

---

# Calibration des calorimètres de CMS pour la reconstruction de flux de particules.

---

## Résumé :

Les énergies des flux de particules dans le détecteur CMS sont mesurées à l'aide d'un calorimètre électromagnétique (ECAL) et d'un calorimètre hadronique (HCAL). Pour reconstruire les flux de particules dans le traqueur, il nous faut connaître au mieux l'énergie de la particule qui a engendrée ces dépôt d'énergies dans les calorimètres.

Pour se faire, j'ai développé durant ce stage des algorithmes qui, connaissant les énergies déposées dans les calorimètres pour un événement, lui prédisent une énergie de calibration ( $e_{calib}$ ) qui se veut la plus proche possible de la vraie énergie en se basant sur des données d'entraînement simulées, c'est à dire un ensemble d'événements qui contiennent l'énergie déposée dans ECAL, dans HCAL et la vraie énergie,  $(e_{cal}, h_{cal}, e_{true})$ .

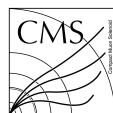
L'idée générale des différentes méthodes proposées est de modéliser ce nuage de points simulés  $(e_{cal}, h_{cal}, e_{true})$ , par une surface  $e_{calib} = f(e_{cal}, h_{cal})$ .

En première approximation, j'ai utilisé une régression linéaire, qui modélise grossièrement le nuage de point et qui met en avant des non-linéarités locales.

Pour prendre en compte les non-linéarités, j'ai maillé le plan  $(e_{cal}, h_{cal})$  en petits carrés et j'ai moyenné les vraies énergies pour obtenir une énergie de calibrations par carré. Cependant, cette méthode de calibration fait apparaître des paliers et est trop dépendante de la répartition des données d'entraînement et de la taille arbitraire des carrés.

Pour lisser cette méthode précédente et enlever ces dépendances, j'ai donc cette fois-ci choisi de travailler en fonction des plus proches voisins : pour un couple  $(e_{cal}, h_{cal})$ , l'énergie calibrée sera dépendante des vraies énergies de ces plus proches voisins.

Cette dernière idée fût la plus prometteuse et reste à être améliorée.



**Mots clefs :** *Calibration, Modélisation, Physique des particules*

Stage encadré par :

**Colin Bernet** colin.bernet@cern.ch

*Bâtiment Paul Dirac*

*4, Rue Enrico Fermi*

*69622 Villeurbanne Cedex*

*Tél. : +33 (0) 4 72 44 84 57*

19 juillet 2017

# Table des matières

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b>   | <b>2</b>  |
| <b>2</b> | <b>Méthodes de calibrations développées pendant le stage</b>                      | <b>2</b>  |
| 2.1      | Explications valables pour toutes les méthodes . . . . .                          | 2         |
| 2.1.1    | Séparation des données . . . . .  | 2         |
| 2.1.2    | Moyenne / moyenne de la gaussienne ajustée ("gaussian fit", "gaussienne fitée") ? | 2         |
| 2.1.3    | Comment est fait un fit ? . . . . .   | 3         |
| 2.2      | Régression Linéaire . . . . .   | 4         |
| 2.3      | Méthode des "legos" . . . . .   | 5         |
| 2.3.1    | Principe général de l'algorithme . . . . .  | 5         |
| 2.3.2    | Résultat de la calibration . . . . .  | 5         |
| 2.4      | Méthode des plus proches voisins (KNN) . . . . .                                  | 7         |
| 2.4.1    | Principe général de l'algorithme . . . . .  | 7         |
| 2.4.2    | Résultat de la calibration . . . . .  | 7         |
| 2.5      | KNN Gaussian Cleaning . . . . .   | 8         |
| 2.5.1    | Principe général de l'algorithme . . . . .  | 8         |
| 2.5.2    | Efficacité du fit . . . . .   | 9         |
| 2.5.3    | Résultat de la calibration . . . . .  | 9         |
| 2.6      | KNN Gaussian Fit . . . . .  | 10        |
| 2.6.1    | Principe général de l'algorithme . . . . .  | 10        |
| 2.6.2    | Résultat de la calibration . . . . .  | 11        |
| <b>3</b> | <b>Comparaison des méthodes</b>   | <b>12</b> |
| 3.1      | Méthodes basées sur KNN . . . . .   | 12        |
| 3.2      | Meilleure méthode . . . . .   | 12        |
| <b>4</b> | <b>Partage du programme</b>   | <b>13</b> |
| <b>5</b> | <b>Annexes</b>  | <b>14</b> |
| 5.1      | Comment créer une calibration ? . . . . .   | 14        |
| 5.2      | Fonctions utiles du programme . . . . .   | 14        |

# 1 Introduction

- annonce decouverte Higgs 2012 par ATLAS [ref] et CMS [ref]. - Ces deux experiences generalistes, installees sur le LHC du CERN, sont toujours en place pour tenter de decouvrir de la nouvelle physique au dela du modele standard. - Les detecteurs ATLAS et CMS sont bases sur les meme principes. - structure cylindrique en couches placée autour du point d'interaction ou les collisions proton-proton sont realisees - but : determiner le type, l'energie et la direction des particules de l'etat final de la collision, pour pouvoir en mesurer les proprietes, et par exemple determiner si une nouvelle particule instable a ete produite. - description du detecteur et du comportement des differentes particules champ magnetique tracker ecal hcal chambre a muons

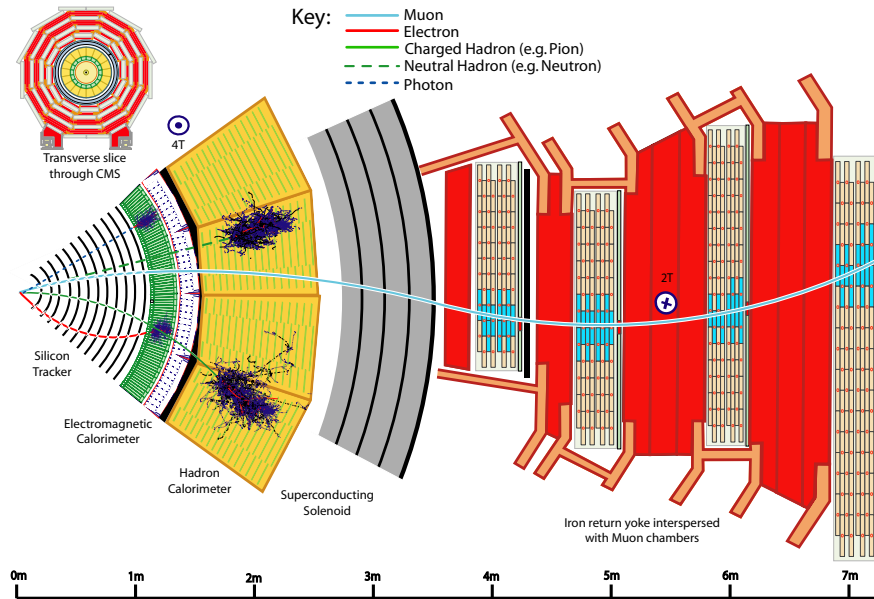


FIGURE 1 – Une esquisse des interactions spécifiques des particules dans une tranche transversale du détecteur CMS.

