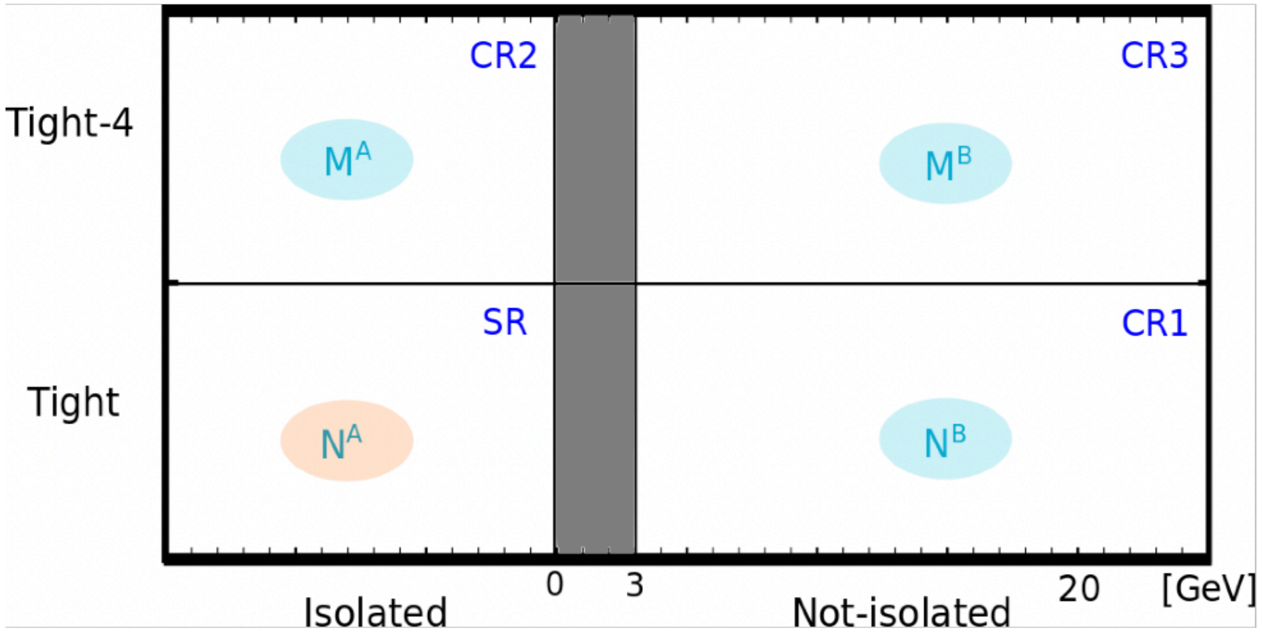
Measurement of the jets faking photons background for the Dark Matter search in the mono photon channel with the ATLAS detector

The 2D sideband method

In the 2D sideband method we consider candidates in a 2D plane (x,y) formed by a photon identification variable x and an isolation variable y.



- Identification:
  - Photons that pass the Tight selection are considered as identified
  - Photons that pass the Tight-4 selection but not the Tight one are considered as not-identified
  - Tight-3 and Tight-5 selections are used to estimate the systematic uncertainty

- Isolation:
  - Track isolation released in the control regions to improve statistics
  - E_{iso} = TopoEtCone40 - 0.022pt - 2.45 GeV
  - track_isolation = pt_{cone20}/pt

- Isolated photon:

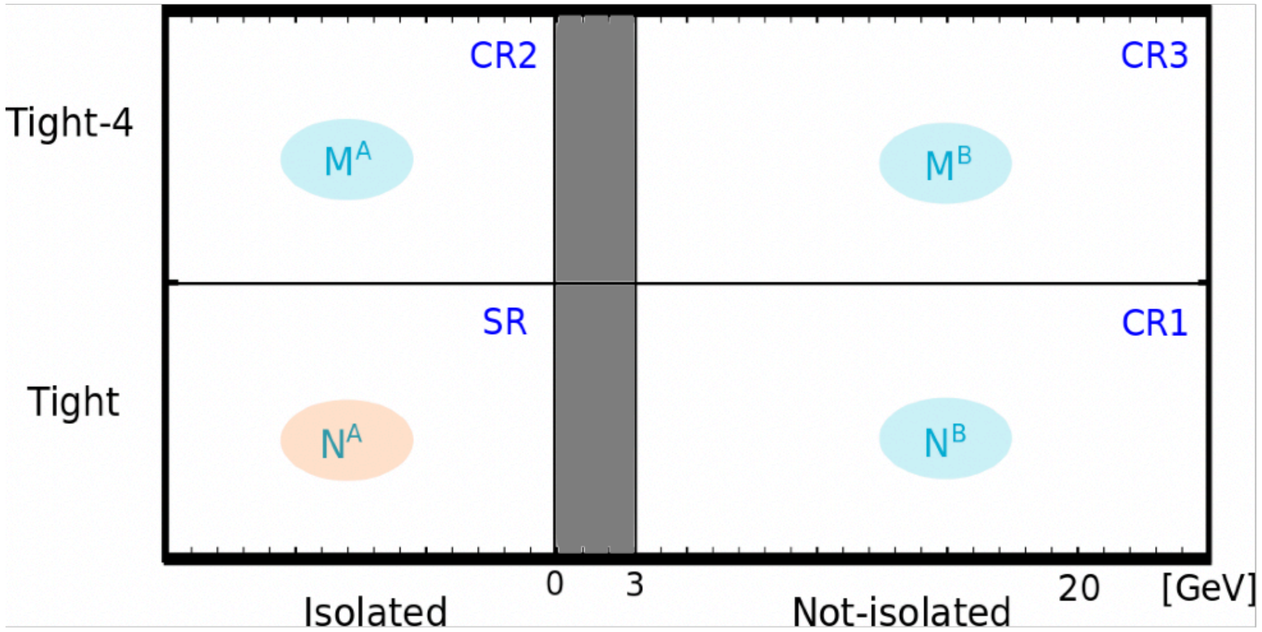
  E_{iso} < 0 GeV AND track_iso < 0.05

- Not-isolated photon:

  E_{iso} > 3 GeV OR track_iso > 0.10

The 2D sideband method

- A gap of 3 GeV in E_{iso} and of 0.05 in track_isolation has been excluded to prevent signal leakage.
- To exclude pathological events we set an upper limit to E_{iso} and track_isolation:
  - E_{iso} < 140 GeV
  - track_isolation < 1



- The plane is so divided in 4 regions. The number of photons in each region is:
  - N^A : tight isolated (signal region)
  - N^B : tight not-isolated (control region 1)
  - M^A : not-tight isolated (control region 2)
  - M^B : not-tight not-isolated (control region 3)

