Chapter 5

Jets faking photons

The jets faking photons background represents approximately the 8% of the total background of the Mono-Photon analysis and it is due to Z + jet or W + jet events in which the jet is mistakenly reconstructed and identified as a photon. The classification of the reconstructed object into a photon or a jet is performed by an algorithm based on the detector inputs. This algorithm necessarily introduces an efficiency for true photons and a jet contamination in a selected sample of photons. The simulations do not describe accurately the performance of this algorithm, so the jet $\rightarrow \gamma$ fraction is estimated with a purely data-driven technique. This is done applying the *Two Dimensional Sideband Method*, which was used in 2015 and 2016 analysis too. In this chapter the method, the validation and the results of the estimation of the jets faking photons background in all the regions of the 2019 ATLAS' Mono-Photon analysis using the full 2015-2018 Run 2 statistics will be presented.

5.1 Basic method

The *Two Dimensional Sideband Method*, also known as the ABCD method, is an almost data-driven technique employed to determine a background contamination in a given signal region. This method relies on counting photon candidates in four regions of a two-dimensional plane defined by an isolation variable and an identification variable (tightness). The identification variable is represented by the value of $TopoEtCone40 - 0.022p_T - 2.45$ GeV, described in Sec. 3.1.4: the candidate photon is considered isolated if it is lower than 0 GeV, not-isolated if it is greater than 3 GeV. Note that the gap of 3 GeV is defined to minimize signal leakage in the CRs. On the other hand the identification variable is divided in two bins corresponding to the Tight and Tight-4 selections described in Sec. 3.1.3. Tight-3 and Tight-5 selections will be used to estimate systematic uncertainties.

The plane is so divided into four regions as shown in Fig. 5.1, where the signal region corresponds to the Tight-Isolated region and the three CRs are assumed to be populated only by background photons. In each region the number of candidate photons is:

- N^A: number of Tight - Isolated candidates;

- N^B: number of Tight - Not-isolated candidates;

- M^A: number of Tight-4 - Isolated candidates;

- M^B: number of Tight-4 - Not-isolated candidates.

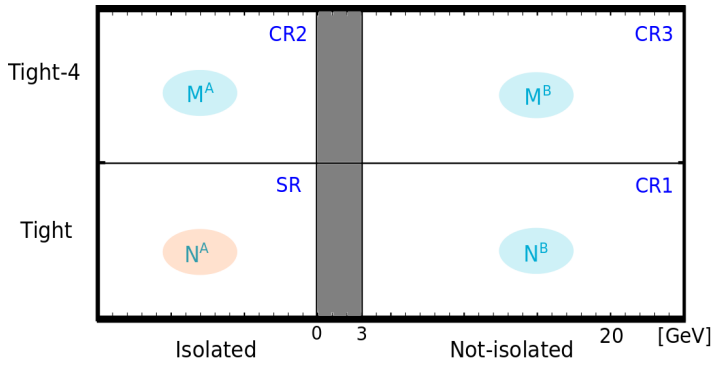

Figure 5.1: Scheme of the regions defined by the Two Dimensional Sideband Method.

For real data the numbers of signal and background photon candidates in the SR are defined as N^A_{sig} and N^A_{bkg} respectively, while for MC simulation as N^A_{sigMC} and N^A_{bkgMC} (and similarly in the CRs).

The method relies on two simplifying hypotheses:

1. the correlation between tightness and isolation is negligible for the background;
2. in the three CRs the number of signal photon candidates is negligible compared to the number of fake candidates:

$$\begin{aligned} N^B_{bkg} &\gg N^B_{sig} \\ M^A_{bkg} &\gg M^A_{sig} \\ M^B_{bkg} &\gg M^B_{sig} \end{aligned} \tag{5.1}$$

The first assumption leads to:

$$\frac{N^A_{bkg}}{N^B_{bkg}} = \frac{M^A_{bkg}}{M^B_{bkg}} \tag{5.2}$$

while the second one implies that:

$$\begin{aligned} N^B &= N^B_{bkg} \\ M^A &= M^A_{bkg} \\ M^B &= M^B_{bkg} \end{aligned} \tag{5.3}$$

Therefore, combining the two hypotheses:

$$N^A_{bkg} = N^B \frac{M^A}{M^B} \tag{5.4}$$

that leads directly to:

$$N^A_{sig} = N^A - N^A_{bkg} = N^A - N^B \frac{M^A}{M^B} \tag{5.5}$$

$$P := \frac{N^A_{sig}}{N^A} = 1 - \frac{N^B}{N^A}\frac{M^A}{M^B} \tag{5.6}$$

where the purity P of the sample is defined as the ratio of the number of signal photons and the total number of candidates in the SR.