- expliquer comment on mesure les propriétés des différentes particules. - hadrons neutres : détectés uniquement par les calorimètres - expliquer en 1 ou 2 paragraphes le fonctionnement d'un calorimètre. rapport  $e/\pi$ , non linéarité - nécessite d'une calibration - une technique classique de calibration consiste à déterminer l'énergie calibrée par une fonction linéaire de l'énergie lue dans le ecal, et de l'énergie lue dans le hcal. Cette technique et ses résultats sont présentés dans la section 2. - ce rapport présente une nouvelle technique de calibration qui permet de prendre en compte la non-linéarité des calorimètres. Cette technique est présentée dans la section 3. - enfin, la section 4 présente une comparaison des deux méthodes.

Le but de ce stage est de trouver une méthode de calibration des calorimètres hadroniques et électromagnétiques de CMS, c'est à dire, pour une particule qui va laisser un dépôt d'énergie  $h_{cal}$ ,  $e_{cal}$  dans chacun des calorimètres, comment approximer sa vraie énergie  $e_{true}$  ?

Cette énergie dite énergie calibrée  $e_{calib}$  sera déterminée à l'aide de particules issues d'une simulation très précise (prenant en compte les défauts des calorimètres) qui serviront de données d'entraînement aux différents algorithmes que j'ai développés durant mon stage.

Le but final de cette calibration sera d'améliorer la reconstruction des flux de particules (particules flow).

## 2 Méthodes de calibrations développées pendant le stage

### 2.1 Explications valables pour toutes les méthodes

#### 2.1.1 Séparation des données

On séparera et traitera différemment les événements qui ont  $e_{cal} = 0$ . Ces événements sont liés à des particules qui ont interagité avec le détecteur hadronique mais pas avec le détecteur électromagnétique (cf Fig.1).

Cette séparation se justifie par le fait que modéliser les dépôts d'énergie dans les deux calorimètres pour en conclure ce qui se passe dans le cas particulier où il n'y a des dépôts que dans un amène un biais. Ainsi, à chaque "création" de calibration, on créera en fait deux modèles.

- limite  $e_{cal} + h_{cal} < 150$

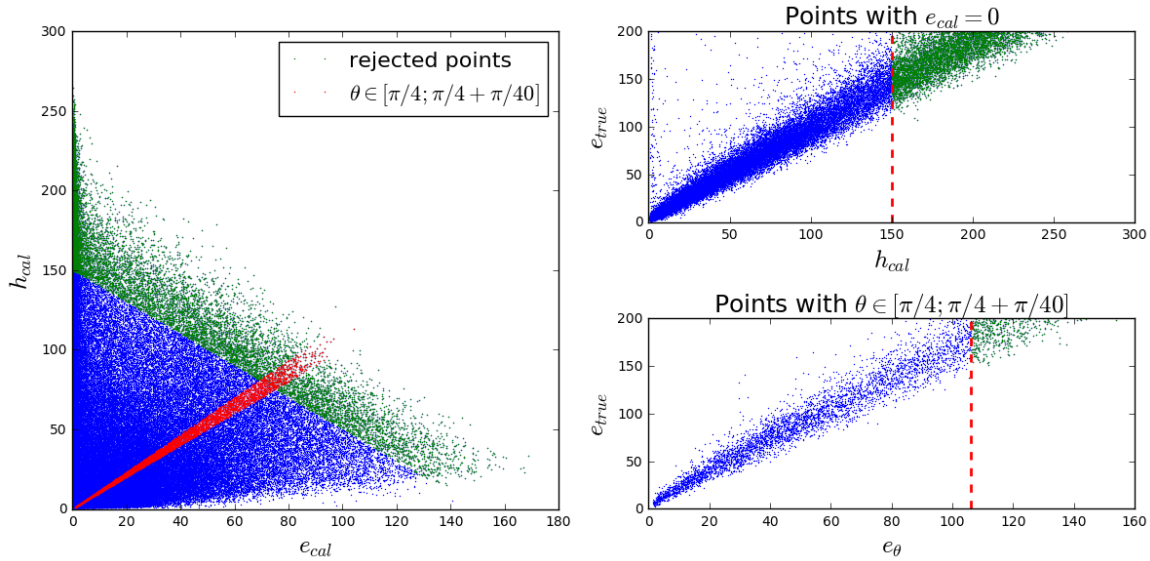


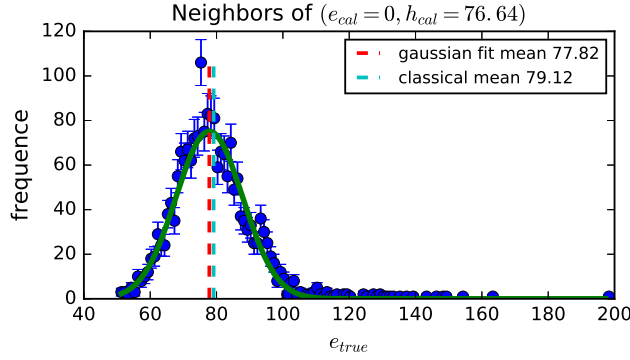
FIGURE 2 – On place une limite à  $e_{cal} + h_{cal} = 150$ . À gauche, ..., en haut à droite, ..., en bas à droite, ...

#### 2.1.2 Moyenne / moyenne de la gaussienne ajustée ("gaussian fit", "gaussienne fitée") ?

À différents moments, nous aurons besoin de calculer des moyennes. Or souvent, la moyenne classique ne serait pas représentative de ce que nous souhaitons montrer car certains points ont des valeurs  $e_{cal}, h_{cal}$  mal estimées car la simulation prend en compte les défauts des calorimètres. Il serait donc alors incorrecte de les prendre en compte pour juger l'efficacité d'une calibration car ils sont complètement incohérents.

Pour résoudre ce problème, nous allons ajuster une gaussienne de la distribution des points à moyenner et choisir considérer que la moyenne à prendre en compte est la moyenne de la gaussienne. Ainsi, les points aberrants totalement écartés du centre de la distribution ne perturberont pas le calcul de la moyenne alors que dans le cas d'une moyenne classique, ils peuvent fortement attirer la moyenne vers eux.

Ces points aberrants peuvent également venir d'une particule qui se serait décomposée avant le calorimètre. Ainsi on trouve près de l'origine, des points à fort  $e_{true}$  et pour de faibles valeurs de  $e_{cal}$  et  $h_{cal}$ , et ces points ne sont pas du tout représentatifs de l'efficacité d'une calibration.