Correlation factor

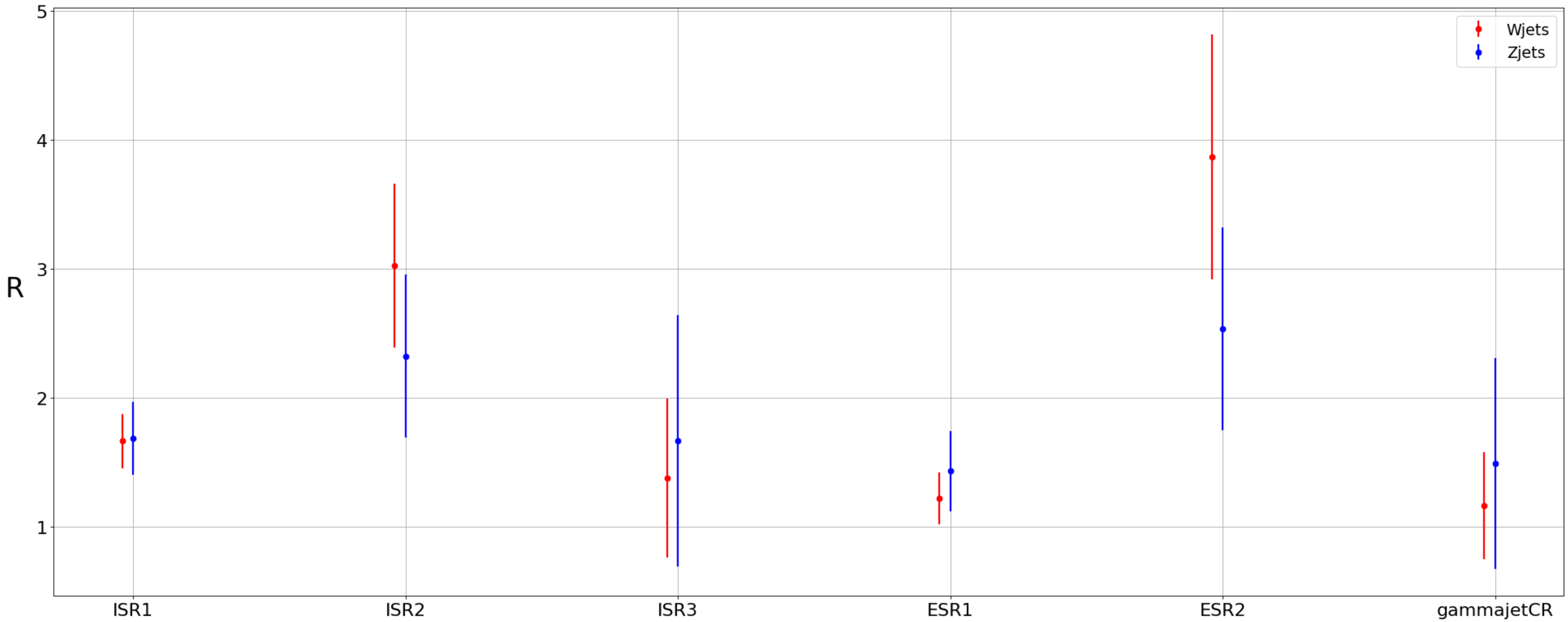
The correlation factors, one for each region of the analysis, have been calculated separately on two samples of Z + jets and W + jets to check eventual process dependence.

Very strange results because of Monte Carlo weights. We managed this weights as follow:

- If |MC weight| > 100, it is rescaled to 1
- If an event has pt < 140 GeV at truth level we excluded it
  - we already select photons with pt > 150 GeV, but some of them may have been reconstructed with higher pt, possibly gaining a very high cross section
- Merged SR and relevant CRs to increase statistics
  - W + jets $\longrightarrow$ SR + 1muCR
  - Z + jets $\longrightarrow$ SR + 2muCR + 2eCR
  - gammajetCR treated separately

Correlation factor

Results in the different regions are the following:

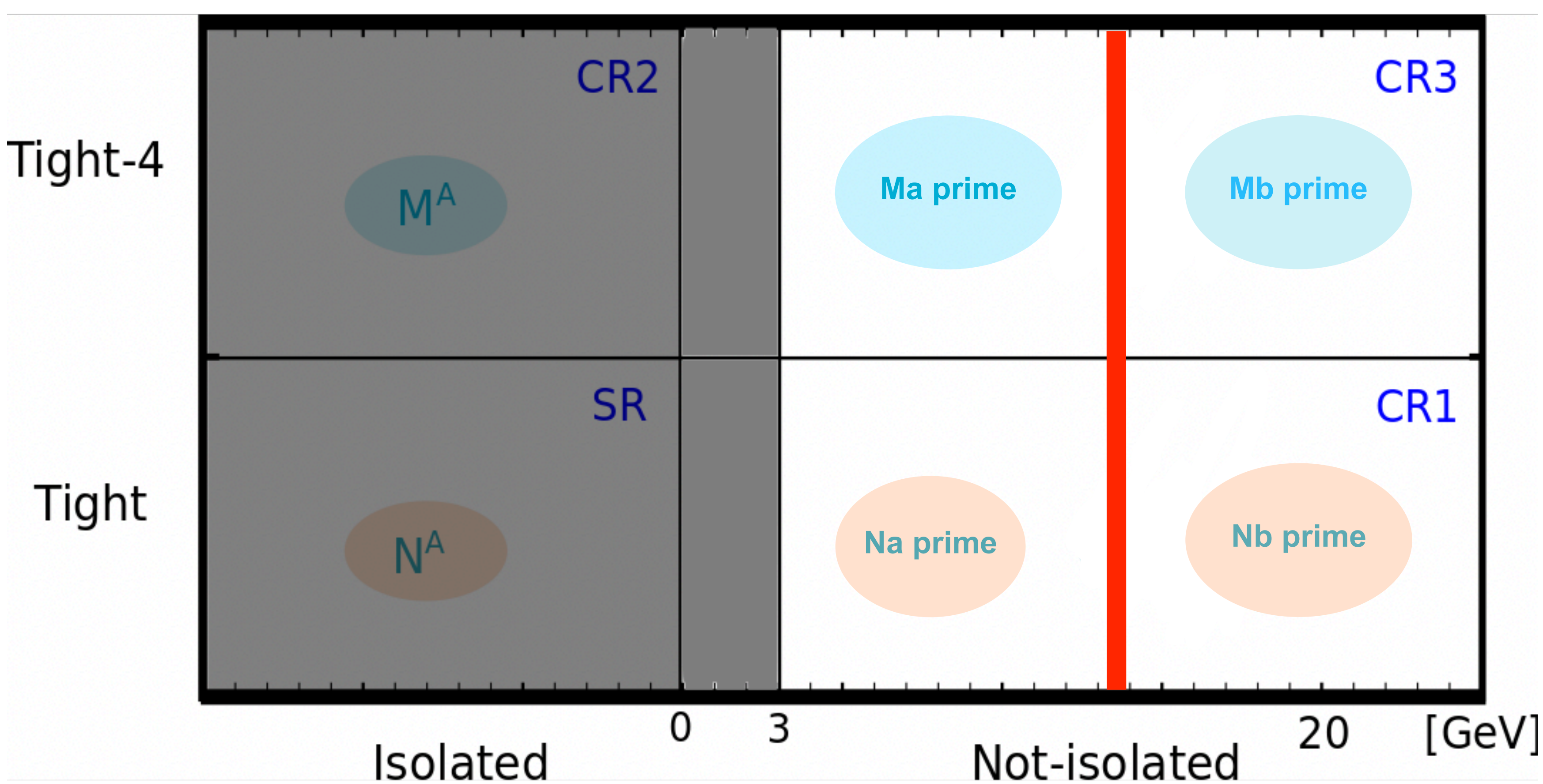


- Good compatibility between W + jets and Z + jets

R prime

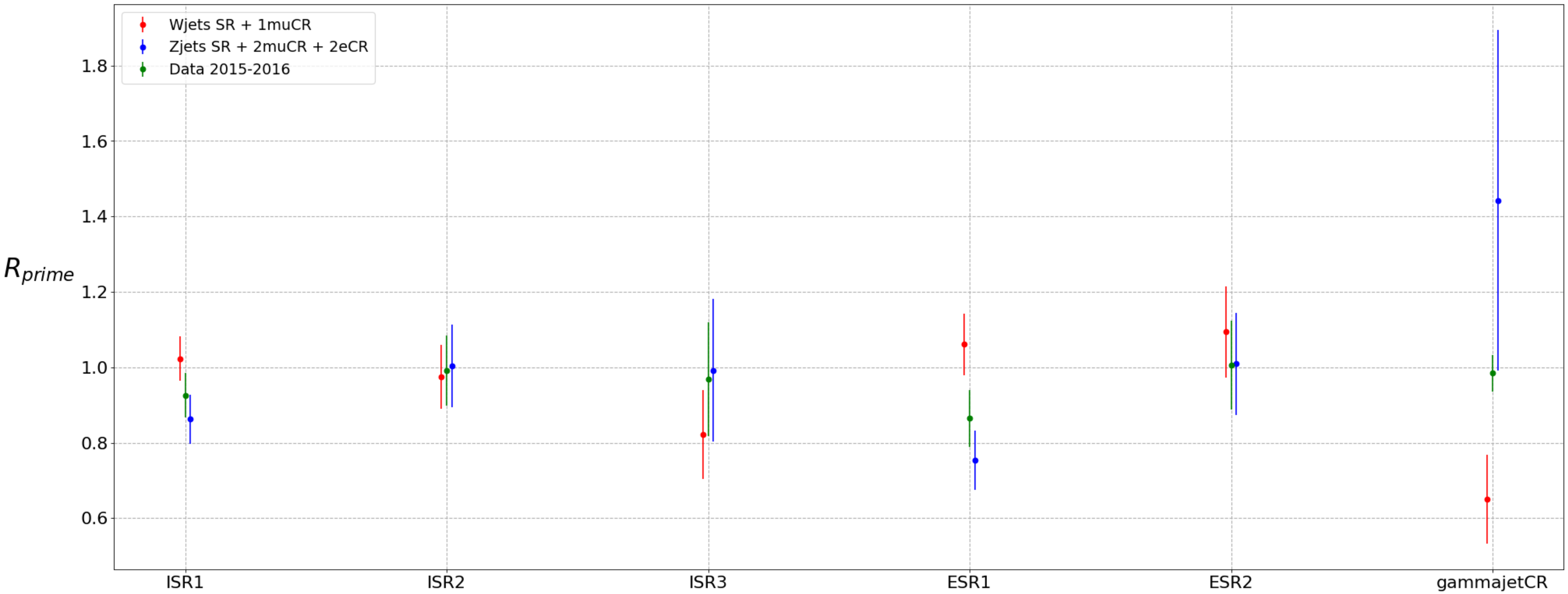
To have a feeling of the robustness the predictions on R we calculated a correlation factor in a completely not-isolated region, where no signal is expected, both for Monte Carlo samples and for 2015-2016 data. The new number of candidates are as follow:

- N^{A}_{prime}: (3 GeV < E_{iso} < 50 GeV AND track_iso < 0.2) OR (E_{iso} < 3 GeV AND 0.1 < track_iso < 0.2)
- N^{B}_{prime}: (50 GeV < E_{iso} < 100 GeV AND track_iso<1.0) OR (E_{iso} < 50 GeV AND 0.2 < track_iso < 1.0)
- Same for M^{A}_{prime}, M^{B}_{prime} but in the Tight-4 region



R prime

Results in the different regions are the following:



- Fair compatibility between MCs and 2015-2016 data

Signal Leakage

The signal leakage coefficients have been calculated separately on three samples of W + γ, Z + γ and Z (vv) + γ, selecting from this samples only reconstructed photons matched with true photons.

