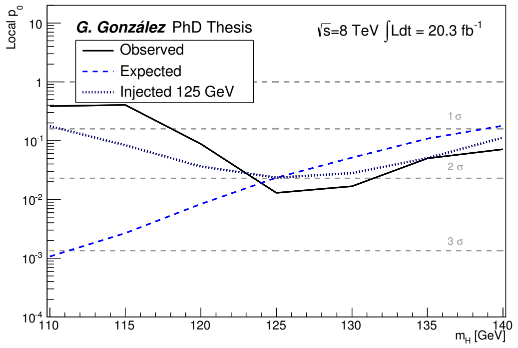


Figure 8.6: Observed (solid) and expected p_0 values as a function of m_H for all channels using the 8 TeV dijet mass analysis dataset. The expected p_0 values are given for the background-only hypothesis in the presence of a SM Higgs boson: for the dashed curve the Higgs boson mass corresponds to each tested mass point in turn; for the dotted curve the Higgs boson mass is 125 GeV.

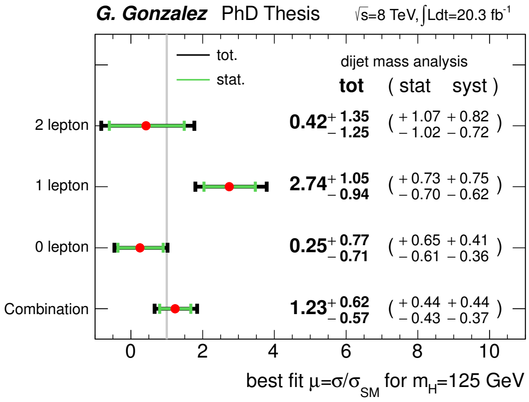


Figure 8.7: The fitted values of the Higgs boson signal strength parameter μ for $m_H = 125\,\text{GeV}$ for the 0-, 1- and 2-lepton channels and the combination of the three channels with the 8 TeV dataset. The individual μ values for the lepton channels are obtained from a simultaneous fit with the signal strength for each of the lepton channels floating independently.

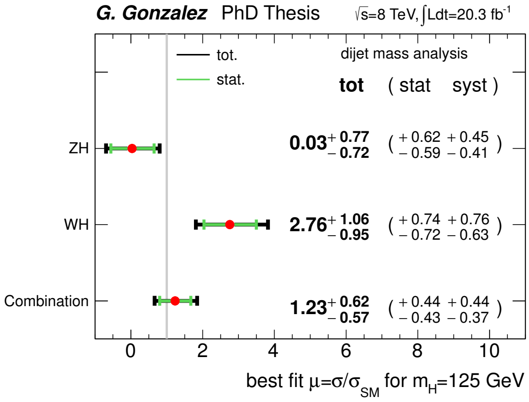
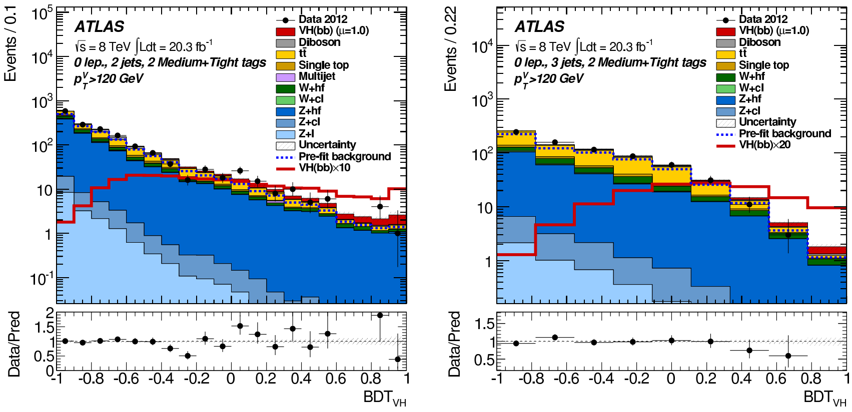


Figure 8.8: The fitted values of the Higgs boson signal strength parameter μ for $m_H = 125\,\text{GeV}$ for the WH and ZH processes and the combination of the WH and ZH processes with 8 TeV dataset. The individual μ values for the $(W/Z)H$ processes are obtained from a simultaneous fit with the signal strength for each of the WH and ZH processes floating independently.

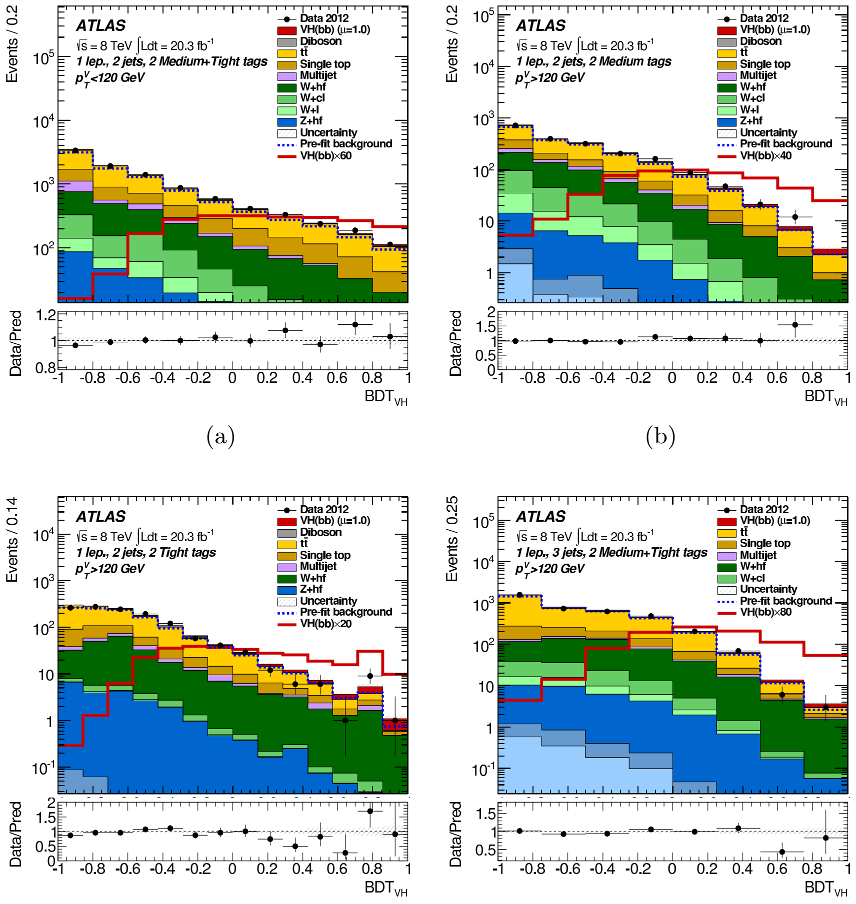
8.3 Multivariate analysis results

The multivariate analysis (see Appendix B) uses the Boosted Decision Tree algorithm output (BDT$_{VH}$) as the discriminant in the analysis regions with two b-jets (2-tag). For the 1-tag control regions, it uses the same MV1c distributions used in the dijet mass analysis. Some BDT$_{VH}$ distributions for the 0-, 1- and 2-lepton channels are presented in Figures 8.9, 8.10 and 8.11, respectively. The Figures show a combination of regions that include the b-jet categories with the highest purity (2-Medium and 2-Tight) or a combination of both; $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{V} < 120\,\mathrm{GeV}$, and events with 2- or 3-jets. Figure 8.9 shows a deficit of events in data compared to the background expectation in the most sensitive bins of the 0-lepton channel. The rest of the distributions present good agreement between number of data and background expected events.



