

Background decomposition in $Z\gamma$ events used in the search for high-mass resonances

N P Rapheeha¹, X Ruan¹, G Mokgatitwane¹, S Dahbi¹ and B Mellado^{1,2}

¹ School of Physics and Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, Wits 2050, South Africa

² iThemba LABS, National Research Foundation, PO Box 722, Somerset West 7129, South Africa

E-mail: ntsoko.phuti.rapheeha@cern.ch

Abstract. The study presents the measurement of purity of $Z\gamma$ and Z +jet background events in the search for high-mass $Z\gamma$ resonances. The study uses events where the Z boson decays into a pair of oppositely charged electrons or muons. The events used consist of 139 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13 \text{ TeV}$, recorded by the ATLAS detector at the CERN's Large Hadron Collider. The measured purity of $Z\gamma$ events depends on the parameter R that gives the correlation between the isolation and identification criteria for jets faking photons in the Z +jet events. A data-driven method that uses $\gamma\gamma$ events collected with the same detector conditions as the $Z\gamma$ events is used to determine the value of R in various bins of the photon transverse momentum. The results are compared against values obtained by computing R using a Z +jet Monte Carlo sample and a data-driven method that uses $Z\gamma$ events to estimate R .

1. Introduction

Many theories of physics Beyond the Standard Model, BSM, predict the existence of new high-mass states that can be observed as experimental signatures at the CERN's Large Hadron Collider. One of such models predicts the existence of a heavy scalar boson H which may participate in the electroweak symmetry breaking or decaying predominantly into a pair of lighter scalar boson, S [1–5].

In collider searches for H decaying into the $Z\gamma$ final state the dominant background events are expected to originate the production of non-resonant $Z\gamma$ events and the subleading background contribution from the production of a Z boson in association with jets. The Z boson decays into a pair of leptons, $\ell\ell$, $\ell = e, \mu$ where e and μ are electron and a muon, respectively. In the $Z\gamma$ events, the photon candidate is a prompt photon, which is characterised by a narrow energy cluster in the electromagnetic calorimeter and it is usually well isolated from hadronic activity. In the Z +jet background events, one jet is misidentified as a photon. The misidentified photon candidate is mainly from the decay of neutral meson, typically a π^0 , carrying a large fraction of the initial parton energy and producing an energy cluster in the electromagnetic calorimeter. The produced energy cluster has non-negligible leakage in the hadronic calorimeter. It is not isolated from hadronic activity as other particles in the same jet deposit energy in the calorimeters near the photon candidate. The photon isolation and identification, ID, variables can therefore be used to estimate the contributions of $Z\gamma$ and Z +jet events in the collected $Z\gamma$ data.

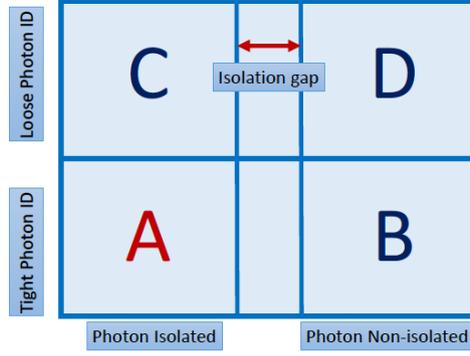


Figure 1. Definition of the ABCD regions based on photon ID and isolation.

The composition of the selected dataset is measured using the data-driven method exploited in the Run 1 SM $h \rightarrow Z\gamma$ (2D-sideband method) with prompt photons in the final state [6] and where h is the SM Higgs boson. The total background yields a smooth $Z\gamma$ invariant mass distribution which can be described by an analytic function. Knowledge of the background composition is very important in performing spurious signal studies, where the bias on the signal yield caused by choice of a particular background function is quantified [7].

1.1. Data-driven background composition estimation with the 2D-sideband method

The 2D-sideband is a counting method that relies on the definition of a two dimensional plane, as shown in Fig 1.1, based on the isolation and identification variables of the prompt photon candidate of the selected $ll\gamma$ triplet [7]. One region (A) with enhanced $Z\gamma$ contribution and three control regions (B,C, D) enriched with Z +jet events are defined in this plane as follows:

- Tight and isolated region (A): the photon candidates are well isolated from hadronic activity and pass the tight selection criteria.
- Tight but non-isolated region (B): the photon candidates are not isolated from the hadronic activity but pass the tight selection criteria.
- Non-tight, isolated region (C): the photon candidates are isolated from the hadronic activity (as in region A) but fail the tight identification criteria but pass some looser identification criteria.
- Non-tight, non-isolated region (D): the photon candidates are non-isolated (as in region B) and pass the same identification requirements of region C.