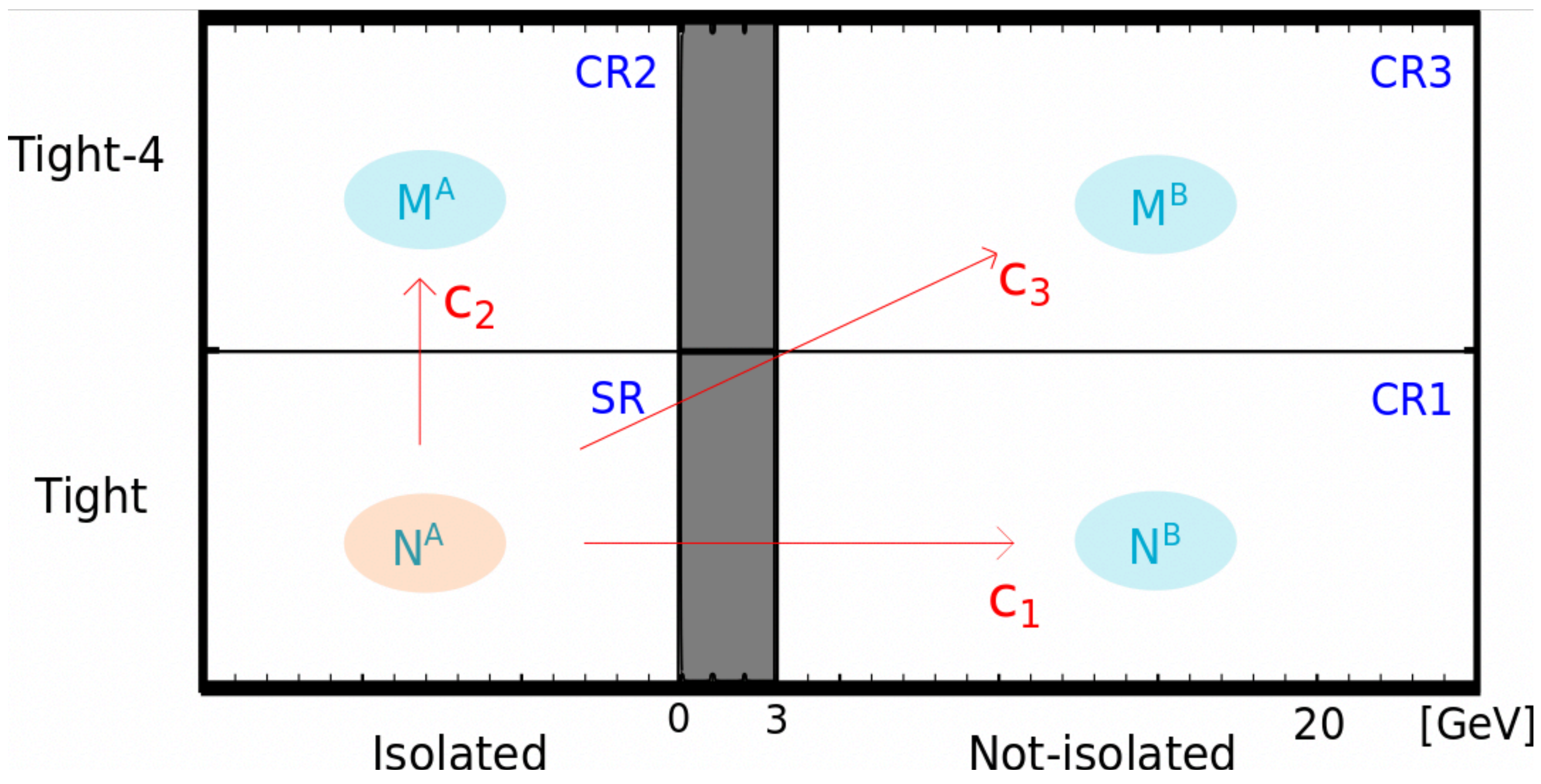
As we did for the correlation factors we merged the regions in which a certain process is dominant. The gammajetCR is treated separately.

- W + γ $\longrightarrow$ SR + 1muCR
- Z + γ $\longrightarrow$ SR + 2muCR + 2eCR
- Z(vv) + γ $\longrightarrow$ SR

Also we expect more leakage with respect to the previous analysis because of three factors:

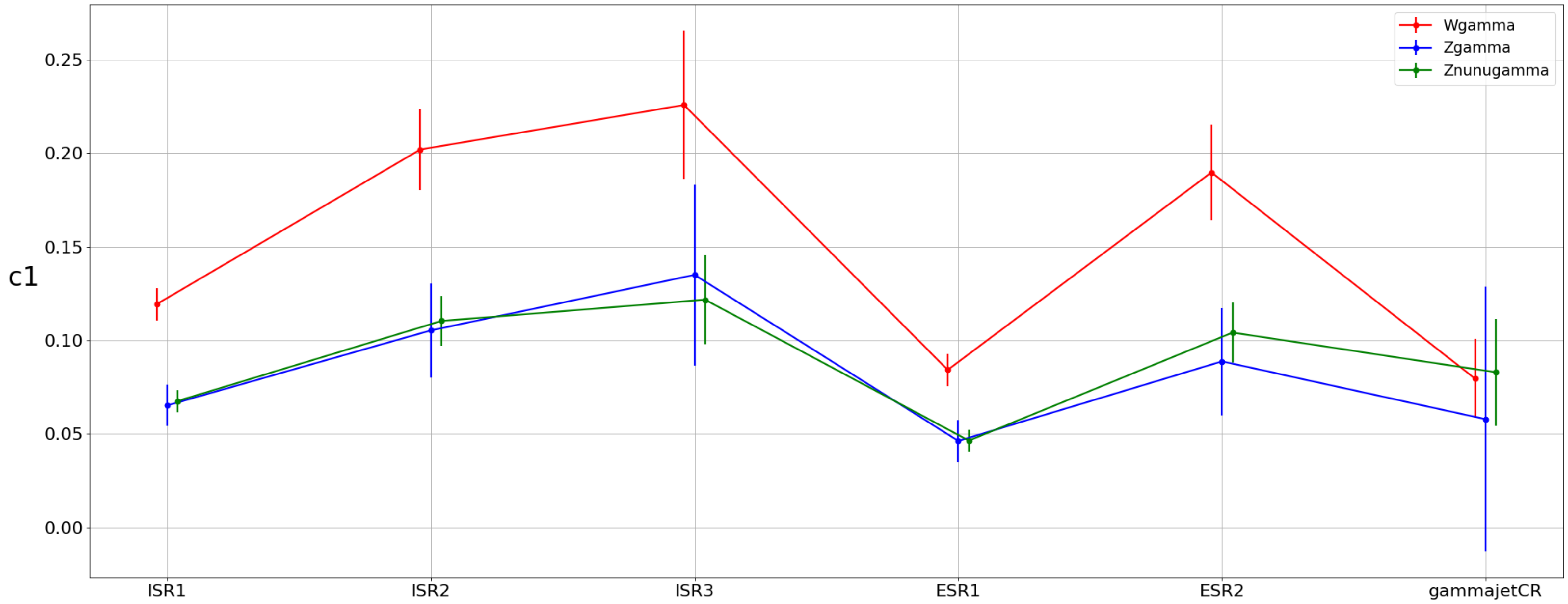
- Track isolation released in the CRs
- Different calorimetric isolation cut
- Different Monte Carlo generation, looser isolation cut at truth level