(a) (b)

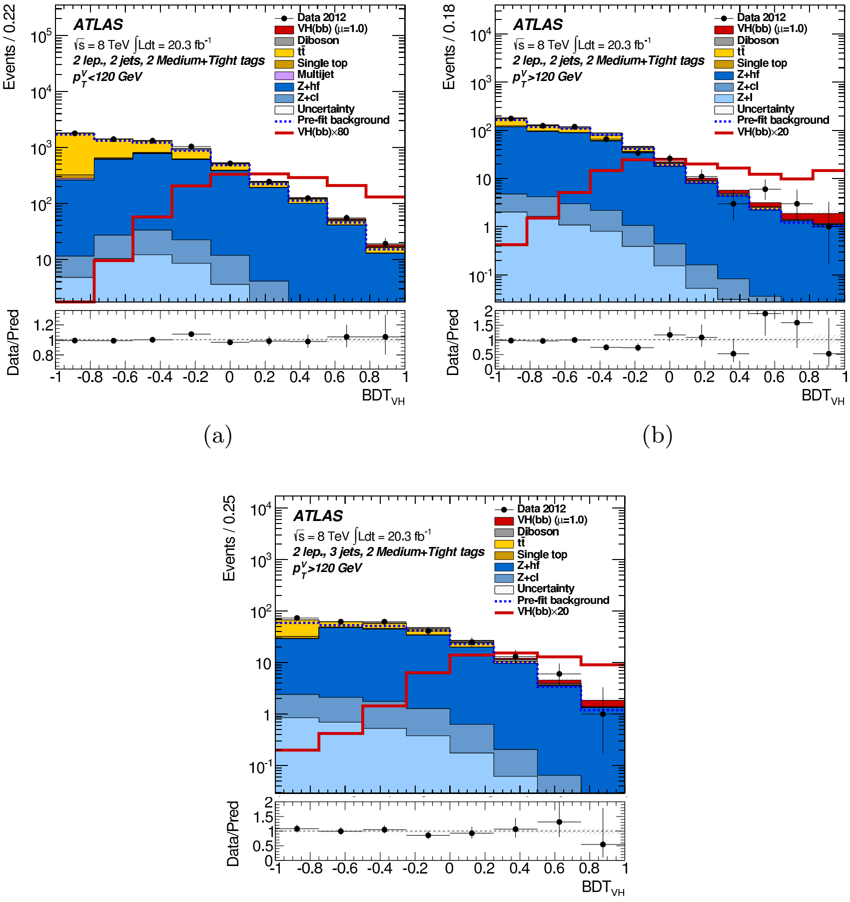
Figure 8.9: The BDT$_{VH}$-discriminant distribution observed in data (points with error bars) and expected (histograms) for the 0-lepton channel combining the 2-tag Medium and Tight b-tagging categories for $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$ for (a) 2-jet events and (b) 3-jet events. The background contributions after the global fit of the MVA are shown as filled histograms. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The size of the combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.



(a) (b)

(c) (d)

Figure 8.10: The BDT_{VH}-discriminant distribution observed in data (points with error bars) and expected (histograms) for the 2-tag signal regions of the 1-lepton channel for (a) 2-jet events with the Medium and Tight b-tagging categories combined and with $p_{\mathrm{T}}^{W} < 120\,\mathrm{GeV}$, (b) MM 2-jet events with $p_{\mathrm{T}}^{W} > 120\,\mathrm{GeV}$, (c) TT 2-jet events with $p_{\mathrm{T}}^{W} > 120\,\mathrm{GeV}$, and (d) MM and TT combined 3-jet events with $p_{\mathrm{T}}^{W} > 120\,\mathrm{GeV}$. The background contributions after the global fit of the MVA are shown as filled histograms. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The size of the combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.



(a)

(b)

(c)

Figure 8.11: The BDT$_{VH}$-discriminant distribution observed in data (points with error bars) and expected (histograms) for the 2-lepton channel combining the 2-tag Medium and Tight b-tagging categories for (a) 2-jet events with $p_{\mathrm{T}}^{Z} < 120\,\mathrm{GeV}$, (b) 2-jet events with $p_{\mathrm{T}}^{Z} > 120\,\mathrm{GeV}$, and (c) 3-jet events with $p_{\mathrm{T}}^{Z} > 120\,\mathrm{GeV}$. The background contributions after the global fit of the MVA are shown as filled histograms. The Higgs boson signal ($m_H = 125\,\mathrm{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The size of the combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

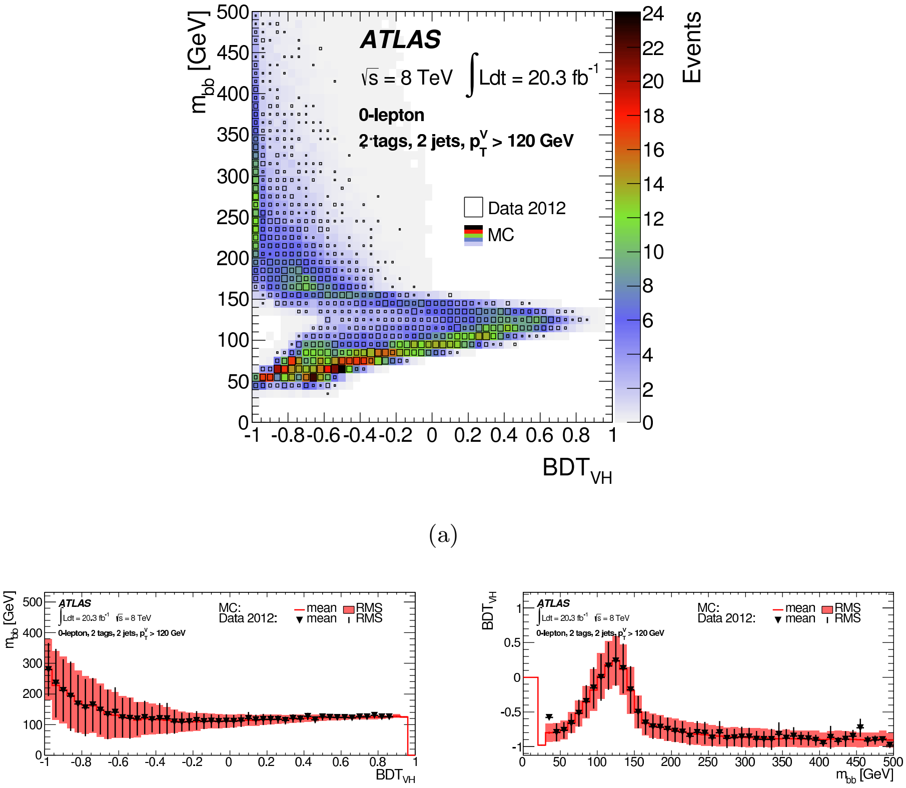
In the multivariate analysis, the observed (expected) 95 % CL upper limits on the cross section times branching ratio is 1.4 (0.8) times the SM expectation. As for the dijet mass analysis, differences between the values indicates an excess of events above the expected background. The observed (expected) significance of the data excess is 1.7σ (2.5σ) with a signal strength of $\mu = 0.65 \pm 0.32(stat.) \pm 0.26(syst.)$ for the Higgs mass $m_H = 125\,\text{GeV}$.