These four regions are populated with the events passing all the object selection requirements of the analysis except the photon identification and isolation requirements. Region A corresponds to the signal region used in the final measurement.

The $Z\gamma$ yield, $N_A^{Z\gamma}$, in region A is estimated from the number of events in data in the four regions, N_k^{data} ($k \in \{A, B, C, D\}$), through the relation:

$$N_A^{Z\gamma} = N_A^{\text{data}} - (N_B^{\text{data}} - c_B N_A^{Z\gamma}) \frac{(N_C^{\text{data}} - c_C N_A^{Z\gamma})}{(N_D^{\text{data}} - c_D N_A^{Z\gamma})} R^{Zj}, \quad (1)$$

where $c_k \equiv \frac{N_k^{Z\gamma}}{N_A^{Z\gamma}}$ are signal leakage fractions that are extracted from the simulated $Z\gamma$ sample and:

$$R^{Zj} \equiv \frac{N_A^{Zj} N_D^{Zj}}{N_B^{Zj} N_C^{Zj}}, \quad (2)$$

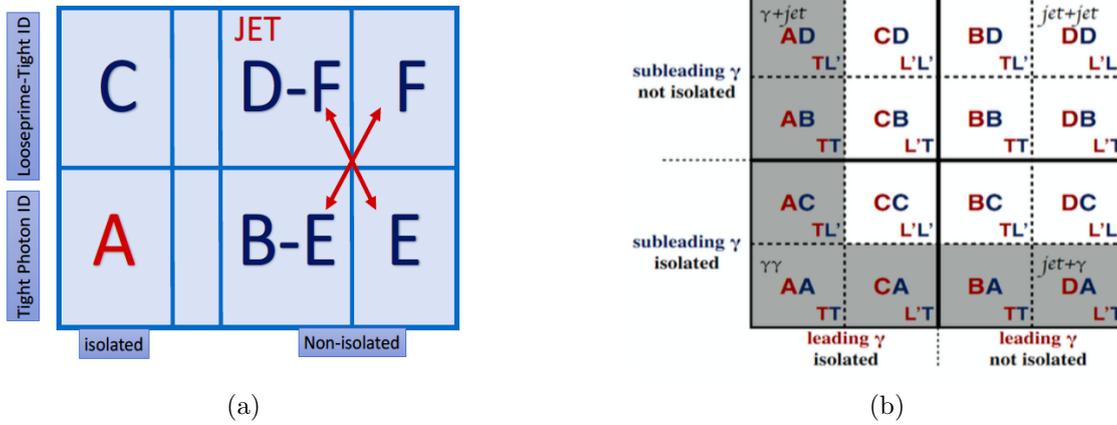


Figure 2. (a) Definition of the EF regions to directly obtain R^{Zj} directly from data. (b) Definition of the 16 regions used in the 2×2 D-sideband method of estimating purity of $\gamma\gamma$ events.

quantifies the correlation between the isolation and identification variables for the jets faking photons in Z +jet events ($R^{Zj} = 1$ in the case of vanishing correlations). The correlation parameter cannot be measured directly from data since it requires background events to fall into the signal region. The parameter R^{Zj} can be obtained from the Z +jet full Monte Carlo, MC, simulated sample, to validate the obtained value of R^{Zj} , data-driven methods are devised.

2. Methodology

2.1. Data-driven estimation of R using $Z + \gamma$ events

The R^{Zj} is estimated by defining two new non-isolated regions called region "E" and region "F", as illustrated in Fig. 2(a). The region E is extra tight, while region F is extra loose as compared to region D. These two regions are defined so that the R^{Zj} from Eq. 2 can be estimated directly from data instead of MC, Eq. 2 becomes Eq. 3:

$$R^{\text{data}, Z\gamma} = \frac{N_{B-E}^{\text{data}} N_F^{\text{data}}}{N_{D-F}^{\text{data}} N_E^{\text{data}}}. \quad (3)$$

The following event selection was used to select $Z\gamma$ events used in the estimation of $R^{\text{data}, Z\gamma}$. The reconstructed $\ell\ell\gamma$ system is required to have an invariant mass greater than 130 GeV and less than 2500 GeV. The invariant mass of the reconstructed $\ell\ell$ candidates is required to be within 15 GeV of the Z boson pole mass. The photon candidate is required to have a minimum p_T of 40 GeV. The tight photon identification and the FixedCutLoose isolation requirements were used for the identification and isolation requirements [8].