A purely data-driven formula for the purity is then obtained even if it requires that both the two hypotheses are satisfied, although this is not always the case.

5.1.1 Correlation in the background

By means of MC simulations it is possible to take into account non-negligible correlations in the background 2D distributions. Assuming that MC and real data are in good agreement for background events, the number of candidates in a given region can be approximated with the corresponding MC prediction. This is done in the following way:

$$\begin{aligned} N^A_{sig} &= N^A - N^A_{bkg} = N^A - N^A_{bkg}\frac{N^B_{bkg}\times M^A_{bkg}/M^B_{bkg}}{N^B_{bkg}\times M^A_{bkg}/M^B_{bkg}} \\ &= N^A - \left(N^B_{bkg}\frac{M^A_{bkg}}{M^B_{bkg}}\right)\left(\frac{N^A_{bkg}}{N^B_{bkg}}\frac{M^B_{bkg}}{M^A_{bkg}}\right) \\ &\approx N^A - \left(N^B_{bkg}\frac{M^A_{bkg}}{M^B_{bkg}}\right)\left(\frac{N^A_{bkgMC}}{N^B_{bkgMC}}\frac{M^B_{bkgMC}}{M^A_{bkgMC}}\right) \\ &\approx N^A - \left(N^B\frac{M^A}{M^B}\right)\left(\frac{N^A_{bkgMC}}{N^B_{bkgMC}}\frac{M^B_{bkgMC}}{M^A_{bkgMC}}\right) \\ &= N^A - \left(N^B\frac{M^A}{M^B}\right)R_{MC} \end{aligned} \tag{5.7}$$

that implicitly defines the *correlation factor* R_{MC} which can be estimated by means of pure background MC simulations. Its statistical uncertainty is obtained by the error propagation:

$$\sigma_{R_{MC}} = R_{MC}\sqrt{\frac{1}{N^A_{bkgMC}} + \frac{1}{N^B_{bkgMC}} + \frac{1}{M^A_{bkgMC}} + \frac{1}{M^B_{bkgMC}}} \tag{5.8}$$

which depends only on the available MC statistic.

5.1.2 Signal leakage

MC simulations can also estimate the signal leakage from the SR to each one of the CRs, releasing thus the second hypothesis. As for the correlation factor, real photons are approximated with MC simulations, assuming non negligible signal contamination in the CRs, the number of events in these regions can be written as:

$$\begin{aligned} N^B &= N^B_{bkg} + N^B_{sig} = N^B_{bkg} + N^A_{sig}\frac{N^B_{sig}}{N^A_{sig}} \\ M^A &= M^A_{bkg} + M^A_{sig} = M^A_{bkg} + N^A_{sig}\frac{M^A_{sig}}{N^A_{sig}} \\ M^B &= M^B_{bkg} + M^B_{sig} = M^B_{bkg} + N^A_{sig}\frac{M^B_{sig}}{N^A_{sig}} \end{aligned} \tag{5.9}$$

Three *signal leakage coefficients* are then defined as:

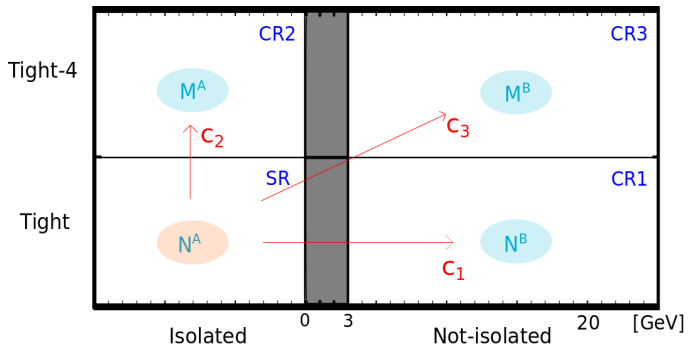

Figure 5.2: Representation of the signal leakage coefficients.