8.4 Correlation between dijet mass and multivariate-based results

The dijet mass and the multivariate analyses use the same data. Correlation studies between the two analyses are performed to check the consistency of the results. Figure 8.12 shows the correlation of the dijet mass versus the multivariate discriminant for a combination of categories in the 0-lepton channel. As expected, the higher values of the BDT_{VH} discriminant are obtained around the Higgs mass resonance at 125 GeV. The correlation between the signal strengths obtained from the two analyses is performed using pseudo-data with the method described in Ref.[103]. A large number of pseudo-data is generated using the simulated events with a signal strength equal to the SM value ($\mu = 1$). Pseudo-data samples are created with the number of events and associated statistical errors set such that they represent the integrated luminosity and the associated Poisson fluctuations in the data. Each of the generated samples is analyzed by both the dijet and the MVA analyses. This allows to compared the $\hat{\mu}$ obtained and to calculate their statistical correlation. The dijet mass and multivariate results are expected to be 67 % correlated, and the observed results are found to be statistically consistent at the level of 8 %.

8.5 The 7 TeV analysis results

For the 7 TeV dataset, only a dijet mass analysis is performed. This dijet analysis is similar but not identical to the one performed for the 8 TeV dataset. The main differences arise from the object identification, reconstruction and calibration, some of them performed independently for the 7 TeV and 8 TeV datasets. As an example, in the 7 TeV analysis the b-tagging calibration is performed for a single operating point. Another difference is that the 100-120 GeV p_{T}^{V} region is not used in the 0-lepton channel. A detailed description of the analysis can be found in Ref. [104]. Figure 8.13 shows the m_{bb} distributions for the most sensitive regions in the 0-, 1- and 2-lepton channels. These correspond to regions with 2-jets both b-tag and $p_{\mathrm{T}}^{V} > 120\,\text{GeV}$. The Figure shows a deficit of events in the m_{bb} region where the Higgs signal is expected in the case of 0- and 1-lepton channels.

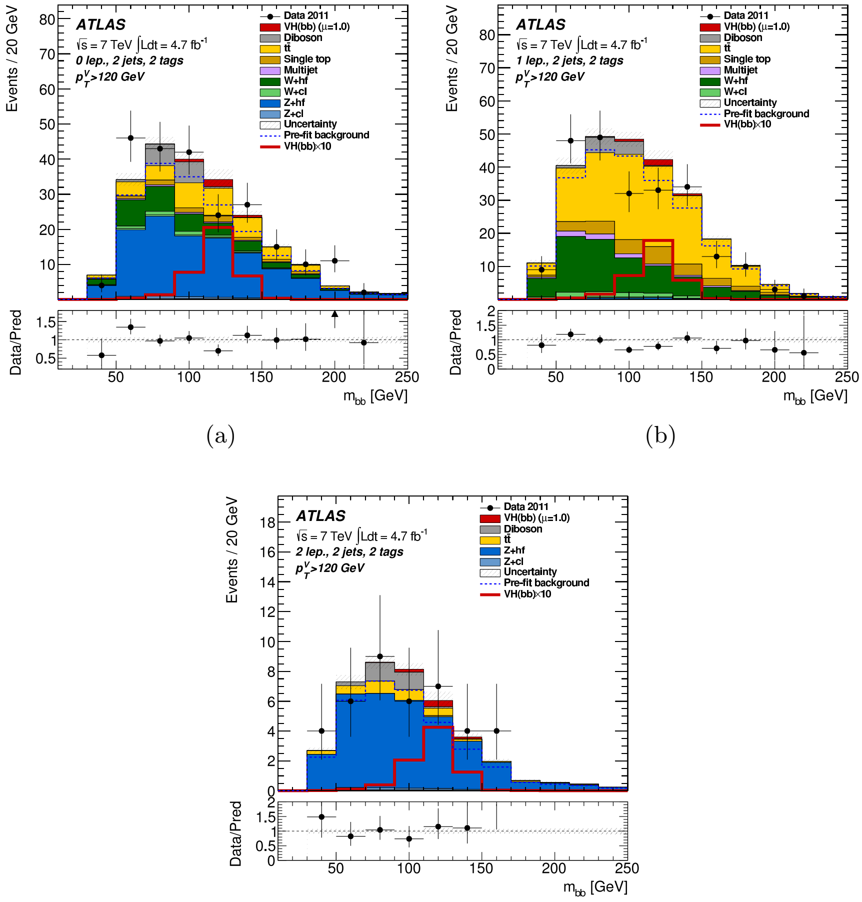


(a)

(b)

(c)

Figure 8.12: Correlations between the dijet mass and the multivariate discriminant (BDT_{VH}) in the 2-jet 2-tag category (LL, MM and TT combined) of the 0-lepton channel for $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$: (a) the dijet mass versus the BDT_{VH} discriminant for the total expected background (shadings indicating the numbers of events) and the data (open boxes with the box size being proportional to the number of events), (b) and (c) show the mean values and RMS of the projections onto the axes of the BDT_{VH} discriminant and dijet mass, respectively, for the total expected background after the global fit of the MVA and the data.



(a)

(b)

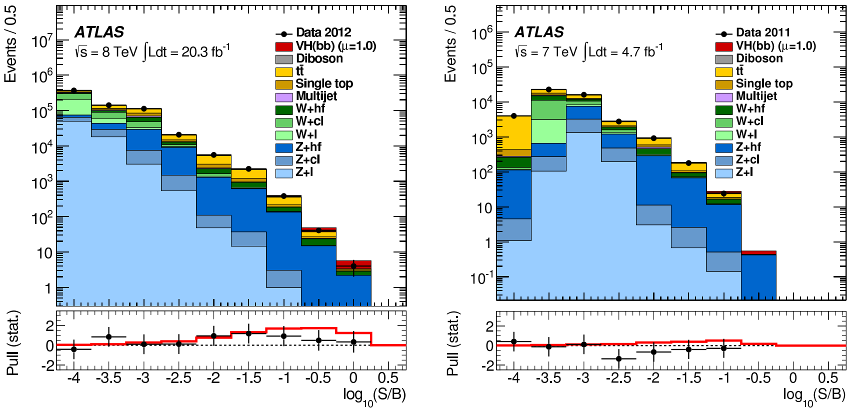
(c)

Figure 8.13: The dijet-mass distributions observed in the 7 TeV data (points with error bars) and expected (histograms) for the 2-jet 2-tag signal regions with the $p_T^V > 120\,\text{GeV}$ intervals combined: (a) 0-lepton channel, (b) 1-lepton channel, and (c) 2-lepton channel. The background contributions after the global fit of the dijet-mass analysis are shown as filled histograms. The Higgs boson signal ($m_H = 125\,\text{GeV}$) is shown as a filled histogram on top of the fitted backgrounds, as expected from the SM (indicated as $\mu = 1.0$), and, unstacked as an unfilled histogram, scaled by the factor indicated in the legend. The dashed histogram shows the total background as expected from the pre-fit MC simulation. The entries in overflow are included in the last bin. The size of the combined statistical and systematic uncertainty on the sum of the signal and fitted background is indicated by the hatched band. The ratio of the data to the sum of the signal and fitted background is shown in the lower panel.

8.6 Combination of 7 TeV and 8 TeV results

The 7 TeV and 8 TeV analyses are considered independent given the differences in the physics objects, event categorization and the estimation of the systematic uncertainties. To provide combine results, a global fit, where the distributions are not aggregate and the NPs are treated as uncorrelated, is performed. The approved ATLAS results for the Higgs searches in the $VH(H \to b\bar{b})$ decay mode are obtained from the combination of the 8 TeV MVA and the 7 TeV dijet mass datasets. The MVA is chosen since it has a better expected sensitivity than the dijet mass analysis (2.5σ to be compared with the 1.9σ given by the dijet mass).

Figure 8.14 shows the data, background and signal yields, where the final discriminant bins in all signal regions are combined into bins of $\log(S/B)$, separately for the 7 TeV and 8 TeV datasets. As previously mentioned, it is observed a deficit of events in the 7 TeV most sensitive bins, while the 8 TeV presents a small excess of events above the expected background.