Ici, on peut voir sur cet exemple que si nous prenons la moyenne classique de  $e_{true}$ , on obtient 79.12, or la moyenne de la gaussienne fitée est de 77.82, au vu de ce que nous avons dit précédemment, nous considérerons que la seconde est plus judicieuse.

### 2.1.3 Comment est fait un fit ?

expliquer :

- barre d'erreur
- minimisation du  $\chi^2$
- un bon  $\chi^2$  réduit ?

## 2.2 Régression Linéaire

Pour s'entraîner à l'utilisation de *SciKit Learn*, j'ai d'abord utilisé la régression linéaire. Il s'agit alors de représenter les relations entre les énergies par :

$$e_{true} = a_1 e_{cal} + a_2 h_{cal} + b \quad (1)$$

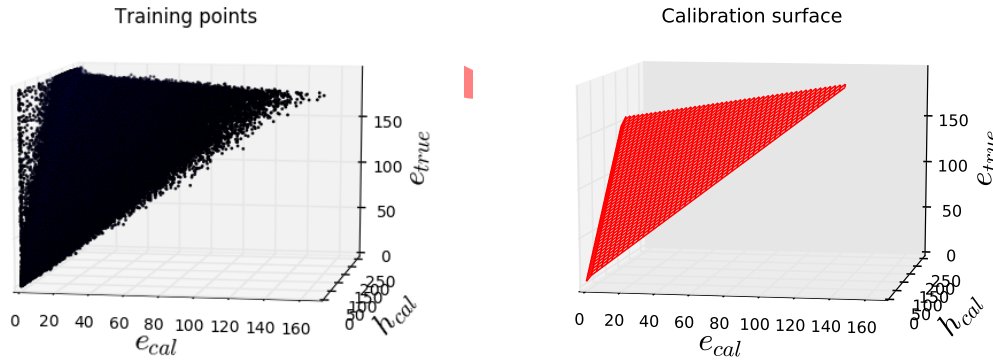


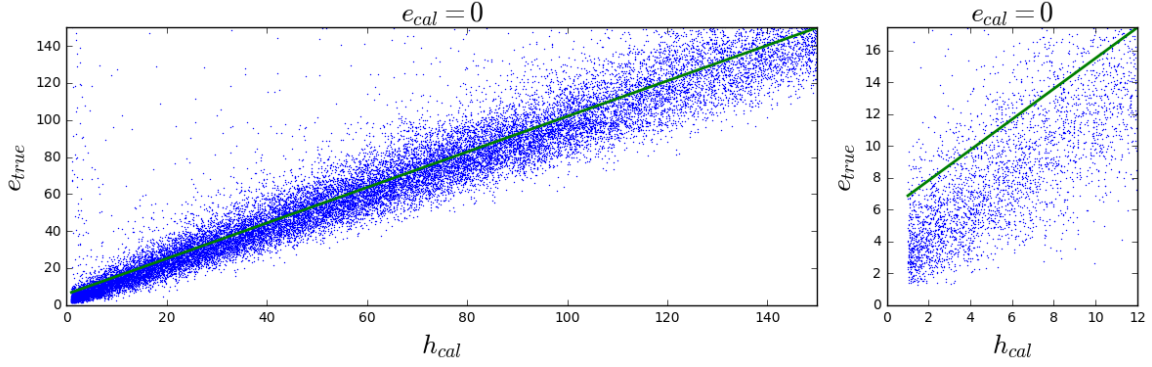
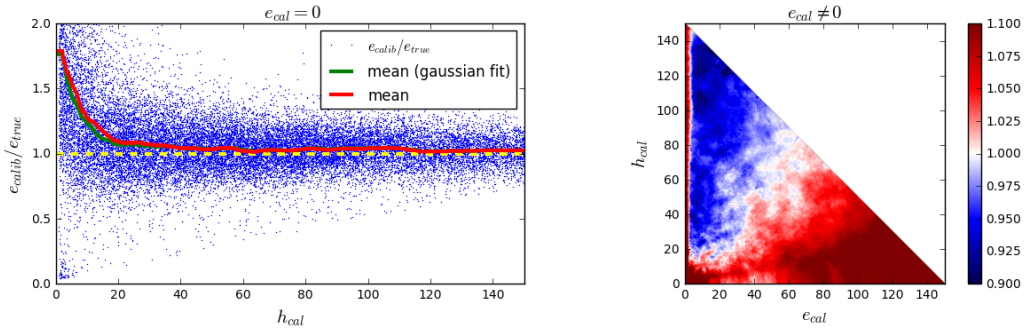
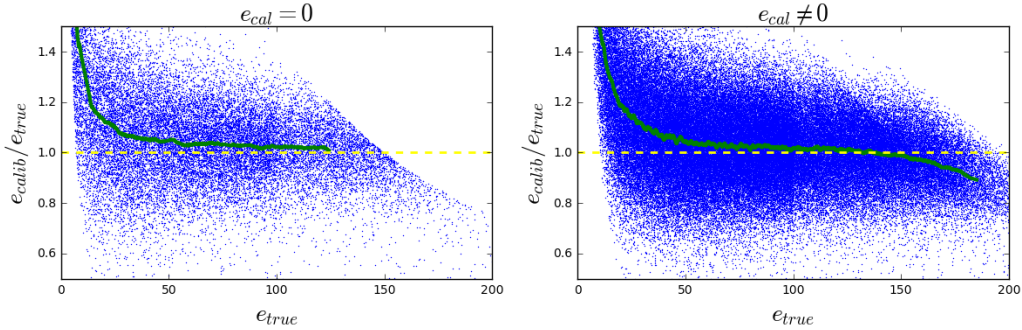
FIGURE 3 – Le nuage de points modélisé (à gauche) par un plan (à droite).

Nous avons ainsi modélisé le nuage de point par un plan, pour voir si cela était réaliste, nous allons d'abord regarder ce qui se passe dans le plan  $e_{cal} = 0$  :

Nous constatons alors que la courbe ne passe pas par le coeur du nuage de point à faible énergie. Pour avoir une vue d'ensemble, nous allons tracer  $e_{calib}/e_{true}$  qui doit être proche de 1 si la calibration est bonne.

En regardant la figure de droite, nous constatons que comme prévu la régression linéaire est mauvaise à faible énergie car en moyenne,  $e_{calib}/e_{true}$  n'est pas proche de 1. Plus intéressant, la figure de droite met en avant les non-linéarités du nuage de point.

Ici nous constatons que à faible et haut  $e_{true}$ , la calibration ne donne pas de bons résultats. En effet, la courbe de la moyenne (fit gaussien) s'écarte très fortement d'une constante égale à 1.

FIGURE 4 – Courbe de calibration pour  $e_{cal} = 0$ .FIGURE 5 –  $e_{calib}/e_{true}$  en fonction de  $e_{cal}$  et  $h_{cal}$ .FIGURE 6 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .

## 2.3 Méthode des "legos"

### 2.3.1 Principe général de l'algorithme

Comme nous l'avons vu précédemment, il faut une calibration qui prenne en compte les non-linéarités. Ici, l'idée est de découper le plan  $(e_{cal}, h_{cal})$  en carré et de calculer la moyenne des  $e_{true}$  dans chaque carré qui sera la valeur  $e_{calib}$ .

Ainsi pour prédire une énergie de  $e_{calib}^i$  pour un point  $(e_{cal}^i, h_{cal}^i)$ , nous allons regarder dans quel carré il se trouve et retourner la valeur d'énergie calibrée correspondante, faisant apparaître ainsi des "legos".



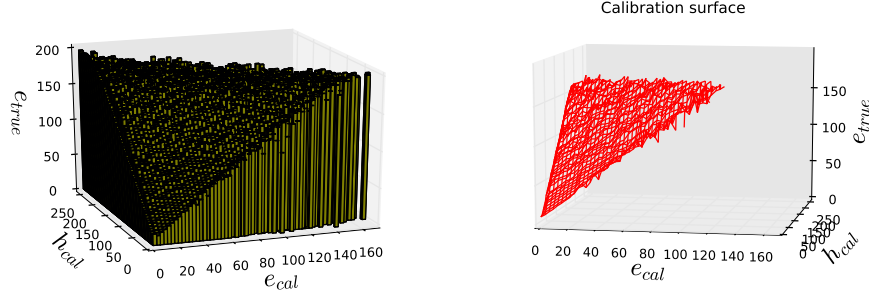


FIGURE 7 – Le nuage de points modélisé par des legos (à gauche) ainsi que la surface correspondante (à droite).  $100 \times 100$  legos.

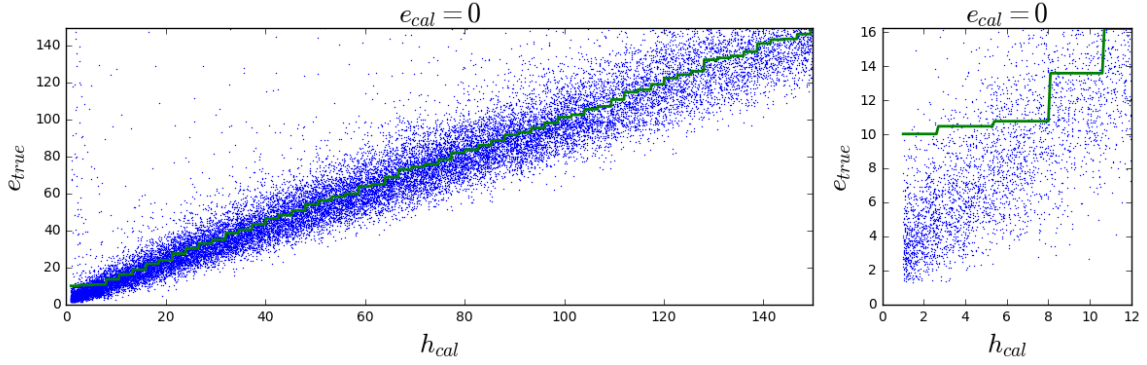


FIGURE 8 – Courbe de calibration pour  $e_{cal} = 0$ .

### 2.3.2 Résultat de la calibration

Bien que cette méthode prenne en compte les linéarité, nous pouvons voir sur les figures ci-dessus qu'il y a un effet de pas, ce qui n'est pas bon car beaucoup trop d'événements ont la même  $e_{calib}$  et la courbe de calibration ne suit pas bien le cœur de distribution, surtout à faible énergie.

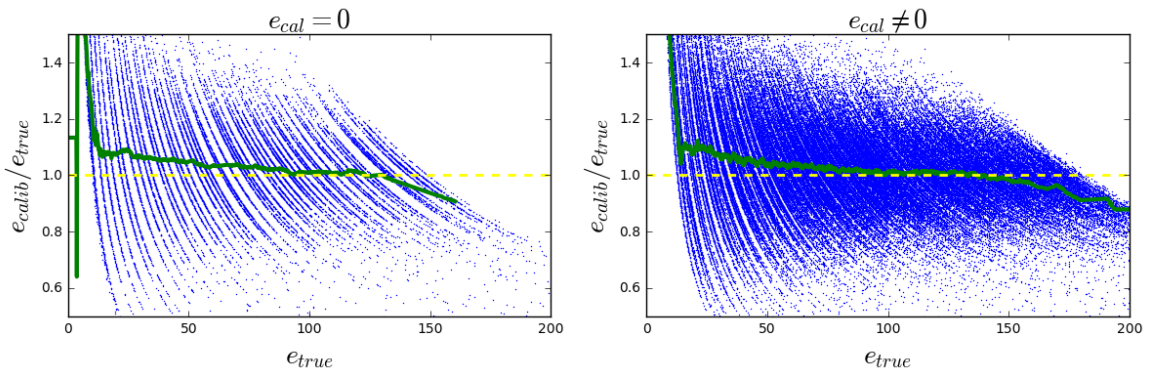
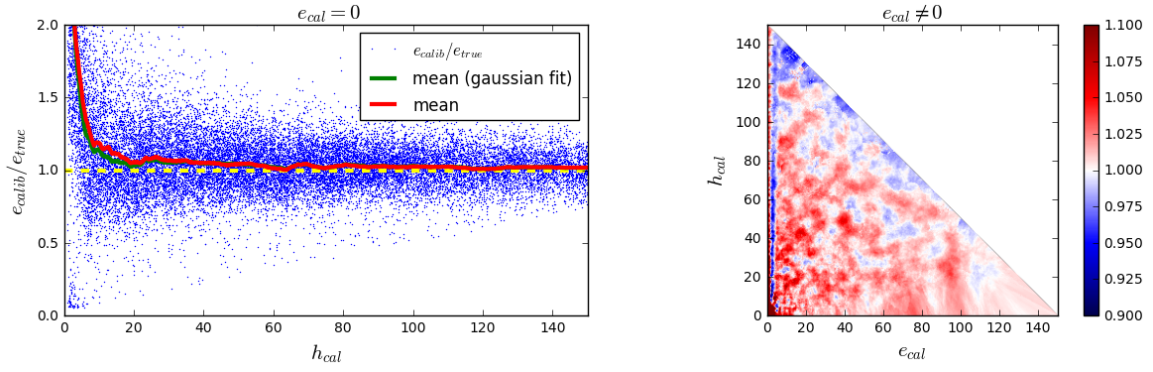


FIGURE 9 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .

Cet effet de pas se retrouve également si l'on trace  $e_{calib}/e_{true}$  en fonction de  $e_{true}$  (Fig. 9) et nous y voyons alors une structure (des hyperboles) liées aux points qui ont la même énergie de calibration (contrairement à la régression linéaire Fig. 6).