Signal Leakage



Signal Leakage c1

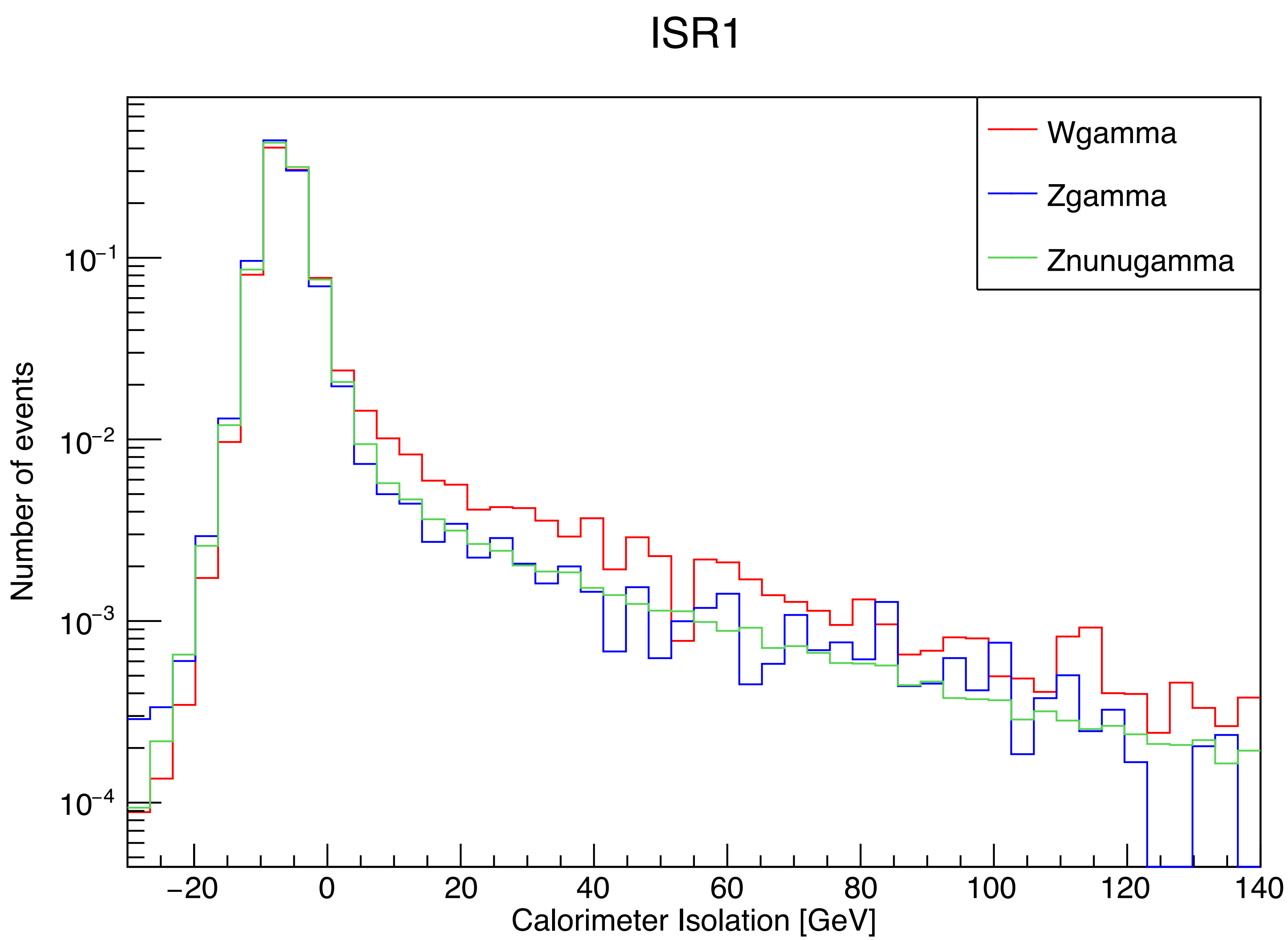
Results in the different regions for the first three samples (W + γ, Z + γ, Z (vv) + γ) are the following:



- Systematic difference between W + γ and Z + γ samples
- Rather clear trend with MET threshold

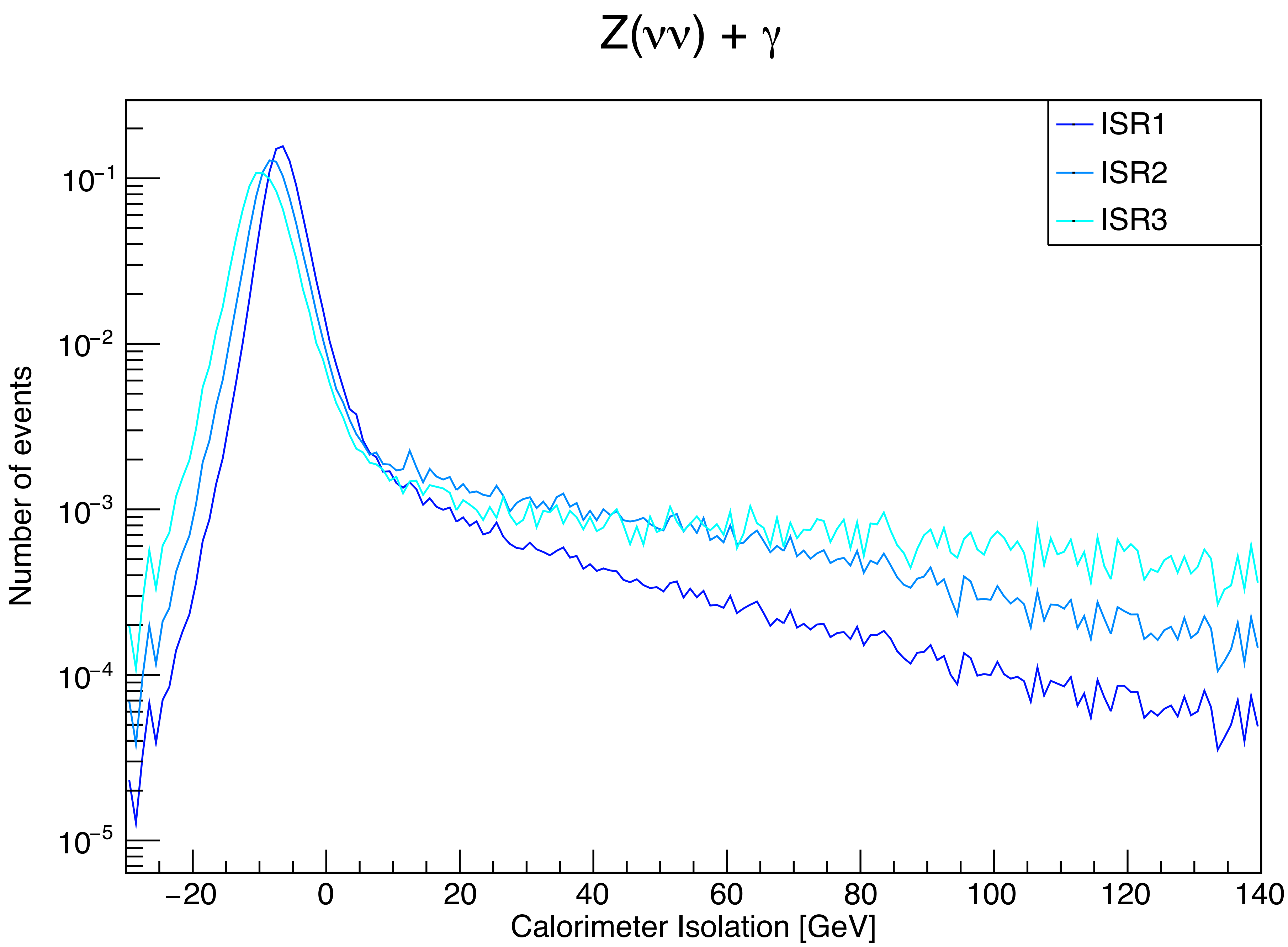
Signal Leakage

- Results for Z + γ and Z (νν) + γ compatible everywhere
- c1 and c3 systematically higher for W + γ
  - Different tails in the calorimeter isolation profile