A sample of Z +jet events was simulated at next-to-leading order, NLO, in quantum chromodynamics using POWHEG [9] showered with PYTHIA8 [10] event generators. The CT10 [11] parton distribution function, PDF, set was in the matrix element. To model the non-perturbative effects, the AZNLO set of tuned parameters [12] alongside the CTEQ6L1 PDF set [13] are used. The MC sample will be used to compute R^{Zj} , as per Eq. 2. The event selection used to select $Z\gamma$ events was used in selecting the Z +jet events.

2.2. Data-driven estimation of R using $\gamma\gamma$ events

To verify the accuracy of the purity computed with $R^{\text{data}, Z\gamma}$ estimated with Eq. 3, a new data driven method is used. The main idea of the method is to use an X +jet sample, where X is

a well defined object, in order to estimate R in the same regions as ones used to define R^{Zj} in Eq. 2. A $\gamma\gamma$ data sample composed of two real photons, 1 real photon and a hadronic jet and two hadronic jets is used. Each photon candidate is classified as either belonging to a category A, B, C or D, depending on whether it fails or passes the identification criteria, as shown in Fig. 2(b). The two candidates are considered sequentially; the 2D-sideband method is first applied to the photon candidate with a leading photon transverse momentum, p_T , to extract events for which the leading p_T candidate is a true photon and then the method is applied to the subleading p_T photon candidate knowing that the leading photon is tightly identified and well isolated. This results in 16 orthogonal regions, shown in Fig. 2(b).

A X +jet sample is formed by fixing either the leading or subleading photon to be well isolated and passing the tight identification requirement to be X . To obtain the desired X +jet sample from the $\gamma\gamma$ data sample, pure $\gamma\gamma$ events are subtracted from the data sample. The number of X +jet events in a region of interest is given by:

$$N_j^{\text{data}} = N_j^{\text{data}} - N_j^{\text{M.C}} \times k\text{-factor}, \quad (4)$$

where j is a region of interest and $k\text{-factor} = N_{AA}^{\text{data}} \times \frac{\text{purity}}{N_{AA}^{\text{MC}}}$. The purity of $\gamma\gamma$ events is computed using the 2×2 D-sideband method used in ATLAS studies [14]. The method performs a data-driven background evaluation by extrapolating the background from the control regions defined in the sidebands of isolation and identification variables in which the photons either pass or fail the Tight ID criteria or fail the isolation selection. The purity of $\gamma\gamma$ events is computed in a region where both the leading and subleading photons are well isolated and identified. The regions used in the 2×2 D-sideband method are shown in Fig. 2 (b).

The data driven correlation parameter in a case where the leading p_T photon candidate is classified as being tight and isolated, $X = \gamma$, is computed as:

$$R^{\gamma+jet} = \frac{N_{AA}^{\gamma+jet}/N_{AB}^{\gamma+jet}}{N_{AC}^{\gamma+jet}/N_{AD}^{\gamma+jet}}. \quad (5)$$

In the case where the subleading p_T photon candidate is selected as X the correlation parameter is defined as:

$$R^{jet+\gamma} = \frac{N_{AA}^{jet+\gamma}/N_{BA}^{\gamma+jet}}{N_{CA}^{\gamma+jet}/N_{DA}^{\gamma+jet}}. \quad (6)$$

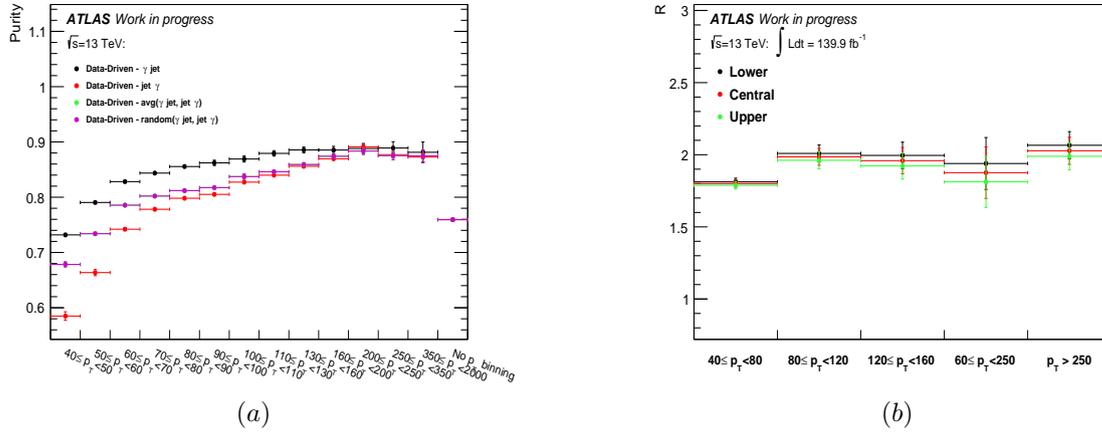
The following event selection was used to select $\gamma\gamma$ events used to compute $R^{\text{data},\gamma\gamma}$. The photon pairs are required to have a reconstructed invariant mass between 130 GeV and 2500 GeV. The leading photon is required to have a minimum p_T of 40 GeV and the subleading photon has a minimum p_T of 30 GeV. The photon isolation and identification requirements are the same as the ones used for $Z\gamma$ events.