Cet illustration montre à nouveau que nous sommes loin d'une répartition des points autour de  $e_{calib}/e_{true} = 1$ .

FIGURE 10 –  $e_{calib}/e_{true}$  en fonction de  $e_{cal}$  et  $h_{cal}$ .

Ici nous constatons malgré tout que nous avons mieux pris en compte la non-linéarité, mais que en majorité, l'énergie calibrée est sur-estimée.

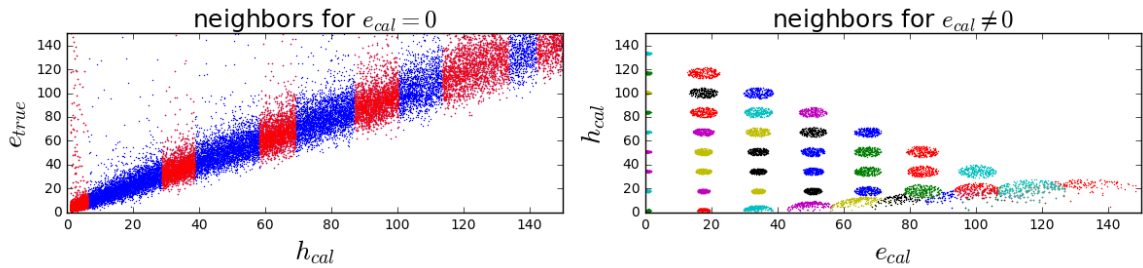
## 2.4 Méthode des plus proches voisins (KNN)

### 2.4.1 Principe général de l'algorithme

Nous utilisons encore des données simulées pour effectuer une calibration, chaque particule simulée  $i$  est vue comme un point d'un espace tridimensionnel possédant des coordonnées  $(e_{cal}^i, h_{cal}^i, e_{true}^i)$ , correspondant respectivement à l'énergie déposée dans le calorimètre électromagnétique, dans le calorimètre hadronique et l'énergie vraie.

Pour trouver l'énergie calibrée d'un point de coordonnées  $(e_{cal}^0, h_{cal}^0)$  :

- on recherche ses  $k$  plus proches voisins dans le plan  $(e_{cal}, h_{cal}) \rightarrow (e_{cal}^i, h_{cal}^i), i \in [1, \dots, k]$

FIGURE 11 –  $n_{voisins} = 2000$  pour  $e_{cal} = 0$ ,  $n_{voisins} = 250$  pour  $e_{cal} \neq 0$ 

- on effectue une moyenne pondérée une moyenne pondérée de l'énergie vraie de ces plus proches voisins  $\rightarrow e_{calib}^0$  : l'énergie calibrée

La moyenne pondérée va donc s'exprimer ainsi :

$$e_{calib}^0 = \frac{\sum_{i=1}^k g(e_{cal}^i, h_{cal}^i) \times e_{true}^i}{\sum_{i=1}^k g(e_{cal}^i, h_{cal}^i)} \quad (2)$$

Dans notre cas nous avons pris pour  $g$  la distribution gaussienne  $g(\vec{x}) = \exp -\frac{1}{2}(\frac{(\vec{x}-\vec{x}^0)^2}{\sigma^2})$ , pour donner plus d'importance aux plus proches des  $k$  plus proches voisins.



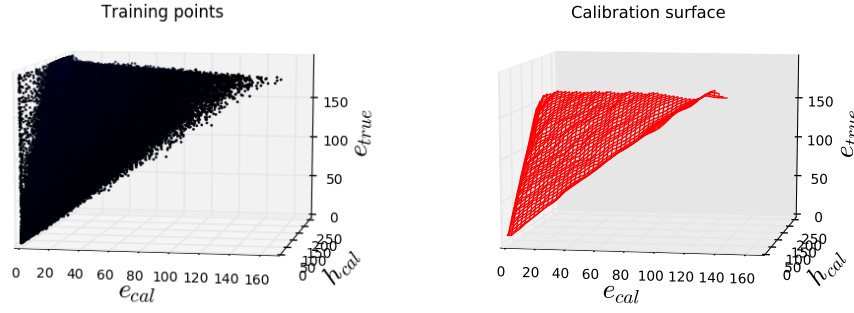
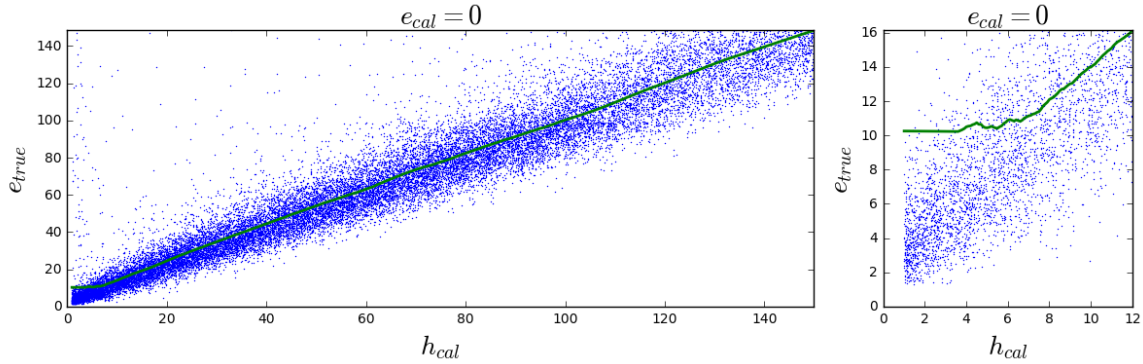


FIGURE 12 – Le nuage de points modélisé (à gauche) par une surface (à droite).

### 2.4.2 Résultat de la calibration

Nous constatons ici que la surface est beaucoup plus lisse que pour la méthode précédente, mais en regardant le cas particulier de  $e_{cal} = 0$ , nous constatons encore une fois que à faible énergie, la courbe de calibration ne passe pas par le coeur de la distribution. Cela vient du fait qu'il y a des points aberrants qui ont une forte énergie vraie mais qui ont une très faible énergie détectée par les calorimètres.

Il nous faut donc un moyen pour ne plus les prendre en compte pour avoir une courbe de calibration plus réaliste.

FIGURE 13 – Courbe de calibration pour  $e_{cal} = 0$ .

Malgré tout, en regardant la Fig. 14, nous constatons que les points sont mieux répartis autour de  $e_{calib}/e_{true} = 1$ .

Nous prenons également en compte les non-linéarité dans ce cas mais nous sur-estimons encore la valeur de l'énergie calibrée (encore une fois, à cause de ces points à fort  $e_{true}$ ).