(a) (b)

Figure 8.14: Event yields as a function of $\log(S/B)$ for data, background and Higgs boson signal with $m_H = 125$ GeV for the (a) 8 TeV data and (b) 7 TeV data. Final discriminant bins in all signal regions are combined into bins of $\log(S/B)$. The signal S and background B yields are expected and fitted, respectively. The Higgs boson signal contribution is shown as expected for the SM cross section (indicated as $\mu = 1.0$). The pull of the data with respect to the background-only prediction is also shown with statistical uncertainties only. The full line indicates the pull of the prediction for signal ($\mu = 1.0$) and background with respect to the background-only prediction.

The signal strengths results for $m_H = 125$ GeV using 7 TeV and 8 TeV data alone, and their combination are shown in Figure 8.15. The signal strength using all lepton channels and datasets is [1]

$$\mu = 0.51 \pm 0.31(stat.) \pm 0.24(syst.). \tag{8.1}$$

For the 8 TeV dataset alone is $\mu = 0.65 \pm 0.32(stat.) \pm 0.26(syst.)$, while for the 7 TeV data alone is $\mu = -1.6 \pm 1.2(stat.) \pm 0.9(syst.)$. The 8 TeV analysis presents higher statistics and cross section of the signal processes, therefore it drives the results when the combination is performed.

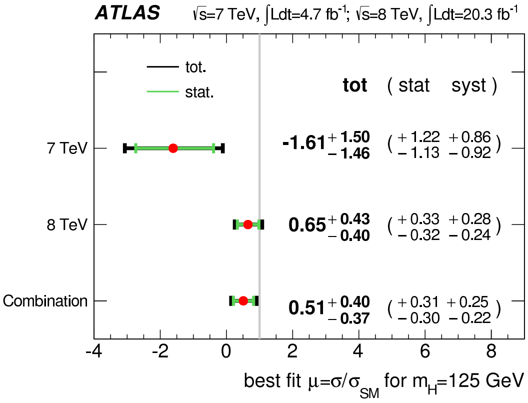


Figure 8.15: The fitted values of the Higgs boson signal strength parameter μ for $m_H = 125$ GeV for the 7 TeV and 8 TeV datasets independently and combined.

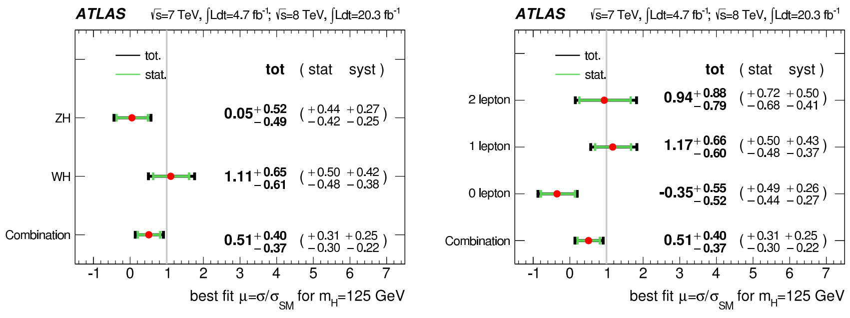
Additional fits where the signal strengths are floated independently for the WH and ZH production processes and for the three lepton channels are shown in Figure 8.16 (a) and (b) respectively. The consistency in the results shown by the WH and ZH processes and their combination is at a 20 % level. In the case of the three lepton signal strengths, the consistency is at the 72 % for the 7 TeV data and 8 % for the 8 TeV data. The low values shown by the 0-lepton μ is also affected by an additional data deficit observed in the most sensitive bins of the BDT discriminant as shown in Figure 8.9. This low 0-lepton μ drives also the low ZH value.

The 95 % CL upper limits on the cross sections times branching ratio for the Higgs boson mass range 110–140 GeV is presented in Figure 8.17. The observed limit for $m_H = 125$ GeV is 1.2 times the SM value, to be compared with an expected limit, in the absence of signal, of 0.8.

The Figure 8.18 shows the p_0 values in the mass range 110-140 GeV obtained from the combination of the 7 TeV and 8 TeV datasets. The result

---

[1] Using the analysis results from the Higgs decay modes with the highest sensitivity and energy resolution ($H \to \gamma\gamma$ and $H \to 4l$), the ATLAS Collaboration obtained a value for the Higgs mass of $m_H = 125.36$ GeV. The signal strength obtained by fitting this mass value is $\mu = 0.52 \pm 0.32(stat.) \pm 0.24(syst.)$.



(a) (b)

Figure 8.16: The fitted values of the Higgs boson signal strength parameter μ for $m_H = 125\,\text{GeV}$ using the 7 and 8 TeV datasets: (a) for the WH and ZH processes and (b) for the three channels. The result of the final combination is shown in each case. The individual signal strengths for the lepton channels are obtained from a simultaneous (for the $(W/Z)H$ processes) are obtained from a simultaneous fit with the signal strength for each of the lepton channels (for each of the WH and ZH processes) floating independently.

at $m_H = 125\,\text{GeV}$ corresponds to an observed excess with a significance of 1.4σ, to be compared to an expectation of 2.6σ.

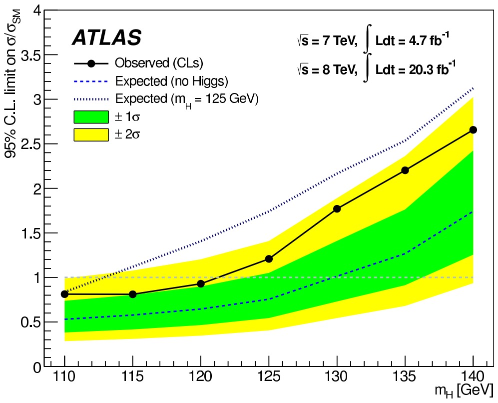


Figure 8.17: Observed (solid) and expected 95 % CL cross section upper limits, normalized to the SM Higgs boson production cross section, as a function of m_H for all channels and data-taking periods combined, as obtained using the dijet-mass analysis for the 7 TeV dataset and BDTs trained at each individual mass for the 8 TeV dataset. The expected upper limit is given for the background-only hypothesis (dashed) and with the injection of a SM Higgs boson signal at a mass of 125 GeV (dotted). The dark and light shaded bands represent the 1σ and 2σ ranges of the expectation in the absence of a signal.

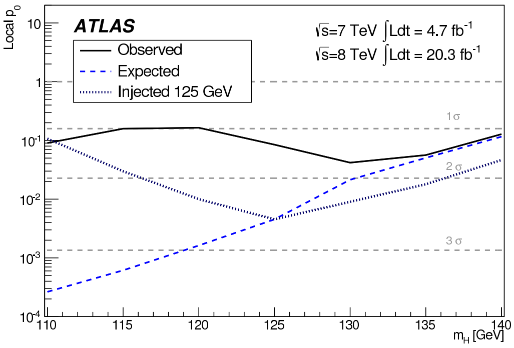


Figure 8.18: Observed (solid) and expected p_0 values as a function of m_H for all channels and data-taking periods combined, as obtained using the dijet-mass analysis for the 7 TeV dataset and BDTs trained at each individual mass for the 8 TeV dataset. The expected p_0 values are given for the background-only hypothesis in the presence of a SM Higgs boson: for the dashed curve the Higgs boson mass corresponds to each tested mass point in turn; for the dotted curve the Higgs boson mass is 125 GeV.