$$\begin{aligned}
\frac{N^B_{sig}}{N^A_{sig}} &:= c_1 \approx \frac{N^B_{sigMC}}{N^A_{sigMC}} \\
\frac{M^A_{sig}}{N^A_{sig}} &:= c_2 \approx \frac{M^A_{sigMC}}{N^A_{sigMC}} \\
\frac{M^B_{sig}}{N^A_{sig}} &:= c_3 \approx \frac{M^B_{sigMC}}{N^A_{sigMC}}
\end{aligned} \tag{5.10}$$

A schematic representation of the meaning of signal leakage coefficients is reported in Fig. 5.2.

Applying these coefficients to Eq. 5.5 leads to:

$$N^A_{sig} = N^A - N^B_{bkg}\frac{M^A_{bkg}}{M^B_{bkg}} = N^A - (N^B - N^A_{sig}c_1)\frac{M^A - N^A_{sig}c_2}{M^B - N^A_{sig}c_3} \tag{5.11}$$

and resolving for N^A_{sig} gives the number of signal photons in the SR corrected for signal leakage in the CRs:

$$N^A_{sig} = \frac{(M^B + N^A c_3 - N^B c_2 - M^A c_1)\left(-1 + \sqrt{1 + \frac{4(c_1c_2 - c_3)(N^A M^B - N^B M^A)}{(M^B + N^A c_3 - N^B c_2 - M^A c_1)^2}}\right)}{2(c_1 c_2 - c_3)} \tag{5.12}$$

5.1.3 Signal yields and purity

It is now possible to account for both signal leakage and correlation in the background at the same time. Starting from Eq. 5.7:

$$N^A_{sig} \approx N^A - \left(N^B_{bkg}\frac{M^A_{bkg}}{M^B_{bkg}}\right)\left(\frac{N^A_{bkgMC}}{N^B_{bkgMC}}\frac{M^B_{bkgMC}}{M^A_{bkgMC}}\right) \tag{5.13}$$

replacing N^B_{bkg} with $N^B - N^A_{sig}c_1$ and so on:

$$N^A_{sig} \approx N^A - \left((N^B - N^A_{sig}c_1)\frac{M^A - N^A_{sig}c_2}{M^B - N^A_{sig}c_3}\right)R_{MC} \tag{5.14}$$

some simple algebra gives N^A_{sig} and dividing by N^A one finds a formula for the purity which accounts for both corrections:

$$P = \frac{\left(M^B + N^A c_3 - N^B c_2 R_{MC} - M^A c_1 R_{MC}\right)}{2N^A(c_1 c_2 R_{MC} - c_3)} \cdot \left(-1 + \sqrt{1 + \frac{4(c_1 c_2 R_{MC} - c_3)(N^A M^B - N^B M^A R_{MC})}{(M^B + N^A c_3 - N^B c_2 R_{MC} - M^A c_1 R_{MC})^2}}\right) \tag{5.15}$$

The derivatives needed for errors propagation are computed with Mathematica [23], a modern technical computing system that enables, among other things, to compute derivatives of an analytical expression of a function.

The systematic uncertainties are estimated varying different assumptions of the method. For example the not-tight control region is moved from Tight-4 to Tight-3 and Tight-5 and two new purities are obtained. The maximum difference between these purities with the nominal one is quoted as systematic uncertainty from the not-tight selection. The same will be done moving the isolation gap from 3 GeV to 2 GeV and 4 GeV, and similarly for the sysematic uncertainties on the four coefficients. The total systematic uncertainty is the square sum of all systematic uncertainties.

5.2 Validation test

To evaluate the consistency of the method, it has been tested on a mixed MC sample of W($\mu\nu$) + γ and W + jets events with known purity. The samples are all MC generated with Sherpa [24] at NLO[4].

The correlation factor and the signal leakage coefficients have been computed on the mixed sample matching the photon candidates with true-level[5] jets (for R_{MC}) and with true-level photons (for c_1,c_2 and c_3). A particular focus is given to the SR - ISR1 as it is the most populated region.