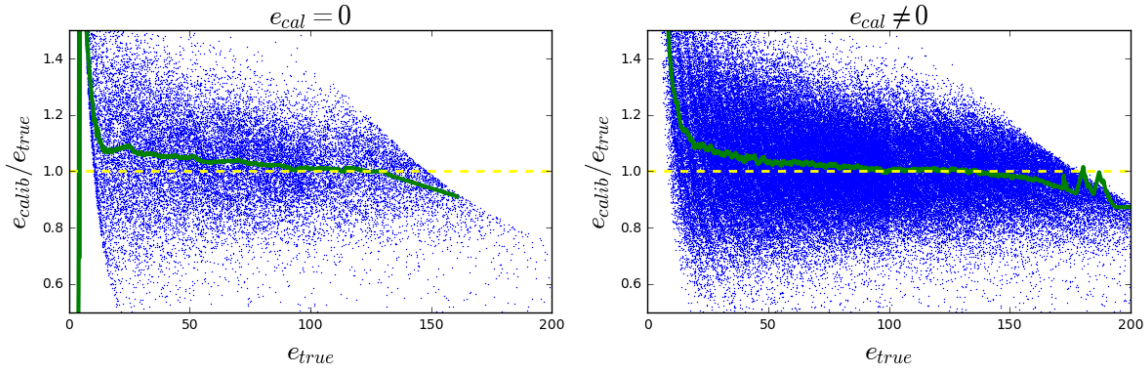
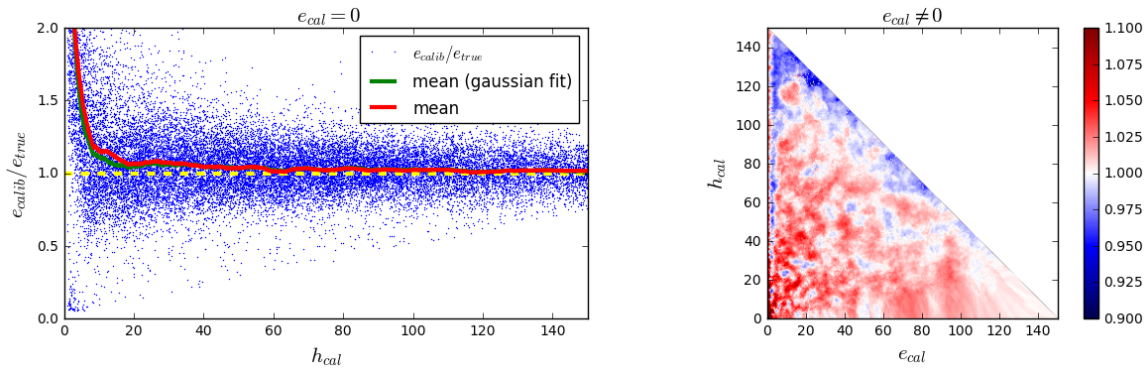
## 2.5 KNN Gaussian Cleaning

### 2.5.1 Principe général de l'algorithme

Cette méthode est assez similaires à la précédente. Elle se base sur la constatation que la distribution en énergie vraie des paquets de plus proches voisins est une distribution gaussienne. Nous allons donc en utilisant la méthode des moindres carrés trouver les paramètres de la gaussienne en question et ne prendre en compte les plus proches voisins dont l'énergie vraie est  $\mu - c\sigma \leq e_{true}^i \leq \mu + c\sigma$  (nous prenons par défaut  $c = 2$ ), avec  $\mu, \sigma$  la moyenne et l'écart type de la distribution gaussienne.

Principe de l'algorithme :

- on considère des points  $(e_{cal}^{0,j}, h_{cal}^{0,j})$  où nous allons évaluer l'énergie calibrée.

FIGURE 14 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .FIGURE 15 –  $e_{calib}/e_{true}$  en fonction de  $e_{cal}$  et  $h_{cal}$ .

- pour chaque  $(e_{cal}^{0,j}, h_{cal}^{0,j})$  :
  - on recherche ses  $k$  plus proches voisins dans le plan  $(e_{cal}, h_{cal}) \rightarrow (e_{cal}^i, h_{cal}^i), i \in [1, \dots, k]$
  - on trouve la gaussienne correspondante  $\mu - c\sigma \leq e_{true}^i \leq \mu + c\sigma$
  - on ne conserve que les voisins dont :  $\mu - c\sigma \leq e_{true}^i \leq \mu + c\sigma$
  - on effectue une moyenne pondérée une moyenne pondérée de l'énergie vraie de ces plus proches voisins  $\rightarrow e_{calib}^0$  : l'énergie calibrée
- on effectue une interpolation pour donner une valeur d'énergie calibrée quelque soit  $(e_{cal}^0, h_{cal}^0)$