3. Results

The values of R obtained using the data-driven $Z\gamma$ method and the Z +jet MC are summarised in Table 1. The obtained values of $R^{\text{data},\gamma\gamma}$ will depend on whether leading or subleading photon in the $\gamma\gamma$ data sample was selected as X when forming a X +jet sample. To eliminate potential bias on the selection of X will be done at random. Figure. 3 (a) shows purities obtained when the leading photon is selected as X (X +jet), the subleading photon selected as X (jet+ X) and when the selection between either photon is done randomly. It is noted that the $\gamma\gamma$ purity obtained when the X candidate is selected at random is the same as the average between X +jet and jet+ X purities. The purity obtained by selection X at random is used to compute values of $R^{\text{data},\gamma\gamma}$.

Table 1. Estimated values of R obtained using the data-driven ABCDEF method, MC Z +jet events and the data-driven $\gamma\gamma$ events.

p_T^γ bin [GeV]	$R^{\text{data},Z\gamma}$	$R^{Zj}(Z+\text{jet MC})$	$R^{\text{data},\gamma\gamma}$
$40 \leq p_T < 80$	1.36 ± 0.03	1.30 ± 0.04	1.80 ± 0.04
$80 \leq p_T < 120$	1.21 ± 0.06	1.55 ± 0.13	1.99 ± 0.08
$120 \leq p_T < 160$	1.28 ± 0.11	1.70 ± 0.27	1.96 ± 0.13
$60 \leq p_T < 250$	1.53 ± 0.16	1.96 ± 0.44	1.88 ± 0.24
$p_T > 250$	1.27 ± 0.20	1.60 ± 0.52	2.03 ± 0.13

**Figure 3.** (a) di-photon purity at different photon p_T bins. (b) The R parameter computed using the central value of $\gamma\gamma$ purity \pm the statistical uncertainty of the purity.

The measured purity of $\gamma\gamma$ events has a significant impact of the computed value of $R^{\text{data},\gamma\gamma}$. Fig. 3(b) shows the values of $R^{\text{data},\gamma\gamma}$ using the central values of the $\gamma\gamma$ purity, lower and upper values of purity. The lower and upper values of the $\gamma\gamma$ purity are obtained by subtracting or adding the statistical uncertainty of $\gamma\gamma$ purity from the central value. The largest difference between R values computed using the central of purity and the values computed using lower and upper values is propagated as the systematic uncertainty of $R^{\text{data},\gamma\gamma}$. The $R^{\text{data},\gamma\gamma}$ values obtained using the $\gamma\gamma$ events are summarised in Table 1. The reported uncertainties of $R^{\text{data},\gamma\gamma}$ consist of the statistical uncertainty and the systematic uncertainty.

The comparison between R values obtained using the three different samples is shown in Figure 4 (a). In the low mass region, where the photon p_T is less than 80 GeV, the R value computed with the $e Z$ +jet MC is much closer to the one computed using the ABCDEF method. As the photon p_T increases, going to higher $Z\gamma$ masses, the R value computed with the Z +jet MC gets closer to R computed with the $\gamma\gamma$ data-driven method.

The comparison of the purity in each photon p_T bin using the different methods of computing R is shown in Figure 4 (b). It can be seen that value of R used does not impact the measured purity that much. It can be seen that as the photon p_T increases the value of purity obtained using the $\gamma\gamma$ data-driven method gets closer to the value obtained using R from the MC.