Results in the SR - ISR1 for the coefficients, with statistical uncertainties, are reported in Table 5.1:

R_{MC}	c_1	c_2	c_3
2.76 ± 0.49	7.08 ± 0.80 %	4.50 ± 0.63 %	0.49 ± 0.21 %

Table 5.1: Coefficients used in the validation test.

The results for the purities with propagated and systematic uncertainties respectively are reported in Table 5.2:

	Purity
True	84.27 %
Calculated	86.41 ± 5.87 ± 7.26 %

Table 5.2: Table of purities calculated to validate the ABCD method.

[4] Events are typically weighted by the process cross section. MC samples at NLO (Next to Leading Order) are characterized by weights assigned to each event to properly reproduce the normalization and kinematics of the process.

[5] MC simulations are obtained in three steps: at first the particles are generated, then they are simulated to interact with the detector and in the end they are reconstructed with the same procedure used for real data. In this process the generation stage is also known as *true-level*.

As can be noticed the calculated purities are compatible within the errors with the expected results.

5.3 Coefficients from MC

Once validated the method on a simplified MC sample it is possible to start analyzing different MCs (all NLO Sherpa samples) to get the coefficients needed by the method and assess the systematic uncertainties.

At first very strange results have been observed due to large negative weights, in few events especially in the CRs of the Mono-Photon analysis. These anomalous weights led to very high unnatural correlation factors.

This difficulty has been managed in the following way:

- weights with an absolute value > 100 are rescaled to 1;
- events with $p_T < 140$ GeV at truth-level are excluded: the analysis selects candidate photons with $p_T > 150$ GeV, but some events can be reconstructed with higher p_T, with respect to true-level p_T, possibly gaining a very high cross section and therefore a high weight;
- regions where a certain sample (W/Z + γ/jets) is dominant have been merged to increase statistics:
  - W + γ/jets $\rightarrow$ SR + 1muCR;
  - Z + γ/jets $\rightarrow$ SR + 2muCR + 2eCR;
  - the gammajetCR has been treated separately as it has a different E_T^{miss} cut.
- the track isolation ($ptcone20/p_T$) has been released in the CRs of the method:
  - not-isolated events are now those which fail the calorimetric isolation OR the track isolation;
  - a gap of 0.05 on the track isolation variable has been excluded to prevent signal leakage.

Upper limits on track and calorimetric isolation have been set, respectively to 1 and 140 GeV, to exclude pathological events.

5.3.1 Signal leakage coefficients

Signal leakage coefficients have been calculated using three signal samples of Z(ll) + γ, Z($\nu\nu$) + γ and W + γ. Results are reported in Fig. 5.3, 5.4 and 5.5.

A systematic difference between c_1 coefficients calculated from W and Z bosons samples is evident in the first plot. Fig. 5.6 shows the normalized calorimetric isolation profiles in the ISR1, where a difference in the tails of the distributions is clearly visible, that results in a systematic difference in the coefficients that accounts for signal leakage from isolated to not-isolated regions. Since the origin of this effect is not fully understood a systematic uncertainty is assigned to c_1 coefficients to cover the differences observed between different samples. Also c_3 is thus affected by this systematic, even if the results are compatible.

It can be noticed also a systematic increase of c_1 and c_3 with the increase of the E_T^{miss} threshold. This behaviour is once again explained by the calorimetric isolation profiles, in

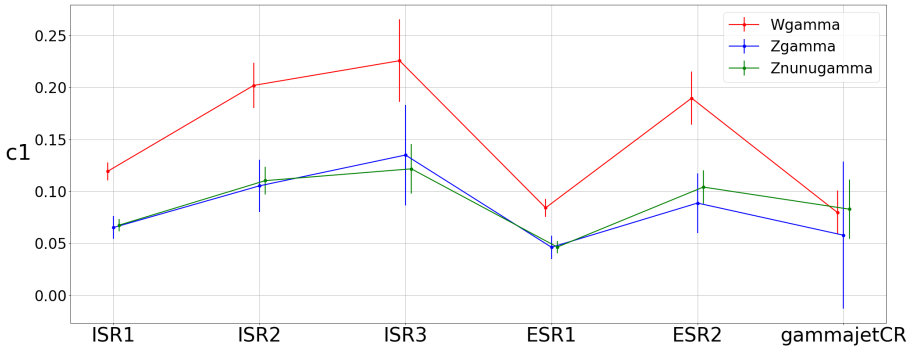

Figure 5.3: Representation of the signal leakage coefficients c_1, accounting for signal leakage in the tight - not-isolated region.