### 2.5.2 Efficacité du fit

### 2.5.3 Résultat de la calibration

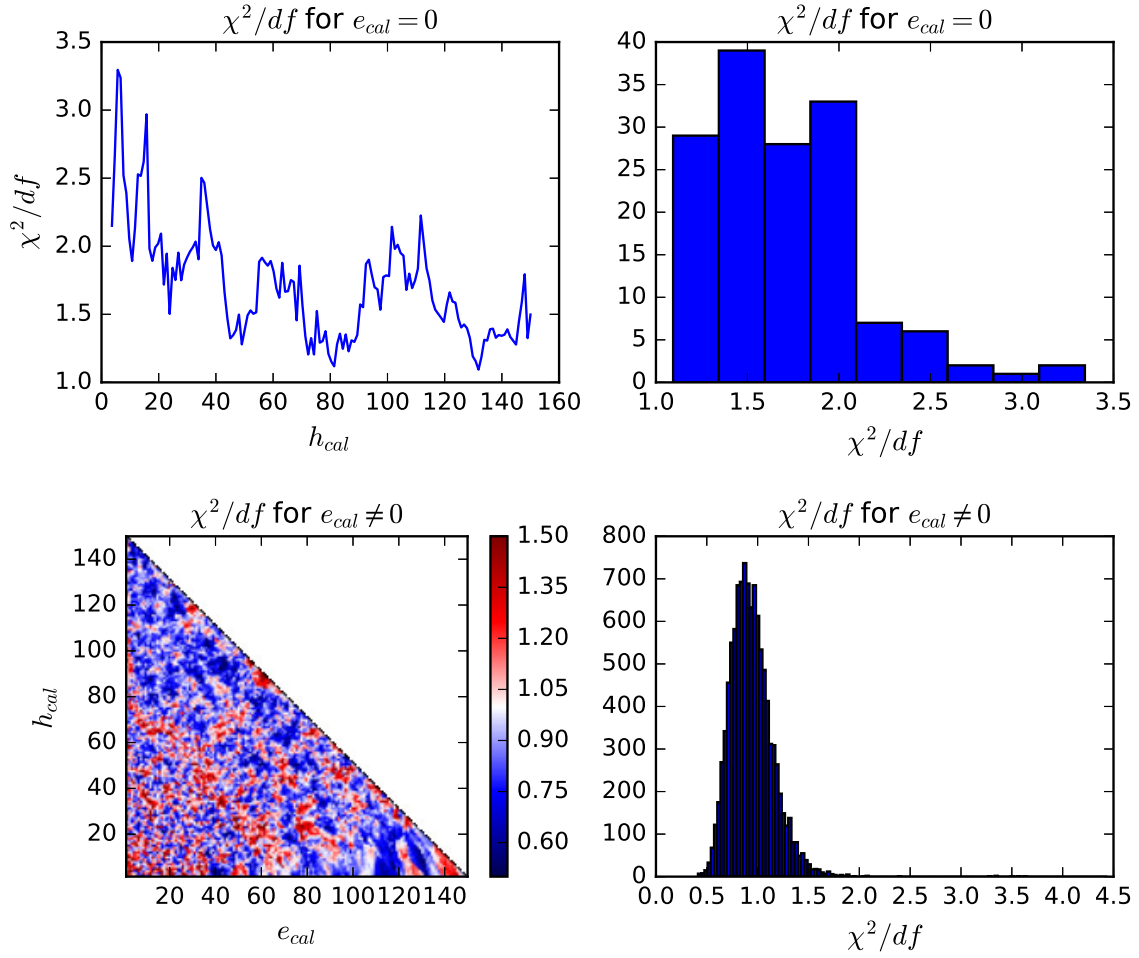
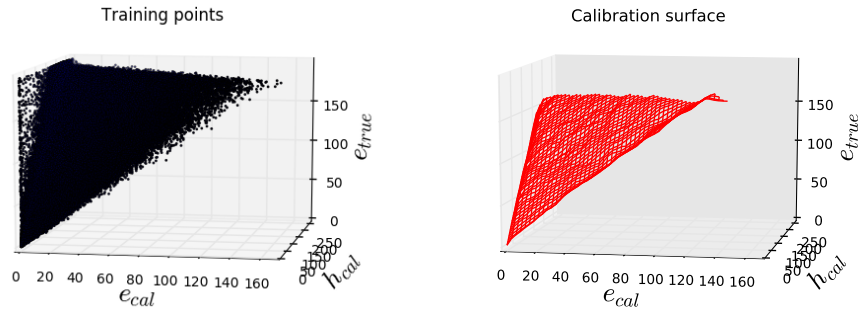
FIGURE 16 – Le  $\chi^2$  réduit pour chaque fit effectué.

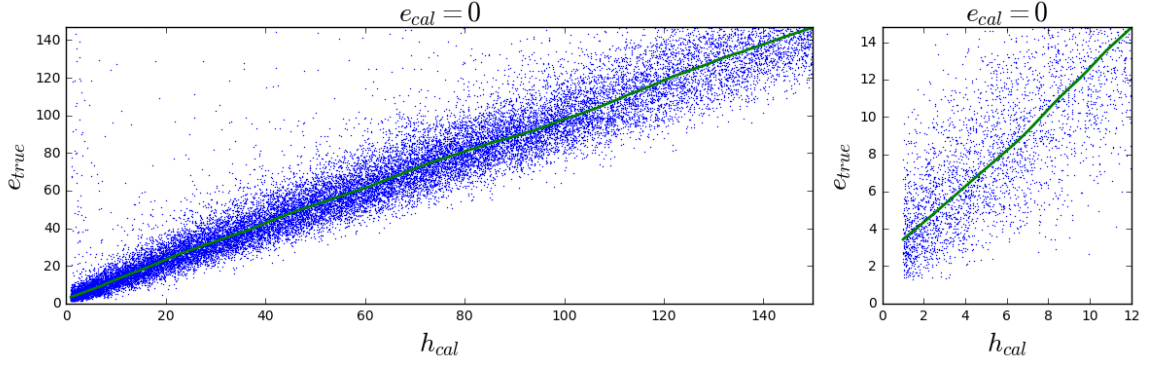
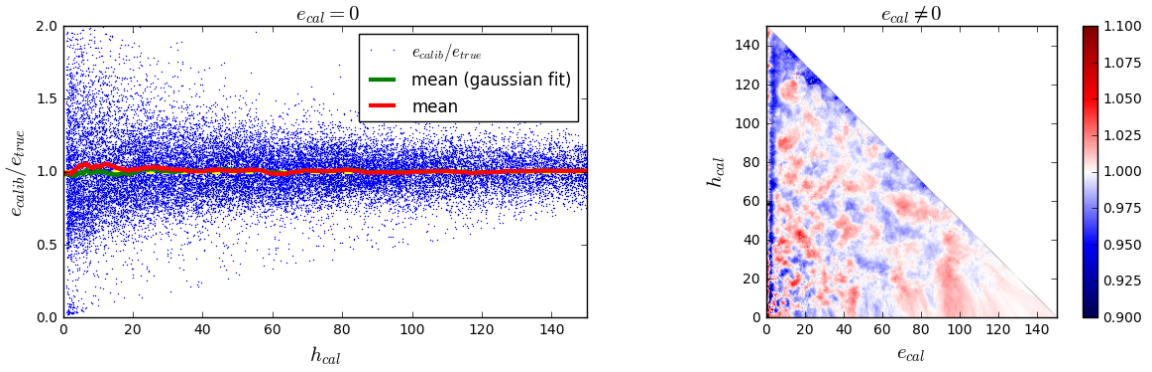
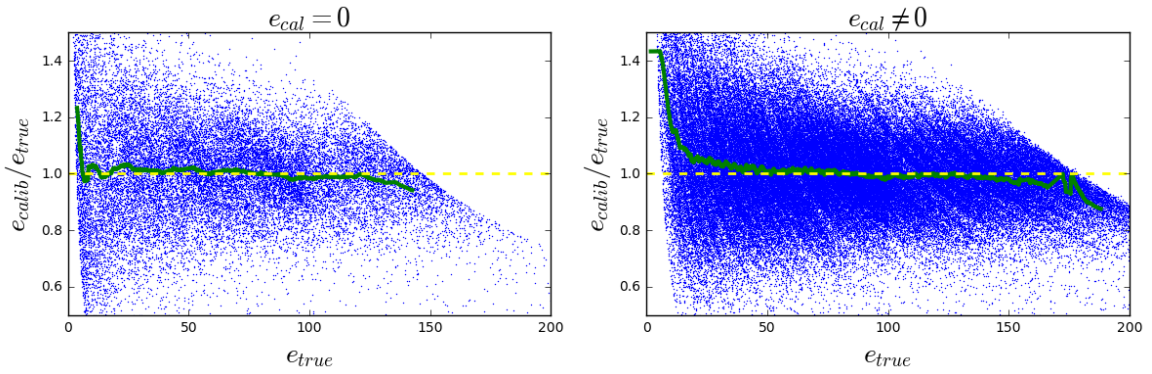
FIGURE 17 – Le nuage de points modélisé (à gauche) par une surface (à droite).