Conclusions

This thesis presents the latest ATLAS results on the search for the Standard Model Higgs boson decaying into a pair of b-quarks. The Higgs boson is searched for in the production channel for which the Higgs is produced in association with a W or Z boson in the final state. This includes the leptonic (electron and muon) decay channels for the vector bosons in addition to the channel with Z decaying into a pair of neutrinos. In particular, this document is mostly focused on the $ZH(Z \rightarrow \nu\nu$ and $H \rightarrow b\bar{b})$ channel. The analysis has considered all the proton-proton collision data collected by the ATLAS experiment at $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$.

The analysis is based on a complex characterization of the events according to the vector boson decay, the number of jets and b-jets in the final state, and the vector boson transverse momentum. This permits a detail study of the main background processes. Two different approaches are considered for the analysis: using the m_{bb} distribution as main discriminating variable or using a neural-net multivariate technique.

The analysis involves a complex fitting strategy of data and background predictions in control and signal regions that contributes to constrain the background processes, to reduce the systematic uncertainties, and to increase the sensitivity to the Higgs signal. The whole procedure is validated by searching for the known diboson $(W/Z)Z$ $(Z \rightarrow b\bar{b})$ process with similar final state. The diboson signal is clearly observed with a significance of $4.9\,\sigma$. Its observed strength (compared to that of the SM) is $\mu_{VZ} = 0.74 \pm 0.09(stat.) \pm 0.14(syst.)$, in agreement with the SM prediction.

The measured output distributions of the multivariate algorithm in each event category are compared with the expected signal and background distributions and an excess of events above the expected background is observed at a Higgs mass m_H of 125 GeV. The observed (expected) significance of this excess is 1.4σ (2.6σ) and the ratio of the measured signal yield to the Standard Model expectation is found to be $\mu = 0.51 \pm 0.31(stat) \pm 0.24(syst)$. These results are consistent among channels and in agreement with those obtained with the analysis that uses m_{bb} distributions as the main discriminating variable. Although an excess is observed, its statistical significance is not large enough to claim neither evidence nor discovery of the process.

The LHC will resume operations in 2015 with an increased centre-of-mass energy of 13 TeV. With the new data, the study of the Higgs couplings to quarks will continue as one of the pillars of the physics program in which the $(WZ)H$ analysis promises to keep playing a central role.

Appendix A

VH analysis in the production channels $WH \rightarrow l\nu b\bar{b}$ and $ZH \rightarrow llb\bar{b}$

This Appendix presents the main characteristics of the channel analyses that target the Higgs signatures $WH \rightarrow l\nu b\bar{b}$ and $ZH \rightarrow llb\bar{b}$. These are not covered in Chapter 6 which is mainly focused on the $ZH \rightarrow \nu\nu b\bar{b}$. As a reminder, due to the number of expected leptons (electrons or muons) in their final states the analysis channels are referred to as 0-lepton, 1-lepton and 2-lepton channel. The main difference between channels is that the 0-lepton veto events with leptons, while the 1- and 2-lepton select them and categorized them into electron and muon sub-channels. Common to the three channels is the selection performed to target the Higgs, but the presence of leptons in the events motivates the following:

- The use of lepton triggers.
- A kinematic selection to constrain the mass of the reconstructed W or Z boson.
- The full reconstruction of events in the 2-lepton channel, allowing to correct the kinematics in the event to balance it, improving the m_{bb} resolution.

Multijet background contributions are estimated using data-driven techniques and different methods are carried out by the lepton channels.

A.1 Object and kinematic selection

As a reminder, the VH analysis uses three inclusive lepton definition that in increasing order of purity are: *loose*, *medium* and *tight*. Loose leptons are described in detail in Section 6.2.1. Medium electrons are defined with the loose requirements and $E_{\mathrm{T}} > 25\,\mathrm{GeV}$, while medium muons correspond to type (1) muons with $E_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| < 2.5$. In addition, tight electrons and muons are selected with track and calorimeter isolations of $iso_{track} < 0.04$ and $iso_{calo} < 0.04$. In the case of electrons, a veryTight likelihood is also required (see Section 5.2). Table A.1 summarizes all lepton requirements.

Flavor	Electrons	Muons		
Type		(1)	(2)	(3)
Loose				
quality	Loose LH			
$\lvert\eta\rvert$	< 2.5	< 2.7	< 0.1	in [2.5, 2.7]
E_{T} (GeV)	> 7	> 7	> 20	> 7
$\lvert d_0\rvert$ (mm)	–	< 0.1	< 0.1	–
$\lvert z_0\rvert$ (mm)	–	< 10	< 10	–
iso_{track}	< 0.1	< 0.1	< 0.1	–
Medium				
E_{T} (GeV)	> 25	> 25	NU	
$\lvert\eta\rvert$	–	< 2.5	NU	
Tight				
quality	VeryTight LH	–		
iso_{track}	< 0.04	< 0.04	NU	
iso_{calo}	< 0.04	< 0.04	NU	

Table A.1: Loose medium and tight lepton requirements. NU stands for "Not Used" and LH for Likelihood (see Section 5.2).

Events containing one tight lepton and no additional loose leptons are assigned to the 1-lepton channel. Events containing one medium lepton and one additional loose lepton of the same flavor, and no other loose leptons, are assigned to the 2-lepton channel.

Table A.2 presents the list of event selection criteria. The events in the 1- and 2-lepton channels are categorized in five p_{T}^{V} intervals , with boundaries at 0, 90, 120, 160, and 200 GeV. In addition, to the common selection of ΔR between the two leading jets, a selection on the mass of the reconstructed W or Z boson is applied. In the the 1-lepton channel, a requirement is imposed on the transverse mass m_{T}^{W}. This requirement reduces the contamination from the $t\bar{t}$ background. To reduce the multijet background, additional cuts on the effective mass M_{eff}, which is the scalar sum p_{T} of the signal jets, leptons, and $E_{\mathrm{T}}^{\mathrm{miss}}$ are applied. In the 2-lepton channel, criteria are imposed on the dilepton invariant mass m_{ll}, which must be consistent with the mass of the Z boson. A requirement is also imposed on $E_{\mathrm{T}}^{\mathrm{miss}}$.

Variable	Dijet Mass Analysis				
$p_{\mathrm{T}}^{j_1}$(GeV)	> 45				
p_{T}^{V} (GeV)	0-90	90-120	120-160	160-200	> 200
$\Delta R(jet_1, jet_2)$	0.7-3.4	0.7-3.0	0.7-2.3	0.7-1.8	< 1.4
1-lepton selection					
M_{eff} (GeV)	> 180		–		
$E_{\mathrm{T}}^{\mathrm{miss}}$ (GeV)	–		> 20		> 50
m_{T}^{W} (GeV)	< 120				
2-lepton selection					
m_{ll} (GeV)	83-99				
$E_{\mathrm{T}}^{\mathrm{miss}}$ (GeV)	< 60				

Table A.2: Event kinematic selection applied in the 1- and 2-lepton analyses.

A.2 Kinematic fit in 2-lepton

The $ZH \to ll b\bar{b}$ has the beneficial feature that the expected signal event topology is fully reconstructed and balanced in the transverse plane, with the possible exception of $E_{\mathrm{T}}^{\mathrm{miss}}$ coming from semileptonic heavy flavor decays. Therefore, one can adjust the energies of the reconstructed particles in order to force the event into a balanced configuration within the event and object resolutions. In particular, the energy calibration of the jets is further improved by a kinematic likelihood fit. This fit includes constrains on the dilepton mass, the transverse components of the $llb\bar{b}$ system momentum, a prior built from the expected true jet p_{T} spectrum in ZH events and dedicated transfer functions to relate the reconstructed jet transverse momenta to their true value. Figure A.1 shows the m_{bb} before and after the kinematic likelihood fit performed in simulation for a signal region of the analysis. Overall, the m_{bb} resolution is improved by 30 % in the 2-lepton channel.

