Multivariate energy regression

G. Cavoto, E. Di Marco, D. Pinci

CYGNO reco and analysis meeting,
14 October 2021

LIME overground data

Energy resolution of clusters from electron recoils was estimated with ^{55}Fe in LEMON and LIME:

- about 15% (20%) in LEMON (LIME) at E = 5.9 keV

Dominating sources implemented in the SIM (gain fluctuations, diffusion in the gas [Z-dependent]) implemented in the SIM

Other sources can be variations of response with Z (saturation, diffusion, etc.), with X-Y (field non-uniformities), etc.

Obvious energy scale correction from optical vignetting already corrected with a X-Y map obtained from white-wall picture

Other sources not completely clear, and factorized corrections can be not the most obvious strategy

Ultimate energy resolution not crucial, but crucial having the right energy scale (i.e. can correct for saturation?)

Multivariate regression

- General principle is to derive a best estimate of the dependent variable (in this case the true cluster energy) given a set of independent variables (position, cluster shape parameters, etc)

- Davide's empiric correction was an energy correction using the projection of the energy scale onto 1 variable (density δ)

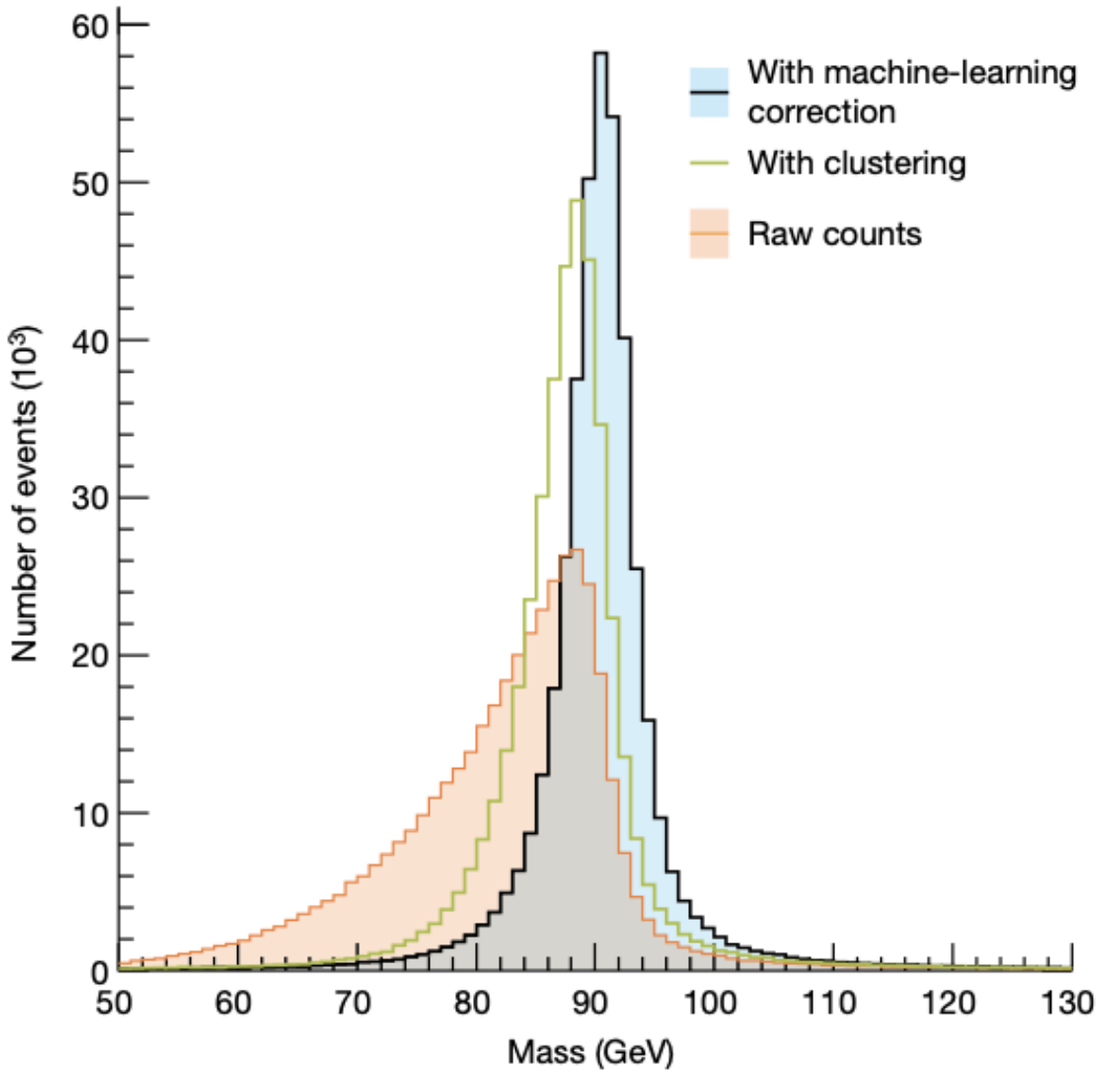
- In an event classification problem this is like using the projected likelihood in several variables (which is fully optimal as long as the correlations between variables are not relevant)