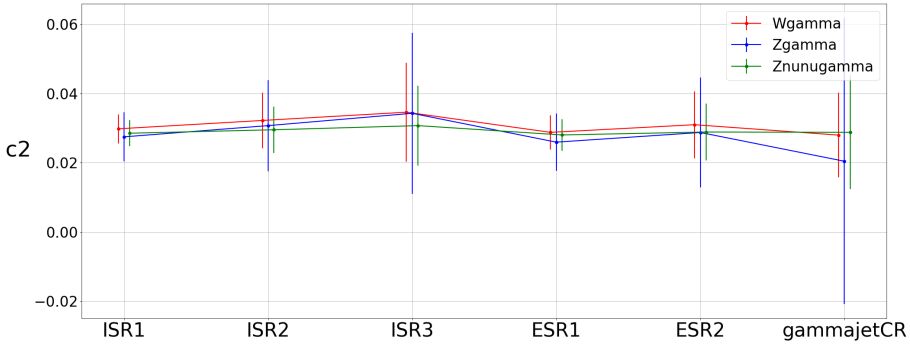

Figure 5.4: Representation of the signal leakage coefficients c_2, accounting for signal leakage in the not-tight - isolated region.

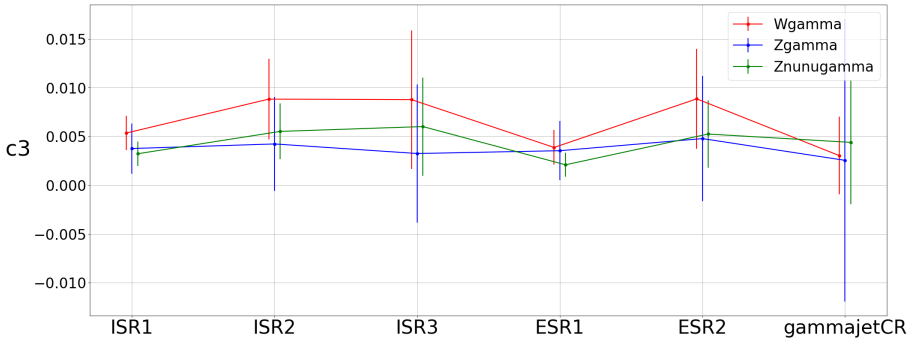

Figure 5.5: Representation of the signal leakage coefficients c_3, accounting for signal leakage in the not-tight - not-isolated region.

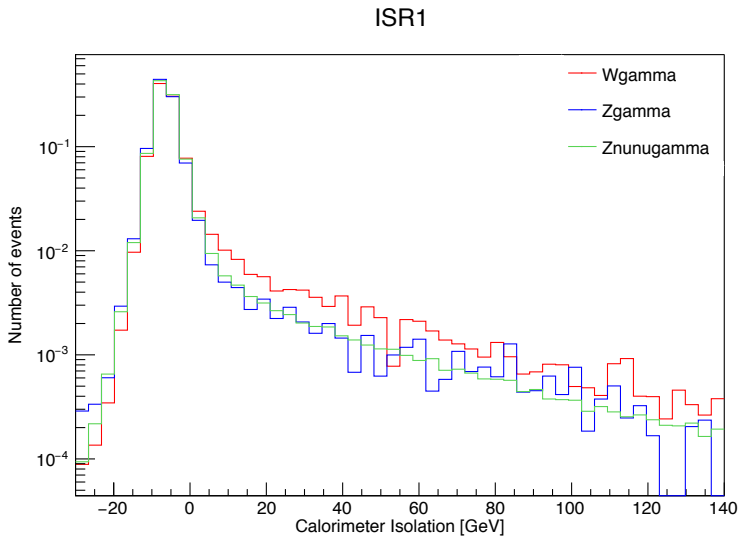

Figure 5.6: Normalized calorimetric isolation profiles in the ISR1 of the three MC samples used to compute the signal leakage coefficients.

particular in the tails. Fig. 5.7 shows the calorimetric isolation profiles of the Z($\nu\nu$) + γ sample in the inclusive signal regions: as can be seen the tails move upward with the increase of the E_T^{miss} threshold, resulting in increasing coefficients values.