A.3 Multijet estimation in the $WH \to l\nu b\bar{b}$ channel

Multijet events mimicking the $WH \to l\nu b\bar{b}$ signal topology is a major background of the 1-lepton analysis. As with most backgrounds of this type, the efficiency is small but the production rate is so large that the background cannot be ignored. As in the 0-lepton case, the multijet background is estimated using a data-driven technique.

The method uses data events to estimate the contribution of multijet events which satisfy the 1-lepton selection. The data sample dominated by multijet events is obtained by modifying the nominal selection

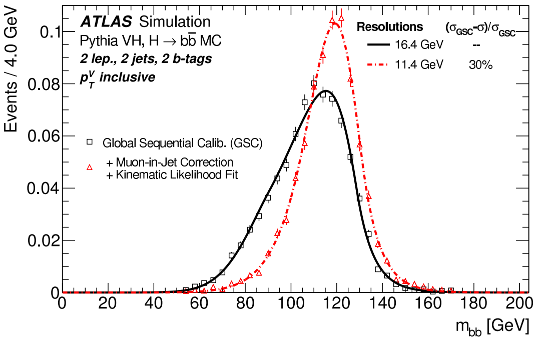


Figure A.1: Dijet-invariant-mass distribution for the decay products of a Higgs boson with $m_H = 125\,\text{GeV}$ in the 2-lepton. The distributions are shown using jets after global sequential calibration (GSC, solid), and after adding muons inside jets and applying the kinematic fit (dash-dotted) (see Section 5.6.4).

to use medium, instead of tight, leptons and loosening both the track and calorimeter-based isolation criteria. Multijet distributions are then defined applying the selection cuts on these data events and subtracting the electroweak contribution. Although the distributions describe the shape of the multijet background, they must be normalized to take into account the difference in efficiency between selections. The normalization of the multijet distributions is obtained from a "multijet fit" to the $E_{\mathrm{T}}^{\mathrm{miss}}$ distributions in the 2- and 3-jet, 1- and 2-tag (LL, MM and TT combined) categories.

A.4 Multijet estimation in the $ZH \to llb\bar{b}$ channel

Multijet events are background to the 2-lepton channel for cases in which the two jets are mistakenly identified as leptons. Photon conversions (detected as electrons) also contribute in the case of electrons. In addition to fake leptons, true leptons from the semi-leptonic decay of heavy flavor hadrons may also contribute. The multijet background is estimated using a data-driven technique. As for the 1-lepton, a data sample dominated by multijet events is obtained by loosening the identification and isolation requirements. The analysis selections are applied to the sample and the resulting data distributions are used to describe the shape of the multijet background in the various analysis regions.

As previously explained, the distributions must be normalized to take into account the difference in efficiency between the two electron selections. The normalization is estimated by fitting the di-electron invariant mass m_{ee} distribution after applying the nominal selection. The fit is performed over the range $40 < m_{ee} < 150\,\mathrm{GeV}$, where the Z+jets and multijet components are free parameters, while the other backgrounds (mostly $t\bar{t}$) are taken from the simulation.

The multijet background in the muon channel is investigated by comparing the data and MC in the $m_{\mu\mu}$ side-bands and it is found to be negligible.

Appendix B

Multivariate Analysis Techniques

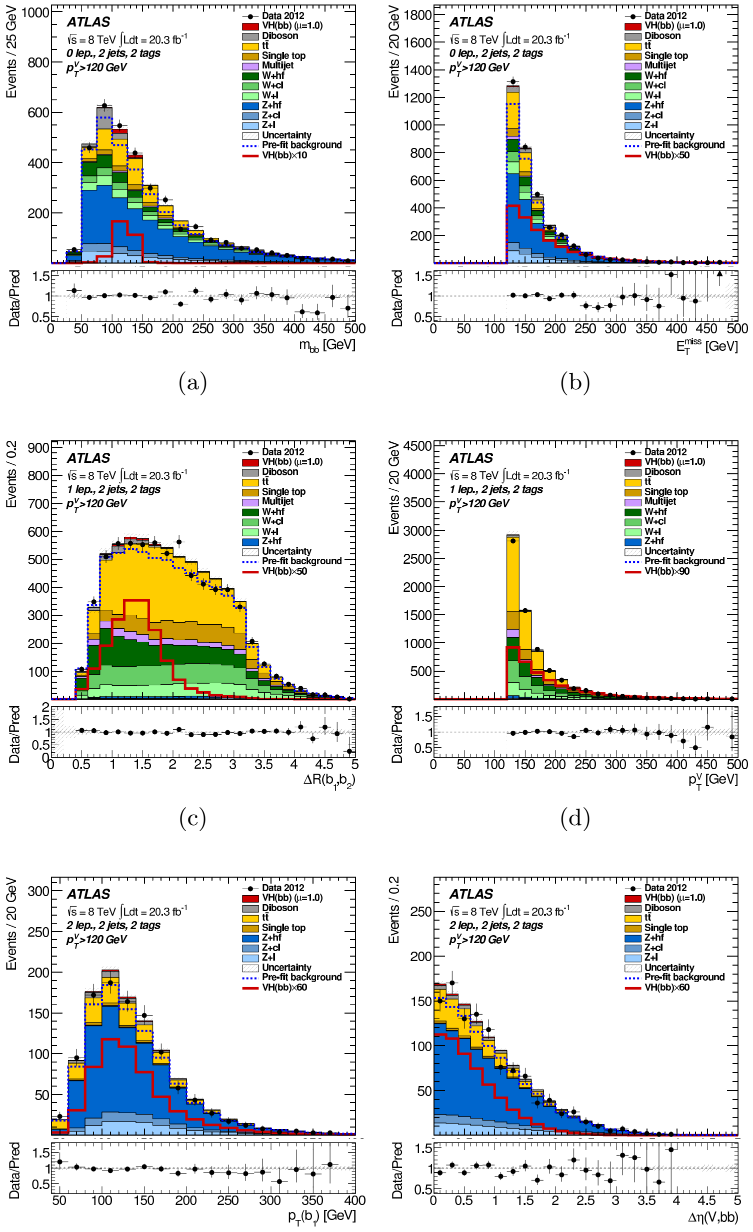
This Appendix describes the general concepts of the multivariate analysis (MVA) techniques and their applications in the VH search. The MVA refer to any data analysis that uses more than one statistical input variable at a time. In particle physics the MVA approach relies on several variables used as input to determine a probability for each event to be originating from a signal or background process. This approach differs from the dijet analysis presented in previous Sections where consecutively cuts are apply on discriminating variables. Boosted Decision Tree (BDT) [105, 106] is the multivariate technique used among all the Toolkit for Multivariate Data Analysis (TMVA) [107].

The BDT algorithm needs to be trained to distinguish signal from background. The selection of variables used in this training can be found in the Table B.1. Variables from the 0-lepton that could need further explanations are: V, used to represent the vector boson which in the 0-lepton analysis is reconstructed from the $E_{\mathrm{T}}^{\mathrm{miss}}$, H, which correspond to the Higgs candidate coming from vectorial sum of the two leading p_{T} jets and M_{eff}, which is the scalar sum p_{T} of the signal jets, leptons, and $E_{\mathrm{T}}^{\mathrm{miss}}$ known as effective mass. Notice that in the 0-lepton analysis the p_{T}^{V} and the $E_{\mathrm{T}}^{\mathrm{miss}}$ correspond to the same variable and $\Delta\phi(V, H)$ correspond to $\Delta\phi(E_{\mathrm{T}}^{\mathrm{miss}}, jj)$. Figure B.1 shows an example of the distributions of signal and background input variables used to train the BDT.