- In a classification problem one can use a multidimensional probability density, Boosted Decision Tree, or Neural Net to take into account the correlations

- We can do the same for multivariate regression

Used in CMS-ECAL

We used it extensively to correct the energy response of the ECAL in CMS wrt many effects (local containment, pileup dependency, etc)

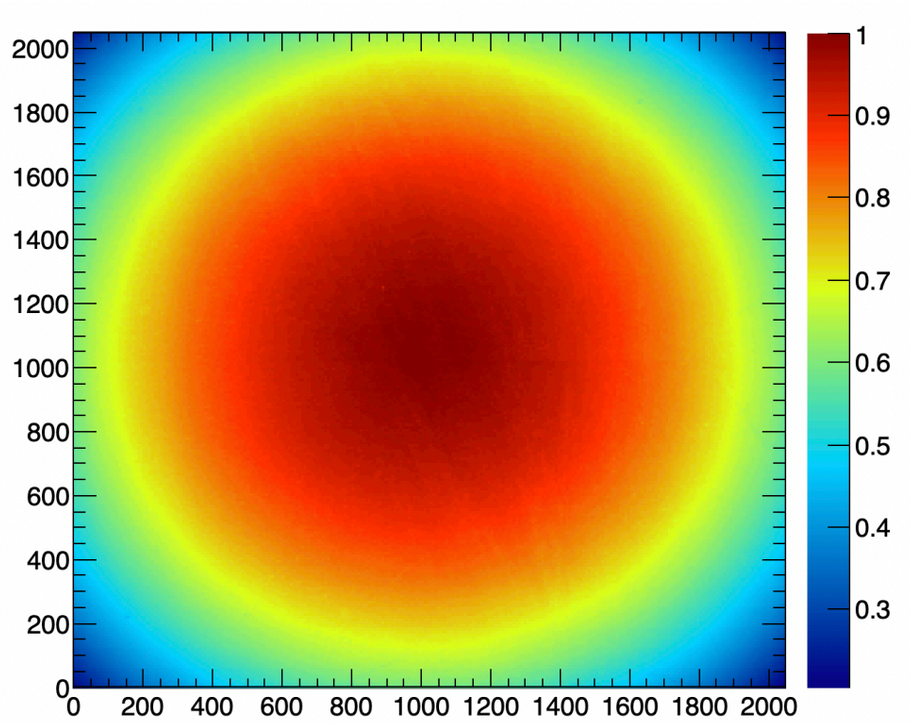


$Z \to e^+e^-$ invariant mass

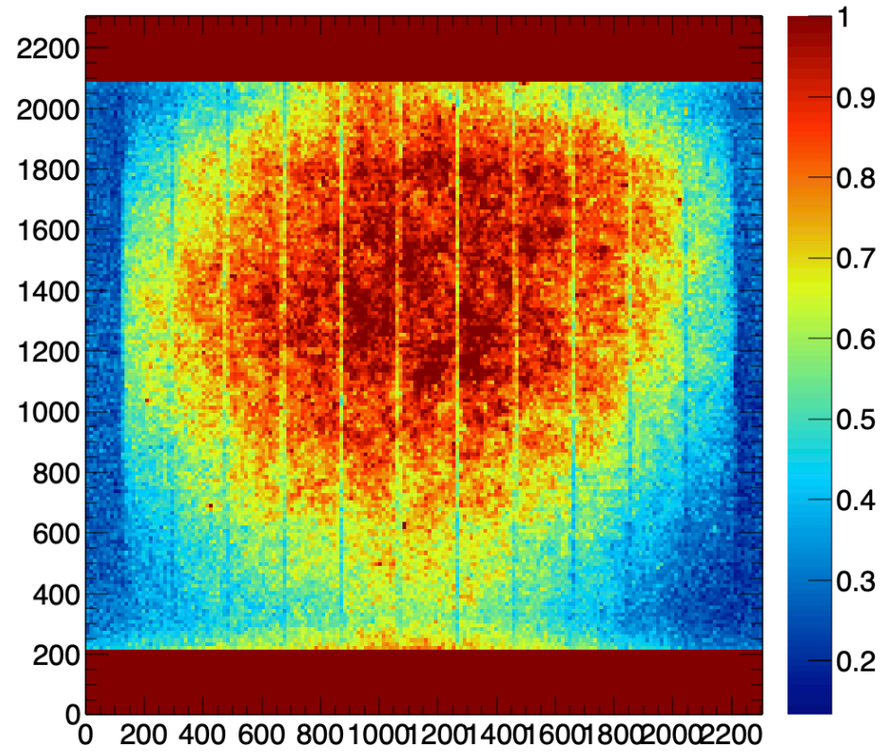
[1] 10.1088/1748-0221/10/06/P06005
[2] 10.1088/1748-0221/10/08/P08010