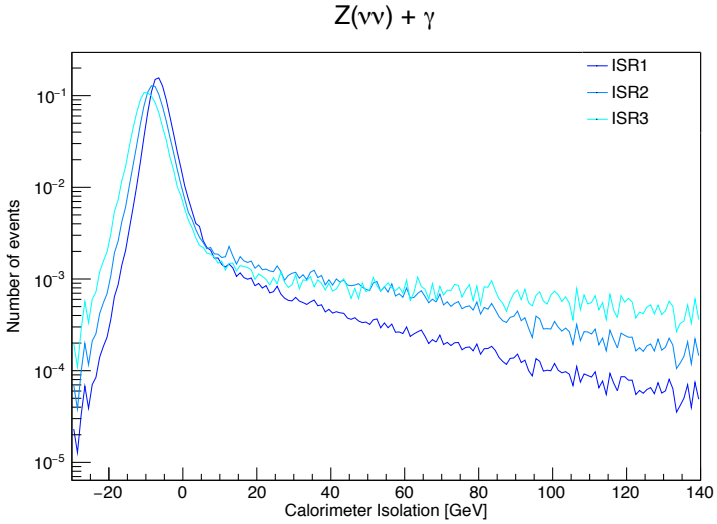

Figure 5.7: Normalized calorimetric isolation profiles, in logarithmic scale, of the Z($\nu\nu$) + γ sample in the ISR1, ISR2 and ISR3.

5.3.2 Correlation factor

The correlation factors have been calculated separately on two background samples of Z + jets and W + jets. Results are reported in Fig. 5.8.

As can be seen in the plot, the results are compatible within the uncertainties and don't show any evident systematics.

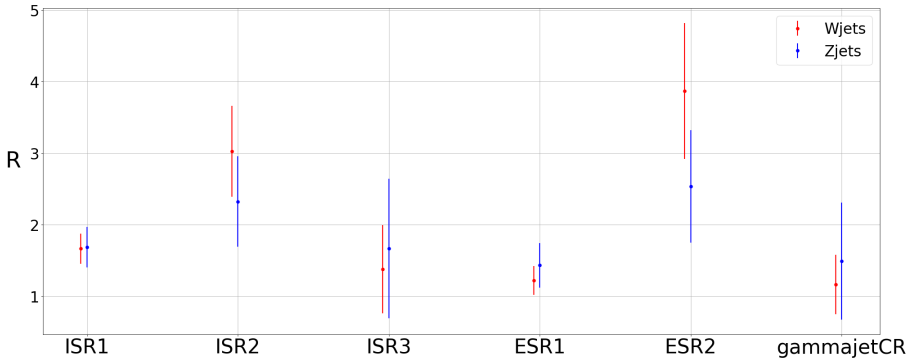

Figure 5.8: Representation of the results obtained for the correlation factor using two samples of Z + jets and W + jets.

5.3.3 R prime

To evaluate the accuracy of the MC description of the correlation factor a special correlation factor (R_{prime}) in a completely not-isolated region has been computed: in this way it is possible to calculate it for both real data and MC without signal contamination problems. The MC used are the same as for the usual correlation factor, while data are the 2015-16 data. The populations in the not-isolated regions are divided by rectangular selections as Fig. 5.9 shows. The cuts on the track isolation and calorimetric isolation are chosen in order to maintain a good population in all these regions and are set to 0.2 and 50 GeV respectively, so that:

- N^A_{prime}: events with (3 GeV $<$ calo isolation $<$ 50 GeV and track isolation $<$ 0.2) or (calo isolation $<$ 3 GeV and 0.1 $<$ track isolation $<$ 0.2)
- N^B_{prime}: events with (50 GeV $<$ calo isolation $<$ 100 GeV and track isolation $<$ 1.0) or (calo isolation $<$ 50 GeV and 0.2 $<$ track isolation $<$ 1.0)

and similarly for M^A_{prime} and M^B_{prime}

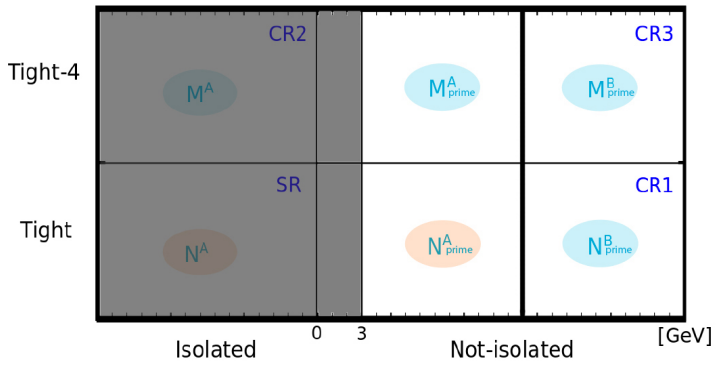

Figure 5.9: Scheme of the new regions defined to compute R_{prime}.

Results are reported in Fig. 5.10. The MC predictions for R_{prime} and are compatible within uncertainties with those extracted from 2015-16 data, except in the gammajetCR where the statistics is very poor. Both data and MCs exhibits a R_{prime} correlation factor close to 1.

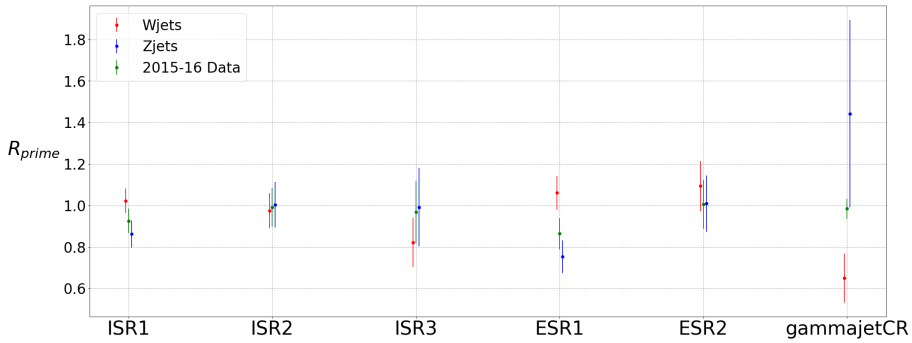

Figure 5.10: Representation of the results obtained for R_{prime} using two samples of Z + jets and W + jets.

5.3.4 Analysis

For all the coefficients (c_1, c_2, c_3 and R_{MC}) a weighted mean over different MC samples is computed with its statistical uncertainties for the five SRs and the gammajetCR. Systematic uncertainties on c_1 and c_3 has been assigned as the RMS of the values from different MC. These uncertainties are propagated to the purity as the difference between the values obtained using c_x varied up and down by the size of its uncertainty.

The 5+1 coefficients are applied to the 20+1 region of the analysis: each value is used in the corresponding SR and in all the Mono-Photon CRs (for example the results in the ISR1 will be used in the SR - ISR1, 1muCR - ISR1, 2muCR - ISR1 and 2eCR - ISR1) except for the coefficients in the gammajetCR that are treated only in this particular CR.

5.4 Results

It is now possible to compute the purities on 2015-16 data and compare them with the previous analysis results.

Statistical uncertainties are computed with the propagation of errors only from the real data populations, while the errors propagated from the coefficients are quoted as systematic errors as they are determined by the limited statistics of the MCs.

Systematic uncertainties on the purities, on the other hand, are obtained as discussed in Sec. 5.1.3 and 5.3.4. Having released the track isolation in CRs of the method, the gap between the isolated and not-isolated regions needs to be redefined. The now rectangular gap is varied respectively on the track isolation and calorimetric isolation as (0.05, 3 GeV) $\rightarrow$ (0.01, 2 GeV) or (0.05, 3 GeV) $\rightarrow$ (0.10, 4 GeV). The different systematic errors are finally summed in quadrature with the statistical errors propagated from the coefficients, giving the final systematic error on the purities. Finally the total error is the sum in quadrature of the total systematic and statistical errors.