In this analysis the BDT algorithm output ranges from -1 to 1. Background events are expected to give values close -1 while signal events close to 1. Figure B.2 shows an example of a 1-lepton BDT output distribution.

Variable	0-Lepton	1-Lepton	2-Lepton
p_{T}^{V}		✓	✓
$E_{\mathrm{T}}^{\mathrm{miss}}$	✓	✓	✓
$p_{\mathrm{T}}^{jet_1}$	✓	✓	✓
$p_{\mathrm{T}}^{jet_2}$	✓	✓	✓
$MV1c(jet_1)$	✓	✓	✓
$MV1c(jet_2)$	✓	✓	✓
m_{jj}	✓	✓	✓
$\Delta R(jet_1, jet_2)$	✓	✓	✓
$\|\Delta\eta(jet_1, jet_2)\|$	✓		✓
$\|\Delta\phi(V, H)\|$	✓	✓	✓
$\|\Delta\eta(V, H)\|$			✓
$M_{eff}(M_{eff\ 3})$	✓		
$min\|\Delta\phi(\ell, jet)\|$		✓	
m_{T}^{W}		✓	
$m_{\ell\ell}$			✓
	Only in 3 Jet Events		
$p_{\mathrm{T}}^{jet_3}$	✓	✓	✓
m_{jjj}	✓	✓	✓

Table B.1: Input variables used in the MVA analysis in the different lepton channels. Some distributions plots can be found Figure B.1.



(a) (b)

(c) (d)

(e) (f)

Figure B.1: Examples of variables input to the MVA in the 2-jet 2-tag category (LL, MM and TT combined) for $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$: (a) 0-lepton channel, dijet mass; (b) 0-lepton channel, $E_{\mathrm{T}}^{\mathrm{miss}}$; (c) 1-lepton channel, $\Delta R(b_1, b_2)$; (d) 1-lepton channel, p_{T}^{W}; (e) 2-lepton channel, $p_{\mathrm{T}}^{b_1}$; (f) 2-lepton channel, $\Delta\eta(V, H)$.

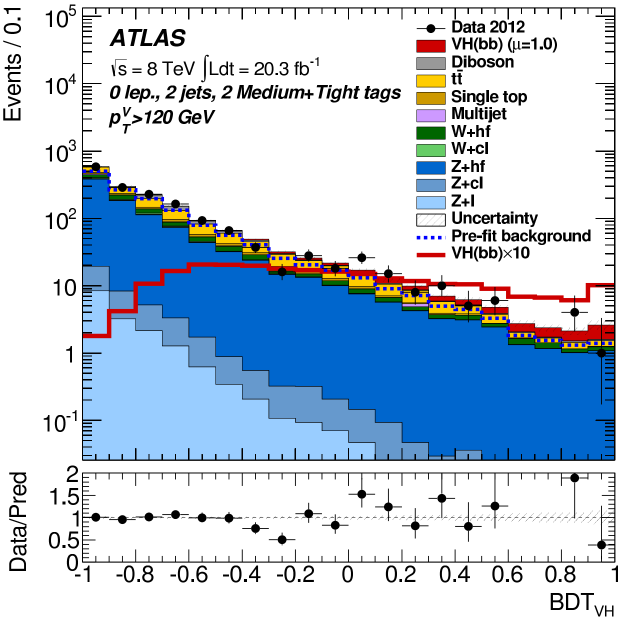


Figure B.2: The BDT-output distribution for the expected background and signal contributions in the 0-lepton channel and the 2-jet 2-tag MM and TT category for $p_{\mathrm{T}}^{V} > 120\,\mathrm{GeV}$.

Bibliography

[1] M. E. Peskin and D. V. Schroeder, *An introduction to quantum field theory*Advanced book program (Westview Press Reading (Mass.), Boulder (Colo.), 1995), Autre tirage : 1997.

[2] F. Halzen and A. Martin, *Quarks & Leptons: An introductory course in modern particle physics* (John Wiley & Sons, New York, USA, 1984).

[3] Particle Data Group, J. Beringer *et al.*, Phys.Rev. **D86**, 010001 (2012).

[4] M. Han and Y. Nambu, Phys.Rev. **139**, B1006 (1965).

[5] M. Gell-Mann, Acta Phys.Austriaca Suppl. **9**, 733 (1972).

[6] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, Phys.Lett. **B47**, 365 (1973).

[7] S. Bethke, Prog.Part.Nucl.Phys. **58**, 351 (2007), hep-ex/0606035.

[8] UA1 Collaboration, G. Arnison *et al.*, Phys.Lett. **B122**, 103 (1983).

[9] UA2 Collaboration, M. Banner *et al.*, Phys.Lett. **B122**, 476 (1983).

[10] UA1 Collaboration, G. Arnison *et al.*, Phys.Lett. **B126**, 398 (1983).

[11] UA2 Collaboration, P. Bagnaia *et al.*, Phys.Lett. **B129**, 130 (1983).

[12] F. Englert and R. Brout, Phys.Rev.Lett. **13**, 321 (1964).

[13] P. W. Higgs, Phys.Lett. **12**, 132 (1964).

[14] P. W. Higgs, Phys.Rev.Lett. **13**, 508 (1964).

[15] A. Djouadi, Phys.Rept. **457**, 1 (2008), hep-ph/0503172.

[16] ATLAS Collaboration, G. Aad *et al.*, Phys.Lett. **B716**, 1 (2012), 1207.7214.

[17] CMS Collaboration, S. Chatrchyan *et al.*, Phys.Lett. **B716**, 30 (2012), 1207.7235.

[18] ATLAS Collaboration, G. Aad *et al.*, Phys.Rev. **D90**, 052004 (2014), 1406.3827.

[19] CERN Report No. ATLAS-CONF-2014-061, 2014 (unpublished).

[20] CMS Collaboration, CERN Report No. CMS-PAS-HIG-13-005, 2013 (unpublished).

[21] ATLAS Collaboration, G. Aad *et al.*, Phys.Lett. **B726**, 120 (2013), 1307.1432.

[22] G. Altarelli and G. Parisi, Nucl.Phys. **B126**, 298 (1977).

[23] M. Dobbs *et al.*, p. 411 (2004), hep-ph/0403045.

[24] T. Sjostrand, p. 51 (2006), hep-ph/0611247.

[25] V. Sudakov, Sov.Phys.JETP **3**, 65 (1956).

[26] M. Mangano and T. Stelzer, Ann.Rev.Nucl.Part.Sci. **55**, 555 (2005).

[27] T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP **0605**, 026 (2006), hep-ph/0603175.

[28] T. Sjostrand, S. Mrenna, and P. Z. Skands, Comput.Phys.Commun. **178**, 852 (2008), 0710.3820.

[29] J. Butterworth, J. R. Forshaw, and M. Seymour, Z.Phys. **C72**, 637 (1996), hep-ph/9601371.

[30] G. Corcella *et al.*, JHEP **0101**, 010 (2001), hep-ph/0011363.

[31] K. Arnold *et al.*, (2012), 1205.4902.

[32] T. Gleisberg *et al.*, JHEP **0902**, 007 (2009), 0811.4622.

[33] S. Catani, F. Krauss, R. Kuhn, and B. Webber, JHEP **0111**, 063 (2001), hep-ph/0109231.

[34] P. Nason, JHEP **0411**, 040 (2004), hep-ph/0409146.

[35] S. Frixione, P. Nason, and C. Oleari, JHEP **0711**, 070 (2007), 0709.2092.

[36] S. Alioli, P. Nason, C. Oleari, and E. Re, JHEP **1006**, 043 (2010), 1002.2581.

[37] B. P. Kersevan and E. Richter-Was, Comput.Phys.Commun. **184**, 919 (2013), hep-ph/0405247.