Inputs

Residual response non-uniformity in x-y



We use this to correct (leading-order effect)



But we see this (i.e. there are other non-uniformities)

- x, y of the cluster baricenter, cluster $x(y)_{max}$, $x(y)_{min}$
- Cluster shapes. Try to correct for diffusion-related (i.e. Z-dependence):
  - slimness, $RMS_{\perp}$, $RMS_{\parallel}$, Gaussian $\sigma_{\perp}$, $\sigma_{\parallel}$, width, length

Selection and sample

Ideally this should be done on SIM. Target would be E_{true}/E_{reco} (or its full PDF, not just an estimator of it, as its mean)

PRO of the SIM: Can be trained on both ERs and NRs (our signals) of whatever energy / condition / prototype

CON of the SIM: Sensitive to data-SIM disagreement of ANY of the regression inputs. At this stage we haven't a reliable, extensive, data-SIM comparison in LIME

- keep in mind for the next future (needs a comparison of ALL the variables)

So right now train on DATA, ^{55}Fe, for which we have a sample with high statistics and high S/B ratio. Target is the known energy (5.9 keV, in raw pixel counts), normalized to the peak position

- Selection: length<100 pix ; width/length>0.6 ; 0.3<integral/9000<1.7 (cut away fake clusters and merged spots), R<900 pixels (avoid highly vignetted region)

Implementation

- Input variables are used to train a multivariate regression using the Gradient Boost Regression (based on a BDT in scikit-learn).

- GBR target is integral/9000 (to have a variable centered at 1)

    - normalization also helps in reducing the phase space of the target variable when training with variable energy clusters

- The loss function are:

    - mean squared errors

    - 50% quantile (median), and 5% and 95% quantiles

        - 50% quantile gives the central prediction, the other two give per-cluster energy resolution estimates (+ and - asymmetric errors)

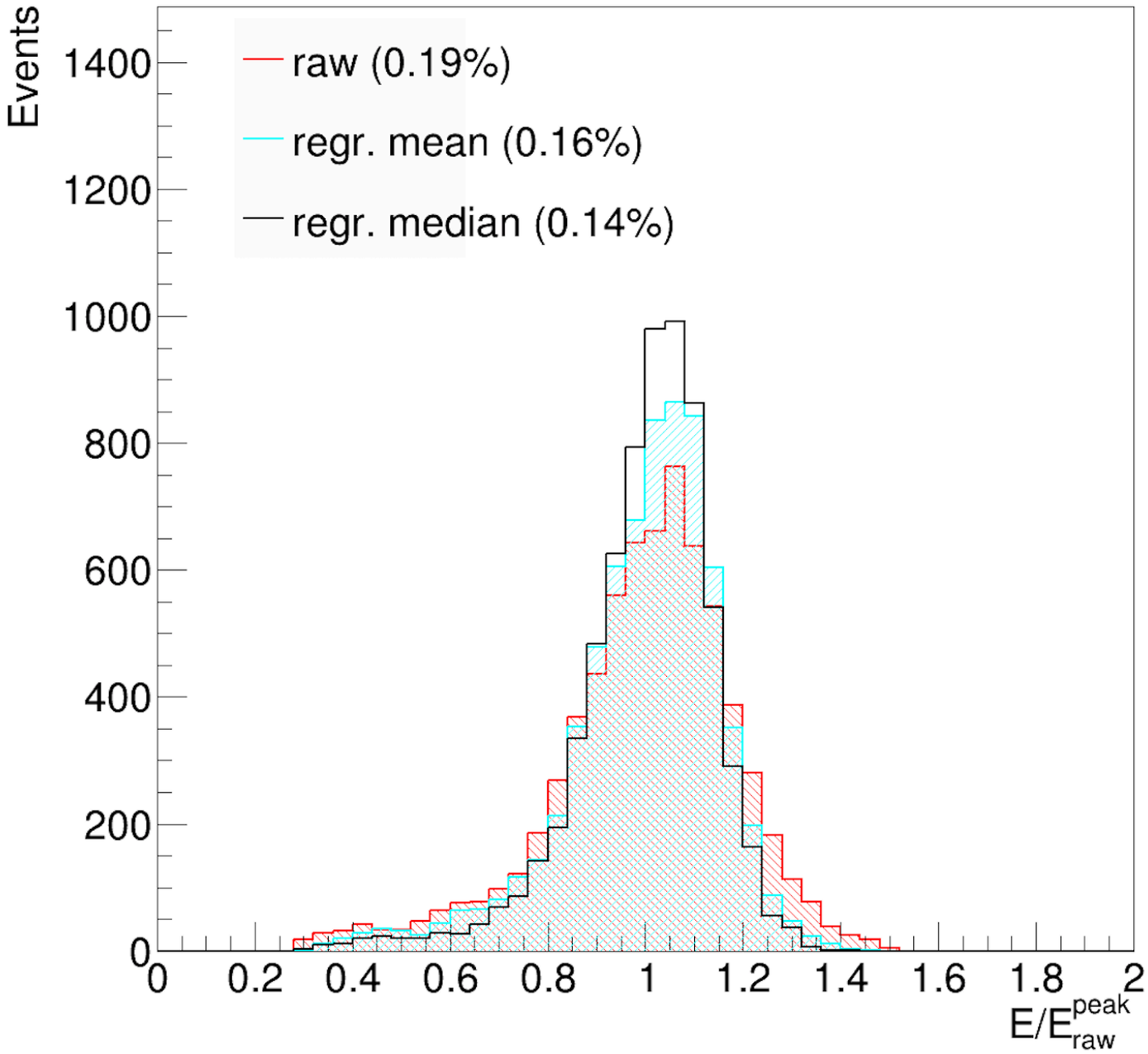
- Detailed training options to be further optimized