Fig. 5.11 reports the results of fake photons numbers, both for the previous analysis and the published analysis [9] on 2015-16 data, calculated as:

$$N_{FakePhotons} = (1 - P)N^A \tag{5.16}$$

where the errors are obtained simply by multiplying the error on P by N^A. The two analysis results are compatible within uncertainties in each region of the analysis, although in the new analysis the systematic uncertainties are larger.

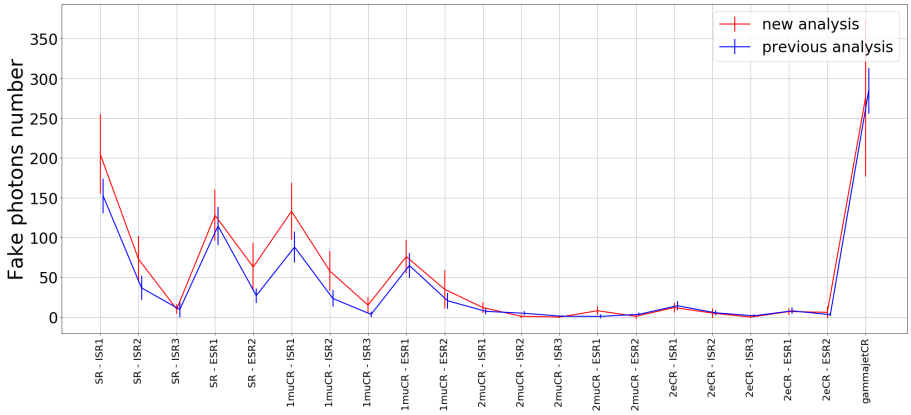

Figure 5.11: Comparison between results of the previous analysis (blu) and of the new analysis (red) on data taken during 2015-16.

Table 5.3 shows the details of the results obtained by the ongoing analysis on 2015-16 data. The results for the fake photons numbers in each region are dominated by the systematic error coming from the tightness control regions and by the statistical uncertainty on R_{MC}, while the uncertainties on the other coefficients don't really affect the total error.

regions	mean	stat.	tightness syst.	isolation syst.	c1 stat.	c2 stat.	c3 stat.	R stat.	c1 syst.	c3 syst.	total syst.	total error
SR - ISR1	205	21	36	3	3	8	0	22	13	0	45	50
SR - ISR2	73	15	19	2	2	6	0	13	6	0	25	29
SR - ISR3	10	4	2	0	1	2	0	4	1	0	5	6
SR - ESR1	128	15	20	5	2	6	0	17	7	0	29	32
SR - ESR2	63	16	20	3	2	6	0	14	5	0	26	30
1muCR - ISR1	133	15	28	5	1	3	0	14	4	0	32	36
1muCR - ISR2	58	12	19	3	1	2	0	10	2	0	22	25
1muCR - ISR3	15	5	6	1	0	1	0	6	0	0	8	10
1muCR - ESR1	76	10	14	2	1	2	0	10	2	0	18	21
1muCR - ESR2	35	9	21	1	1	2	0	7	2	0	22	24
2muCR - ISR1	12	4	5	0	0	1	0	1	1	0	6	7
2muCR - ISR2	1	2	1	0	0	1	0	0	0	0	1	2
2muCR - ISR3	-0	0	0	0	0	0	0	0	0	0	0	1
2muCR - ESR1	8	3	3	1	0	0	0	1	1	0	4	5
2muCR - ESR2	1	3	1	0	0	1	0	0	0	0	1	3
2eCR - ISR1	12	5	3	1	0	0	0	1	1	0	4	6
2eCR - ISR2	5	4	4	0	0	0	0	1	0	0	4	6
2eCR - ISR3	0	1	0	0	0	0	0	0	0	0	0	1
2eCR - ESR1	7	3	3	0	0	0	0	1	1	0	3	4
2eCR - ESR2	6	5	5	1	0	0	0	1	0	0	5	7
gammajetCR	276	18	26	20	8	32	1	85	5	0	97	99

Table 5.3: Results for fake photon numbers of the current analysis of 2015-16 data with statistical (stat.) and systematic (syst.) errors.