[38] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, JHEP **0307**, 001 (2003), hep-ph/0206293.

[39] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, JHEP **0701**, 013 (2007), hep-ph/0611129.

[40] S. Frixione and B. R. Webber, (2008), 0812.0770.

[41] CERN Report No. ATL-PHYS-PUB-2011-011, ATL-COM-PHYS-2011-818, CMS-NOTE-2011-005, 2011 (unpublished).

[42] A. L. Read, J.Phys. **G28**, 2693 (2002).

[43] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, Eur.Phys.J. **C71**, 1554 (2011), 1007.1727.

[44] A. Wald, Transactions of the American Mathematical Society **54**, 426 (1943).

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

[46] CMS Collaboration, S. Chatrchyan *et al.*, JINST **3**, S08004 (2008).

[47] ALICE Collaboration, K. Aamodt *et al.*, JINST **3**, S08002 (2008).

[48] LHCb Collaboration, J. Alves, A. Augusto *et al.*, JINST **3**, S08005 (2008).

[49] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C70**, 787 (2010), 1004.5293.

[50] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C70**, 723 (2010), 0912.2642.

[51] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C70**, 1193 (2010), 1007.5423.

[52] ATLAS, G. Aad *et al.*, Eur.Phys.J. **C73**, 2518 (2013), 1302.4393.

[53] S. van der Meer, (1968).

[54] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C70**, 823 (2010), 1005.4568.

[55] GEANT4, S. Agostinelli *et al.*, Nucl.Instrum.Meth. **A506**, 250 (2003).

[56] I. Bird, Annual Review of Nuclear and Particle Science **61**, 99 (2011).

[57] ATLAS Collaboration, CERN Report No. ATLAS-CONF-2010-072, ATLAS-COM-CONF-2010-073, 2010 (unpublished).

[58] ATLAS Collaboration, CERN Report No. ATLAS-CONF-2010-069, 2010 (unpublished).

[59] R. Fruhwirth, W. Waltenberger, and P. Vanlaer, J.Phys. **G34**, N343 (2007).

[60] W. Lampl *et al.*, CERN Report No. ATL-LARG-PUB-2008-002. ATL-COM-LARG-2008-003, 2008 (unpublished).

[61] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C74**, 3071 (2014), 1407.5063.

[62] CERN Report No. ATLAS-CONF-2014-032, 2014 (unpublished).

[63] ATLAS Collaboration, G. Aad *et al.*, (2014), 1407.3935.

[64] M. Cacciari, G. P. Salam, and G. Soyez, JHEP **0804**, 063 (2008), arXiv:hep-ex/0802.1189.

[65] ATLAS Collaboration, G. Aad *et al.*, (2014), 1406.0076.

[66] M. Cacciari and G. P. Salam, Phys.Lett. **B659**, 119 (2008), arXiv:hep-ex/0707.1378.

[67] CERN Report No. ATLAS-CONF-2013-083, 2013 (unpublished).

[68] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C73**, 2304 (2013), 1112.6426.

[69] CERN Report No. ATL-PHYS-PUB-2014-014, 2014 (unpublished).

[70] CERN Report No. ATLAS-CONF-2014-046, 2014 (unpublished).

[71] M. zur Nedden *et al.*, CERN Report No. ATLAS-COM-CONF-2014-003, 2014 (unpublished).

[72] CERN Report No. ATLAS-CONF-2013-082, 2013 (unpublished).

[73] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C72**, 1844 (2012), 1108.5602.

[74] CERN Report No. ATLAS-CONF-2013-064, 2013 (unpublished).

[75] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C71**, 1630 (2011), arXiv:hep-ex/1101.2185.

[76] H.-L. Lai *et al.*, Phys.Rev. **D82**, 074024 (2010), 1007.2241.

[77] J. Pumplin *et al.*, JHEP **0207**, 012 (2002), hep-ph/0201195.

[78] ATLAS Collaboration, CERN Report No. ATL-PHYS-PUB-2011-008, 2011 (unpublished).

[79] ATLAS Collaboration, CERN Report No. ATL-PHYS-PUB-2011-009, 2011 (unpublished).

[80] K. Melnikov and F. Petriello, Phys.Rev. **D74**, 114017 (2006), hep-ph/0609070.

[81] M. Czakon, P. Fiedler, and A. Mitov, Phys.Rev.Lett. **110**, 252004 (2013), 1303.6254.

[82] N. Kidonakis, Phys.Rev. **D83**, 091503 (2011), 1103.2792.

[83] N. Kidonakis, Phys.Rev. **D81**, 054028 (2010), 1001.5034.

[84] N. Kidonakis, Phys.Rev. **D82**, 054018 (2010), 1005.4451.

[85] P. Nason and G. Zanderighi, Eur.Phys.J. **C74**, 2702 (2014), 1311.1365.

[86] ATLAS Collaboration, G. Aad *et al.*, (2014), hep-ex/1407.0371.

[87] ATLAS Collaboration, CERN Report No. ATLAS-CONF-2013-099, 2013 (unpublished).

[88] N. Kidonakis, p. 831 (2012), hep-ph/1205.3453.

[89] ATLAS Collaboration, G. Aad *et al.*, Eur.Phys.J. **C73**, 2306 (2013), 1210.6210.

[90] P. Golonka and Z. Was, Eur.Phys.J. **C45**, 97 (2006), hep-ph/0506026.

[91] G. Luisoni, P. Nason, C. Oleari, and F. Tramontano, JHEP **1310**, 083 (2013), 1306.2542.

[92] LHC Higgs Cross Section Working Group, S. Heinemeyer *et al.*, (2013), 1307.1347.

[93] A. Djouadi, J. Kalinowski, and M. Spira, Comput.Phys.Commun. **108**, 56 (1998), hep-ph/9704448.

[94] J. Ohnemus and W. J. Stirling, Phys.Rev. **D47**, 2722 (1993).

[95] H. Baer, B. Bailey, and J. Owens, Phys.Rev. **D47**, 2730 (1993).

[96] O. Brein, A. Djouadi, and R. Harlander, Phys.Lett. **B579**, 149 (2004), hep-ph/0307206.

[97] M. Ciccolini, S. Dittmaier, and M. Kramer, Phys.Rev. **D68**, 073003 (2003), hep-ph/0306234.

[98] A. Denner, S. Dittmaier, S. Kallweit, and A. Muck, PoS **EPS-HEP2011**, 235 (2011), 1112.5258.

[99] L. Altenkamp, S. Dittmaier, R. V. Harlander, H. Rzehak, and T. J. Zirke, JHEP **1302**, 078 (2013), 1211.5015.

[100] LHC Higgs Cross Section Working Group, S. Dittmaier *et al.*, (2011), 1101.0593.

[101] L. Moneta *et al.*, PoS **ACAT2010**, 057 (2010), 1009.1003.

[102] W. Verkerke and D. P. Kirkby, eConf **C0303241**, MOLT007 (2003), physics/0306116.

[103] G. Bohm and G. Zech, *Introduction to statistics and data analysis for physicists* (Deutsches Elektronen-Synchrotron, Hamburg, 2010).

[104] CERN Report No. ATLAS-CONF-2013-079, 2013 (unpublished).

[105] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees* (Wadsworth and Brooks, Monterey, CA, 1984).

[106] Y. Freund and R. E. Schapire, Experiments with a new boosting algorithm, in *Proceedings of the Thirteenth International Conference on Machine Learning (ICML 1996)*, edited by L. Saitta, pp. 148–156, Morgan Kaufmann, 1996.

[107] A. Hocker *et al.*, PoS **ACAT**, 040 (2007), arXiv:physics/0703039.