Data used

July 30th runs with ^{55}Fe with V_{GEM1} = 440 V at different Z values: 46, 36, 26, 6cm

- here focusing on Energy, but a dedicated regression can target Z (exploit dependency of cluster shapes - through diffusion, mainly) to give a per-cluster Z estimate
  - a straightforward way to make a 3D reconstruction. Need more Z points to test this (e.g. data taken in April and analyzed by Donatella)
- About 8000 clusters used, 90% for training, 10% for testing
- A thought for the future:
  - data with the source moved on all 4 sides of LIME would help covering uniformly the XY plane with spots

Inclusive Results



Regression gives significant improvement to the energy resolution

- 10% for mean regression and 14% for quantile regression in quadrature

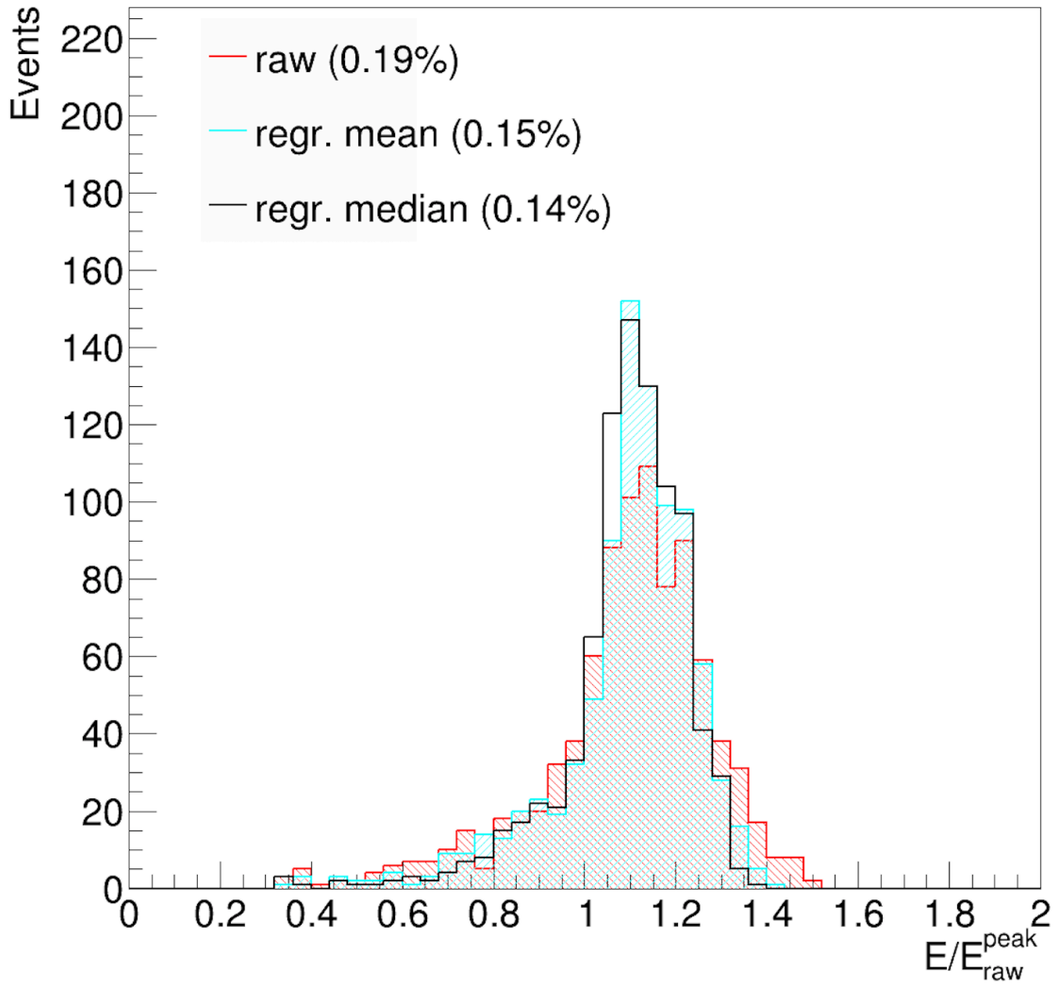
It is necessary to check the robustness vs E