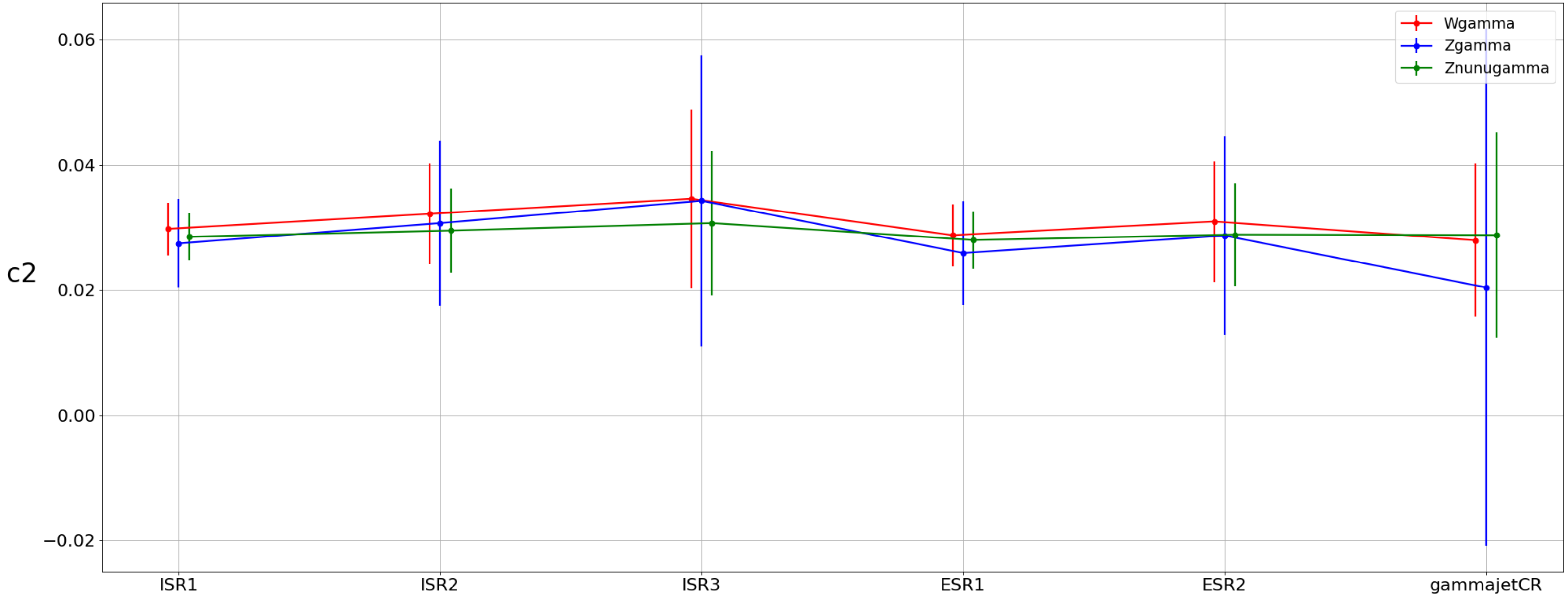


Signal Leakage

- MET dependence

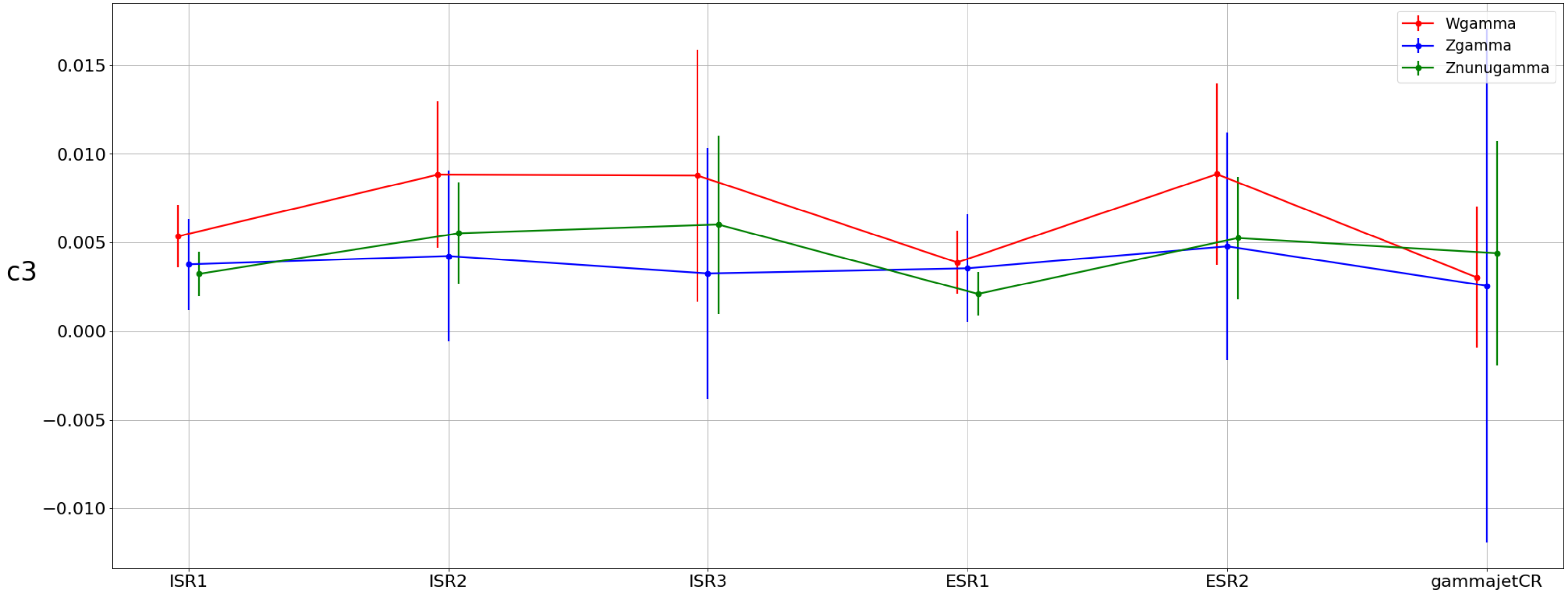


Signal Leakage c2



- Good compatibility everywhere

Signal Leakage c3



- Systematic difference between W + γ and Z + γ samples, but compatible results everywhere
- Rather clear trend with MET threshold