4. Conclusions

The background decomposition study in search for $Z\gamma$ events that can decay from a potential heavy scalar resonance has been performed. The purity of $Z\gamma$ events in data was measured using

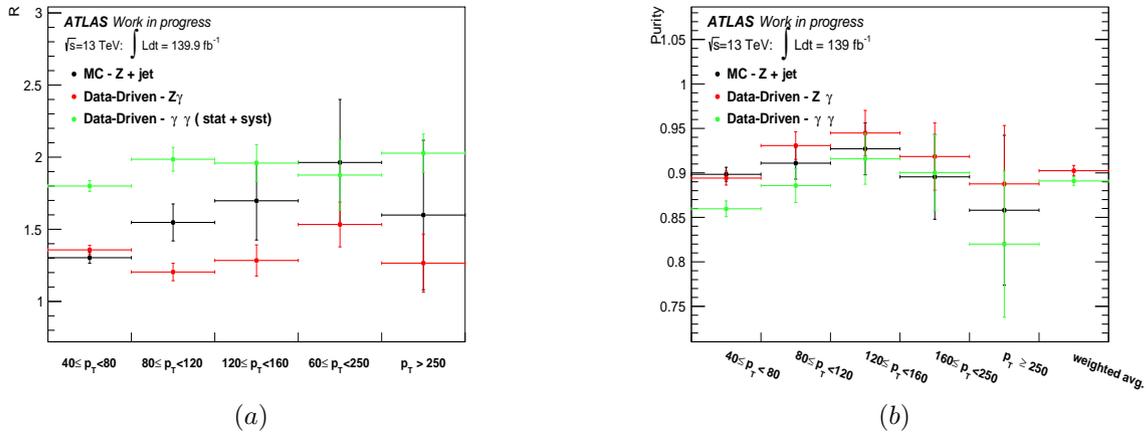


Figure 4. (a) Computed R values in different photon p_T bins. (b) The observed $Z\gamma$ purity using R values computed using the different methods in different photon p_T bins.

R computed with a data-driven method that uses $\gamma\gamma$ events, a data-driven method that uses $Z\gamma$ events and Z +jet MC. It was observed that the purity obtained using R^{Zj} computed with the Z +jet MC and $R^{\text{data},Z\gamma}$ to be comparatively the same for photons with p_T less than 80 GeV. As the photon p_T increases above 80 GeV the purity of $Z\gamma$ measured with R^{Zj} obtained with the Z +jet MC gets closer to the purity calculated with $R^{\text{data},\gamma\gamma}$ obtained from the data-driven $\gamma\gamma$ method. The three methods of estimating R resulted in the measured $Z\gamma$ that is relatively the same, except for the minor differences highlighted above.

References

- [1] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B and Reed R G 2015 (*Preprint* 1506.00612)
- [2] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B, Reed R G and Ruan X 2016 *Eur. Phys. J. C* **76** 580 (*Preprint* 1606.01674)
- [3] Buddenbrock S, Cornell A S, Fang Y, Fadol Mohammed A, Kumar M, Mellado B and Tomiwa K G 2019 *JHEP* **10** 157 (*Preprint* 1901.05300)
- [4] Hernandez Y, Kumar M, Cornell A S, Dahbi S E, Fang Y, Lieberman B, Mellado B, Monnagotla K, Ruan X and Xin S 2021 *Eur. Phys. J. C* **81** 365 (*Preprint* 1912.00699)
- [5] von Buddenbrock S, Ruiz R and Mellado B 2020 *Phys. Lett. B* **811** 135964 (*Preprint* 2009.00032)
- [6] ATLAS Collaboration 2011 *Phys. Rev. D* **83**(5) 052005
- [7] ATLAS Collaboration 2020 Recommendations for the Modeling of Smooth Backgrounds Tech. rep. CERN Geneva
- [8] ATLAS Collaboration 2019 *Journal of Instrumentation* **14** P12006–P12006 (*Preprint* 1908.00005)
- [9] Alioli S, Nason P, Oleari C and Re E 2008 *Journal of High Energy Physics* **2008** 060–060 (*Preprint* 0805.4802)
- [10] Sjostrand T, Mrenna S and Skands P Z 2008 *Computer Physics Communications* **178** 852–867 ISSN 0010-4655 (*Preprint* 0710.3820)
- [11] Lai H L, Guzzi M, Huston J, Li Z, Nadolsky P M, Pumplin J and Yuan C P 2010 *Phys. Rev. D* **82** 074024 (*Preprint* 1007.2241)
- [12] ATLAS Collaboration 2014 *JHEP* **09** 145 (*Preprint* 1406.3660)
- [13] Pumplin J, Stump D R, Huston J, Lai H L, Nadolsky P M and Tung W K 2002 *JHEP* **07** 012 (*Preprint* hep-ph/0201195)
- [14] ATLAS Collaboration 2012 *Phys. Rev. D* **85**(1) 012003