- check with SIM
- check with multi-E X-ray source data

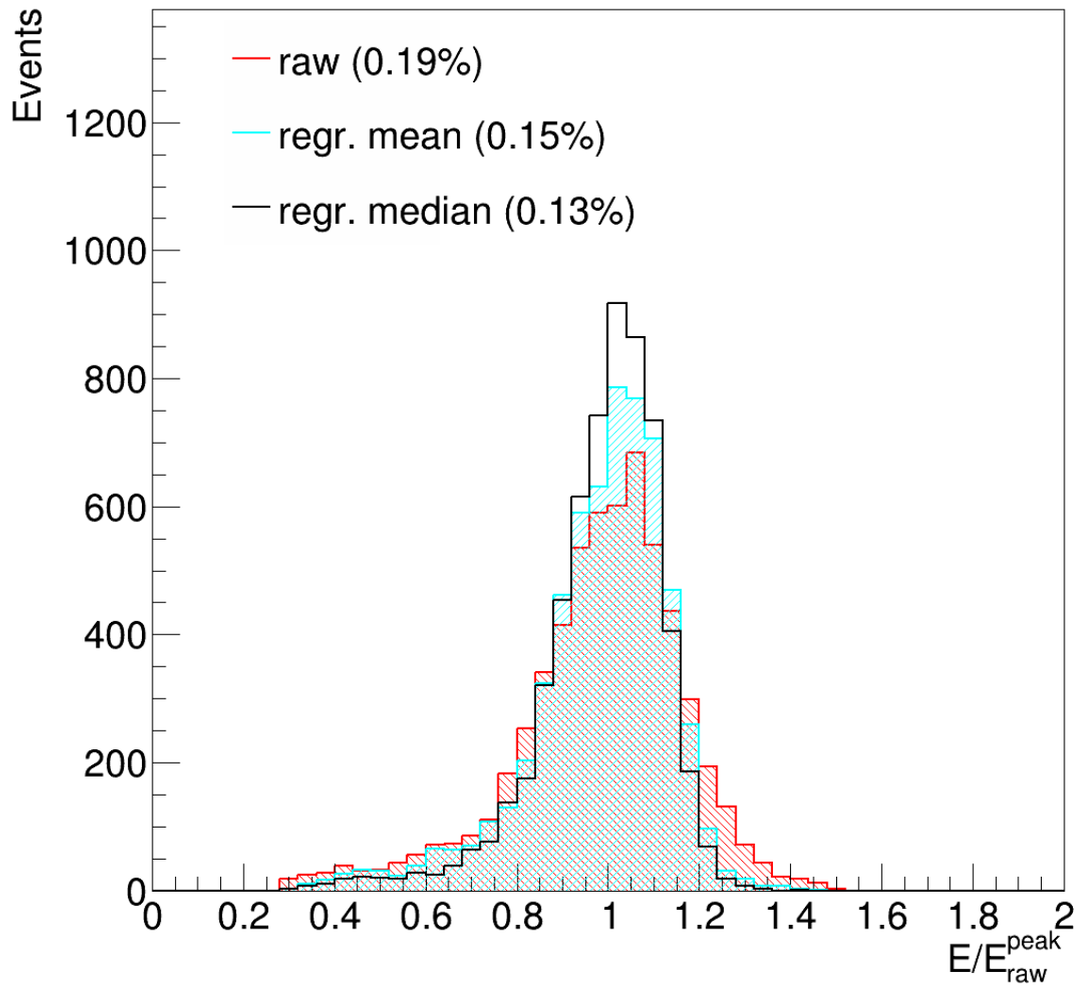
Results vs R

It works similarly in the very center and in the outer ring.