Comments

- Correlation factor R
  - Compatible results everywhere for W + jets and Z + jets samples
  - No evident systematics, only statistical uncertainty

- Signal leakage
  - MET dependence of c1 and c3, not expected
  - Systematic differences of c1 and c3 between W + γ and Z + γ samples, not fully understood, both systematic and statistical errors
  - c2 no evident systematics, only statistical uncertainty

Coefficients Analysis

All the coefficients (c1, c2, c3 and R) are treated the same way:

- After the merge each coefficient have 5 (SR) + 1 (gjCR) values for every sample
- In each region we take the weighted mean (nominal value and statistical uncertainty) of the 2 (R) or 3 (c1, c2, c3) values from different samples
- To account for systematics we also add a systematic error using the following formula (c1,c3 only!):

$$\sigma_{syst} = \sqrt{\frac{1}{N(N-1)} \sum (coeff_i - mean)^2}$$

We then obtain 5 + 1 valued for each coefficient with statistical and a systematic errors:

- Each value is used in a specific SR and in all the corresponding CRs
- The last one is used only in the gammajetCR

We then have 21 coefficients with statistical and a systematic errors

Purity Analysis

We can then compute the purities for the 2015-16 data.

Statistical uncertainties:

- Complete formula and derivatives
- Statistical errors propagated ONLY for Na, Nb, Ma, Mb: statistics of data and MC are different
- Separately we propagate the error on each coefficient

Systematics on the purities from the coefficients, for each coefficient:

- In each region we calculate the nominal purity from the nominal coefficient
- We calculate:

  coeff_up = coeff + σ_{syst}

  coeff_down = coeff - σ_{syst}

- Compute the 2 purities
- Take the differences with the nominal purity
- Quote the error as the maximum of these two differences

Purity Analysis

Systematics on the purities from tightness and isolation are calculated as for the coefficients, moving the not-tight selection or the isolation parameters, recalculating all the coefficients and purities and taking the maximum difference from the nominal purity as systematic error.

We move tightness and isolation as follow:

- Tightness:
  - Tight-4 $\longrightarrow$ Tight-3
  - Tight-4 $\longrightarrow$ Tight-5
- Isolation, gap between the isolated and not-isolated region (E_{iso}, track isolation):
  - (3 GeV, 0.05) $\longrightarrow$ (2 GeV, 0.01)
  - (3 GeV, 0.05) $\longrightarrow$ (4 GeV, 0.10)

Purity Analysis

In all the regions the main sources of error are:

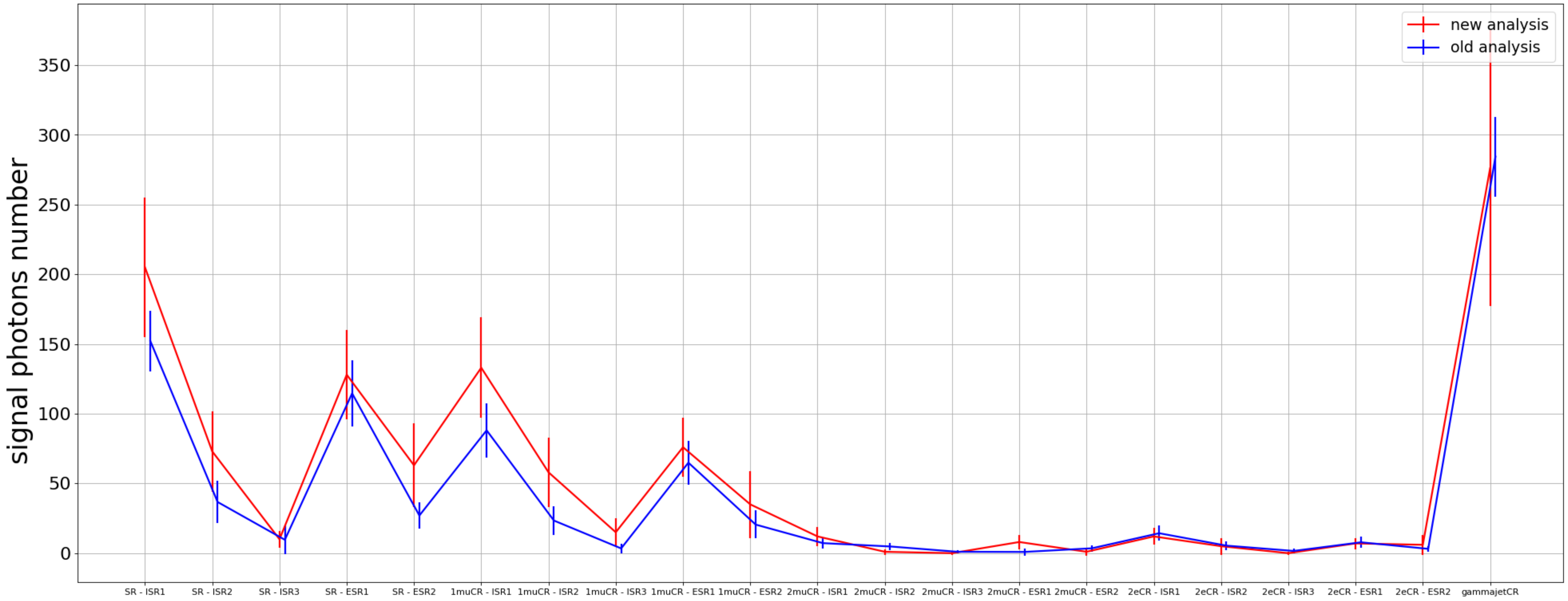
- Tightness systematics
- Correlation factor R, statistical uncertainty

The total systematic error is the square sum of all the systematics discussed and the statistical uncertainties on coefficients.

The total error is computed as the square sum of the total systematics and the statistical uncertainty of data.

Purity Analysis

Results are compatible with the previous analysis everywhere within the total error.



Purity Analysis

Number of signal photons in each region of the analysis.