So it is not just correcting for the residual non-uniformity of the drift field

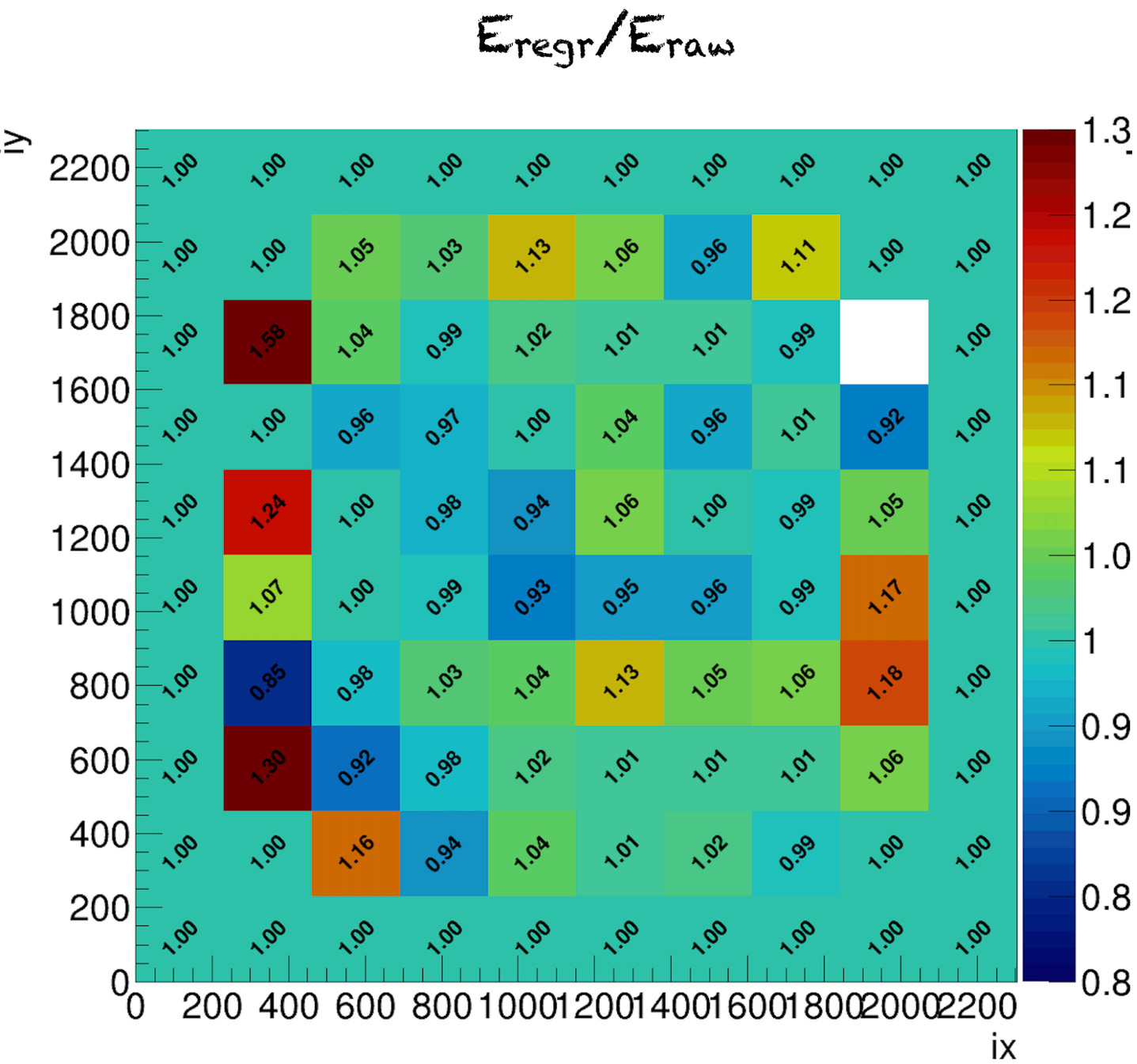


R < 300 pixels



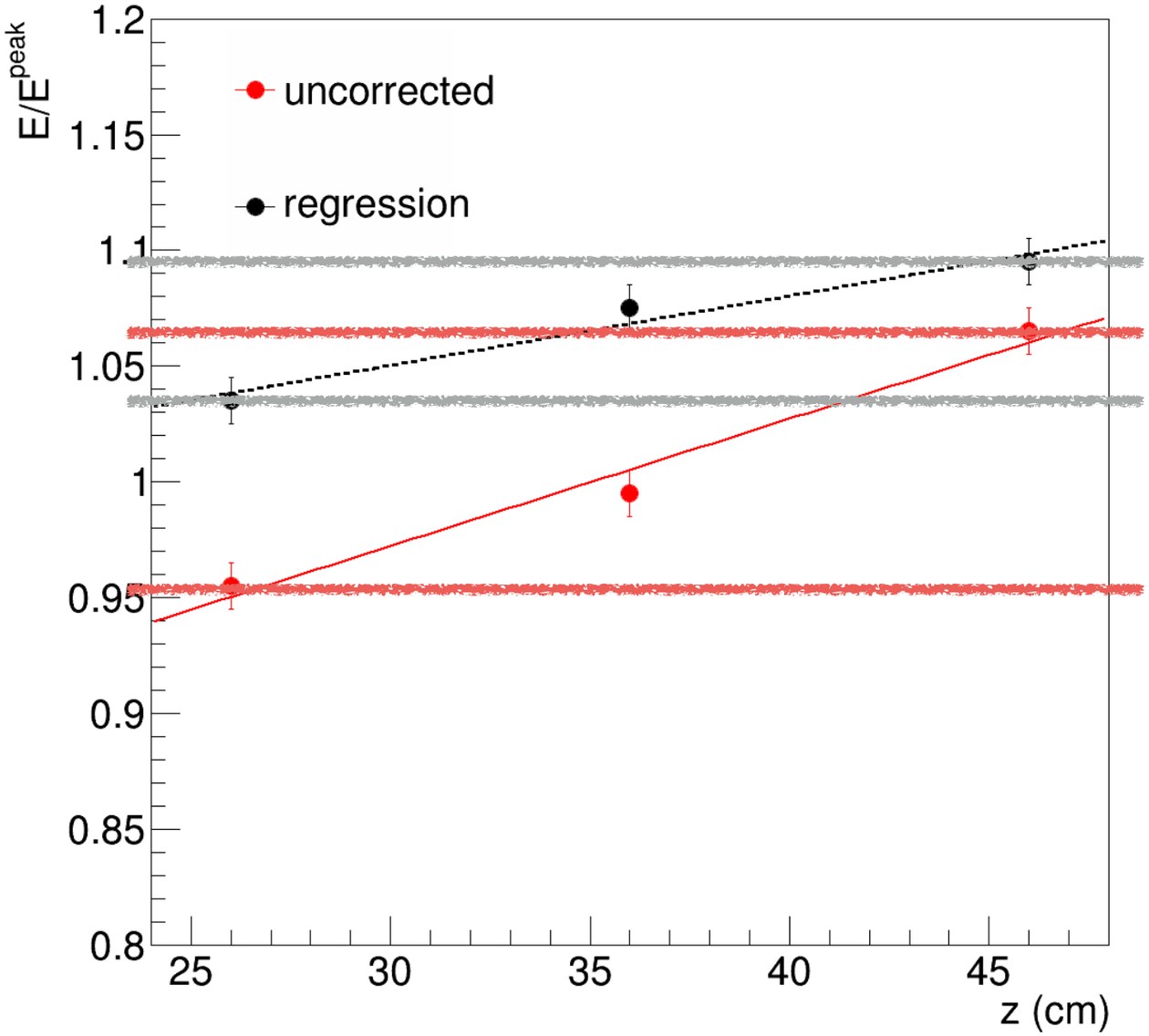
R > 300 pixels

Correction: x-y dependence



- It is enhancing the bottom part of the sensor (by ~[1-6]%)
  - we guessed from the cosmics - integrated map
- It is correcting residual vignetting at high R
- To be checked with statistical error and per-cluster uncertainty
- Closure test: E_{regr} should ideally be flat at 1, but not closing perfectly. Checking why...

Saturation correction?



Energy response using RAW energy estimate varies 11% from 25 to 45cm

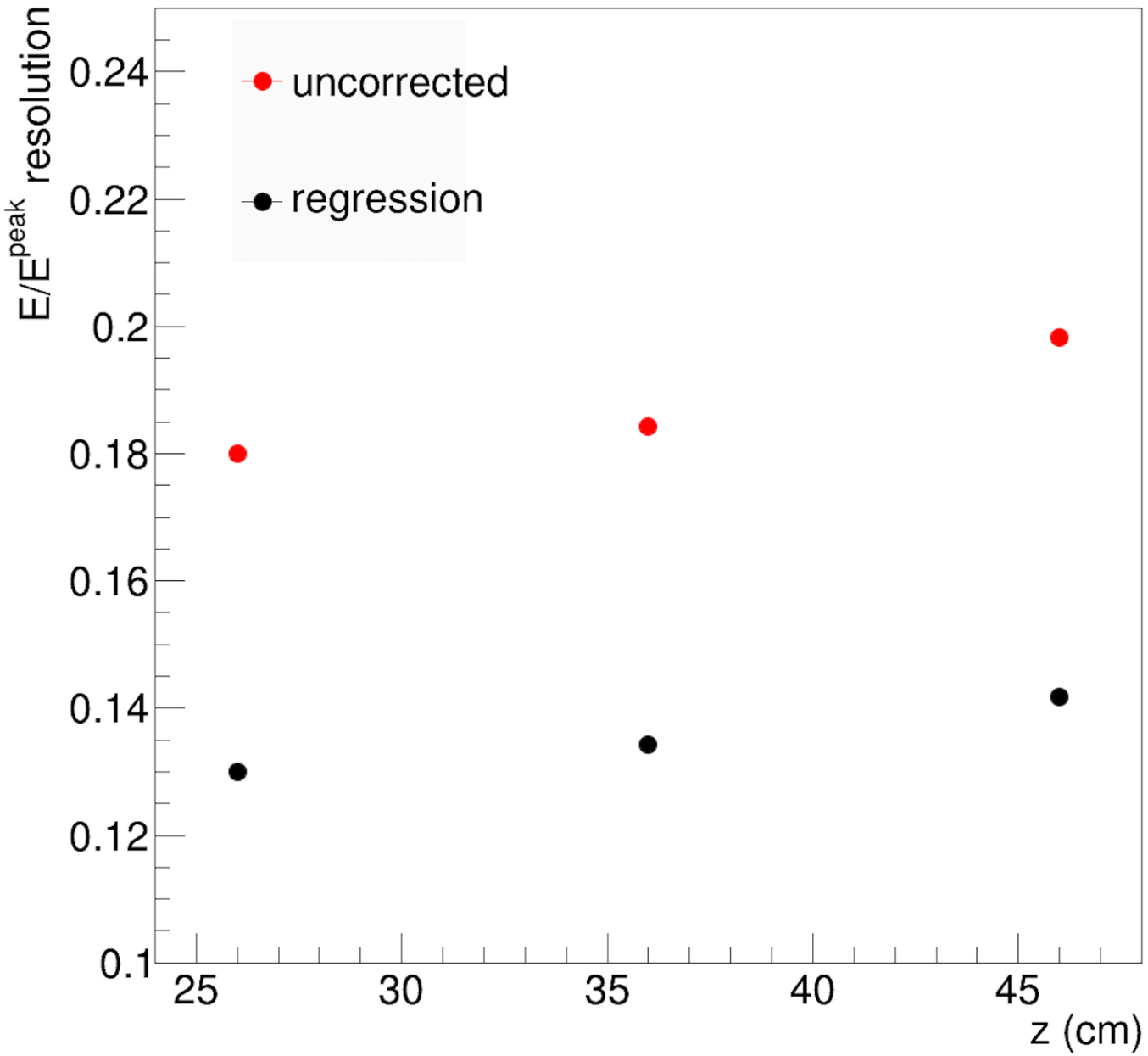
- two possible competing effects: saturation (main) and transport efficiency + diffusion

Dependency reduced to 6% with regression

- need to validate the method with more points in Z (April ^{55}Fe data)

Caveat: interplay with simulation of saturation! (check data-MC again...)

Resolution vs Z



Improvement in resolution substantial at any z

Conclusions

- Multivariate regression gives a large improvement in energy resolution for clusters from 5.9 keV photons

- Regression can give a resolution estimate as well which can be used for categorization, or on a per-cluster basis

- Closure test vs X-Y should be better checked and understood

- There is room for improvements in the training parameters

- Inputs of the regression should be carefully checked between data and MC. If there is a good agreement, it opens the road of training with MC (different kinematics, ERs and NRs...)

- Understand with more granular Z-scan if it is effectively correcting for saturation

- This approach could be used to estimate per-cluster Z, with its uncertainty (3D track reconstruction)