	SR ISR1	SR ISR2	SR ISR3	SR ESR1	SR ESR2	1mu CR ISR1	1mu CR ISR2	1mu CR ISR3	1mu CR ESR1	1mu CR ESR2	2mu CR ISR1	2mu CR ISR2	2mu CR ISR3	2mu CR ESR1	2mu CR ESR2	2e CR ISR1	2e CR ISR2	2e CR ISR3	2e CR ESR1	2e CR ESR2	gjCR
Number of signal ph	205	73	10	128	63	133	58	15	76	35	12	1	0	8	1	12	5	0	7	6	276
Statistical uncertainty	21	15	4	15	16	15	12	5	10	9	4	2	0	3	3	5	4	1	3	5	18
Tightness systematics	36	19	2	20	20	28	19	6	14	21	5	1	0	3	1	3	4	0	3	5	26
Isolation systematics	3	2	0	5	3	5	3	1	2	1	0	0	0	1	0	1	0	0	0	1	20
c1 statistical	3	2	1	2	2	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	8
c2 statistical	8	6	2	6	6	3	2	1	2	2	1	1	0	0	1	0	0	0	0	0	32
c3 statistical	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
R statistical	22	13	4	17	14	14	10	6	10	7	1	0	0	1	0	1	1	0	1	1	85
c1 systematics	13	6	1	7	5	4	2	0	2	2	1	0	0	1	0	1	0	0	1	0	5
c3 systematics	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total systematics	45	25	5	29	26	32	22	8	18	22	6	1	0	4	1	4	4	0	3	5	97
Total error	50	29	6	32	30	36	25	10	21	24	7	2	1	5	3	6	6	1	4	7	99

Conclusions

- Samples: problems with negative weights
  - Track isolation released in the CRs
  - Merging of the CRs dominated by the same boson decay
  - γ + jets (Sherpa LO and NLO) samples tried but excluded, too strange behavior
    - Consider to try a Phythia sample
- Coefficients: systematic differences for c1 and c3
  - Not fully understood
  - Weighted mean with a systematic error (currently under evaluation, probably conservative)
- Purities: calculated for 2015-2016 data
  - Compatible with the old analysis within the total error
  - Much grater errors, mainly from tightness and R (currently under evaluation)
- Few details are under evaluation, almost everything ready for the analysis of 2017 and 2018 data.

Backup

The 2D sideband method

We start with two simplifying hypothesis:

1. The correlation between x and y is negligible for the background
2. The number of signal candidates is negligible compared to the number of fake candidates is negligible in the control regions:

$$N_{\rm bkg}^B \gg N_{\rm sig}^B$$
$$M_{\rm bkg}^A \gg M_{\rm sig}^A$$
$$M_{\rm bkg}^B \gg M_{\rm sig}^B$$

As a consequence of the first assumption we can assume that:

$$N_{\rm bkg}^A / N_{\rm bkg}^B = M_{\rm bkg}^A / M_{\rm bkg}^B$$

and as a consequence of the second assumption:

$$N^B = N_{\rm bkg}^B$$
$$M^A = M_{\rm bkg}^A$$
$$M^B = M_{\rm bkg}^B$$

The 2D sideband method

Therefore combining the two hypothesis:

$$N^A_{\rm bkg} = N^B_{\rm bkg} \times M^A_{\rm bkg}/M^B_{\rm bkg} = N^B \times M^A/M^B$$

which provides a fully data-driven technique to estimate the background.

The signal in region A is therefore:

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

so that the purity is:

$$P = N^A_{\rm sig}/N^A = 1 - \frac{N^B}{N^A}\frac{M^A}{M^B}$$

which is valid only if the two assumptions are satisfied.

The 2D sideband method

If one or both of the two hypothesis are not satisfied we can correct the formula for the purity using Monte Carlo

1. Non negligible correlation in the background between x and y

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

2. Signal leakage in background control regions

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

where

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

The 2D sideband method

Substituting we obtain:

$$N_{\mathrm{sig}}^{A} \quad = \quad \frac{(M^B + N^A c_3 - N^B c_2 R_{\mathrm{MC}} - M^A c_1 R_{\mathrm{MC}})(-1 + \sqrt{1 + \frac{4(c_1 c_2 R_{\mathrm{MC}} - c_3)(N^A M^B - N^B M^A R_{\mathrm{MC}})}{(M^B + N^A c_3 - N^B c_2 R_{\mathrm{MC}} - M^A c_1 R_{\mathrm{MC}})^2}}}{2(c_1 c_2 R_{\mathrm{MC}} - c_3)}$$

and the purity straightforward.

We can also approximate this formula if:

$$\left| \frac{4(c_1 c_2 R_{\mathrm{MC}} - c_3)(N^A M^B - N^B M^A R_{\mathrm{MC}})}{(M^B + N^A c_3 - N^B c_2 R_{\mathrm{MC}} - M^A c_1 R_{\mathrm{MC}})^2} \right| \ll 1$$

so that:

$$N_{\mathrm{sig}}^{A} \quad = \quad \left(N^A - N^B \frac{M^A}{M^B} R_{\mathrm{MC}} \right) \frac{1}{1 + \frac{c_3 N^A - c_2 N^B R_{\mathrm{MC}} - c_1 M^A R_{\mathrm{MC}}}{M^B}}$$

more reliable when we compute the derivatives for the propagation of uncertainty.

The systematic uncertainties are evaluated moving the calorimeter isolation threshold from 3 to 2 and 4, and the not-tight region from Tight-4 to Tight-3 and Tight-5, than taking the maximum difference from the original purity. The total systematic uncertainty is computed as the square sum of the two uncertainties.

Control Regions

The Signal Region (SR), is divided in 5 regions with different MET cuts:

	ISR1	ISR2	ISR3	ESR1	ESR2
MET [GeV]	> 150	> 225	> 300	150 - 225	225 -300

For each of this regions we define 3 control regions, each enriched with different processes:

- one muon CR (1muCR) $\longrightarrow$ W (μν) + γ
- two muon CR (2muCR) $\longrightarrow$ Z (μμ) + γ
- two electron CR (2muCR) $\longrightarrow$ Z (ee) + γ

We define one more control region, the gammajetCR, dominated by jet + γ process. This region can't be divided as the other because it has a different MET cut (85 GeV < E_T^{miss} < 110 GeV).

Validation

To validate the method we used a mixed Monte Carlo sample of W(μν) + γ and W + jets, with known purity:

- The correlation factor (R_{MC}) and the signal leakage coefficients (c1, c2, c3) are calculated on the same sample
- Focus on the ISR1 of the SR
- Two different calculi of the purity, taking and not taking into account Monte Carlo weights

In this region the coefficients are:

	R_{MC}	c_1	c_2	c_3
Not Weighted	1,60 ± 0,24	6,98 ± 0,42 %	2,94 ± 0,27 %	0,31 ± 0,09 %
Weighted	2,76 ± 0,49	7,08 ± 0,80 %	4,50 ± 0,63 %	0,49 ± 0,21 %

At first the not weighted purity wasn't compatible with the result we expected, the reason is that the approximation made is not satisfied. We then used the complete formula calculating all the derivatives for the propagation of the uncertainty with Mathematica.

The results with propagated and systematic uncertainties respectively are:

	Not weighted purity	Weighted purity
Expected	96,96%	84,27%
Calculated	97,05 ± 0,91 ± 1,56 %	86,41 ± 5,87 ± 7,26%

Compatible with the expected purities.