## 2.6 KNN Gaussian Fit

### 2.6.1 Principe général de l'algorithme

Ici, il s'agit du même principe que précédemment mais nous allons considérer que la valeur de  $e_{calib}$  est la moyenne de la gaussienne. Principe de l'algorithme :

- on considère des points  $(e_{cal}^{0,j}, h_{cal}^{0,j})$  où nous allons évaluer l'énergie calibrée.
- pour chaque  $(e_{cal}^{0,j}, h_{cal}^{0,j})$  :
  - on recherche ses  $k$  plus proches voisins dans le plan  $(e_{cal}, h_{cal}) \rightarrow (e_{cal}^i, h_{cal}^i), i \in [1, \dots, k]$

FIGURE 18 – Courbe de calibration pour  $e_{cal} = 0$ .FIGURE 19 –  $e_{calib}/e_{true}$  en fonction de  $e_{cal}$  et  $h_{cal}$ .FIGURE 20 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .

- on trouve la gaussienne correspondante  $\rightarrow \sigma, \mu$
- $\rightarrow e_{calib}^0 = \mu$
- on effectue une interpolation pour donner une valeur d'énergie calibrée quelque soit  $(e_{cal}^0, h_{cal}^0)$

### 2.6.2 Résultat de la calibration

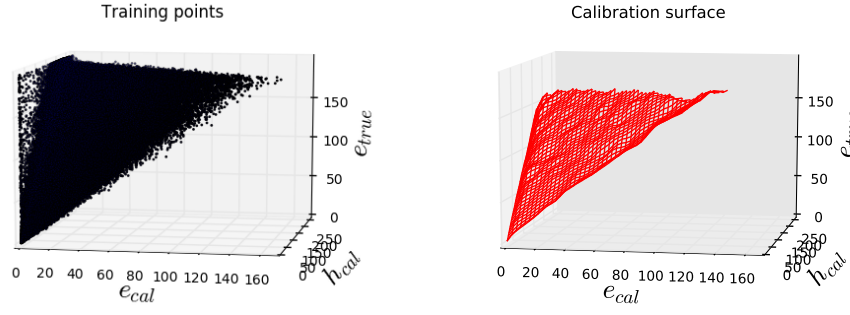
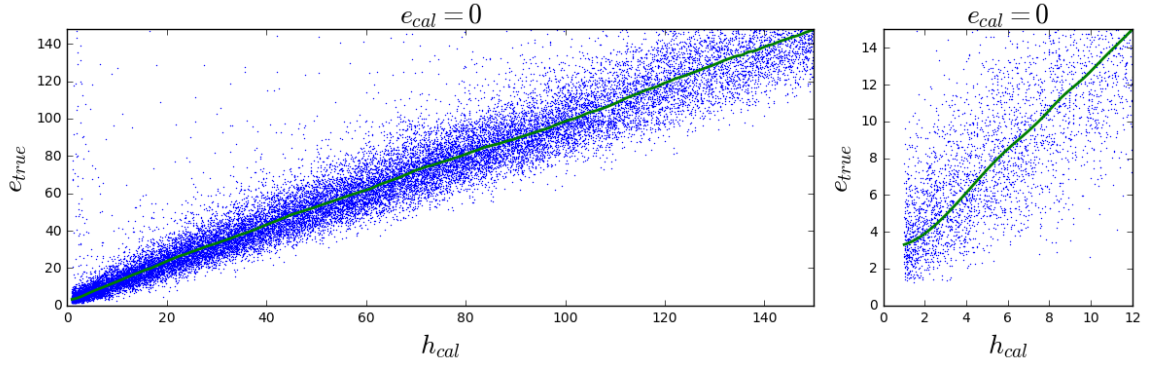
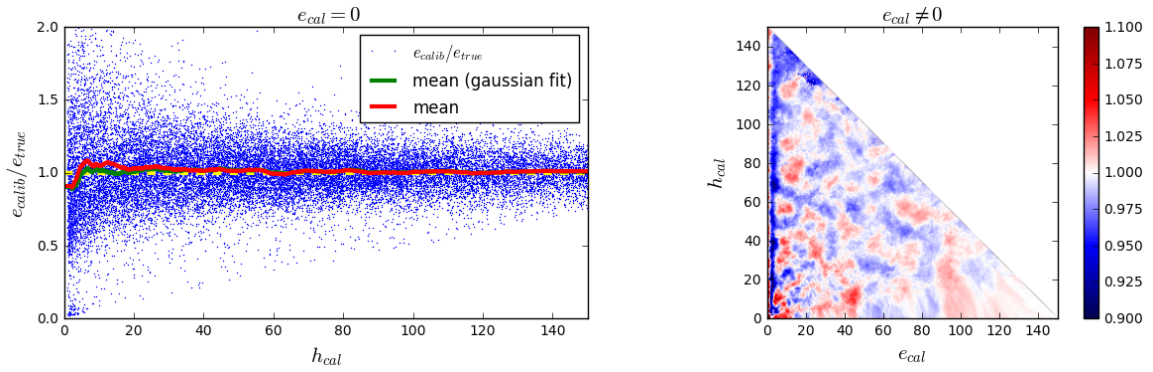


FIGURE 21 – Le nuage de points modélisé (à gauche) par une surface (à droite).

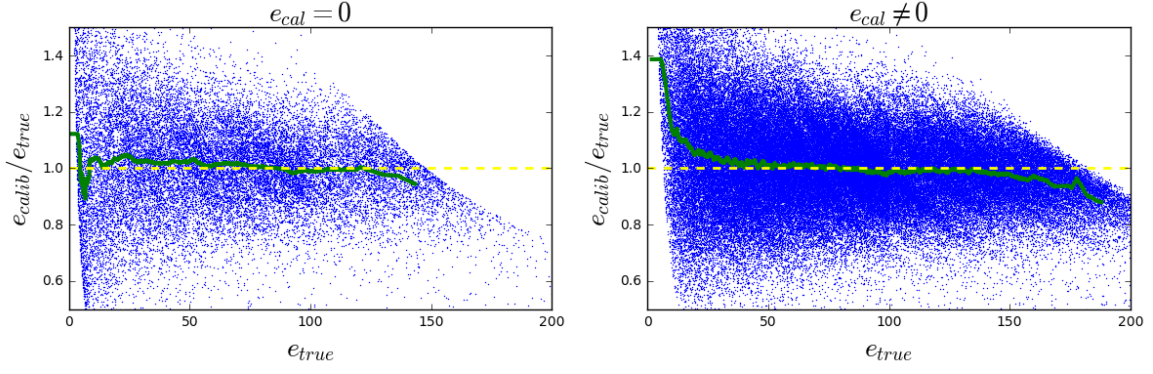
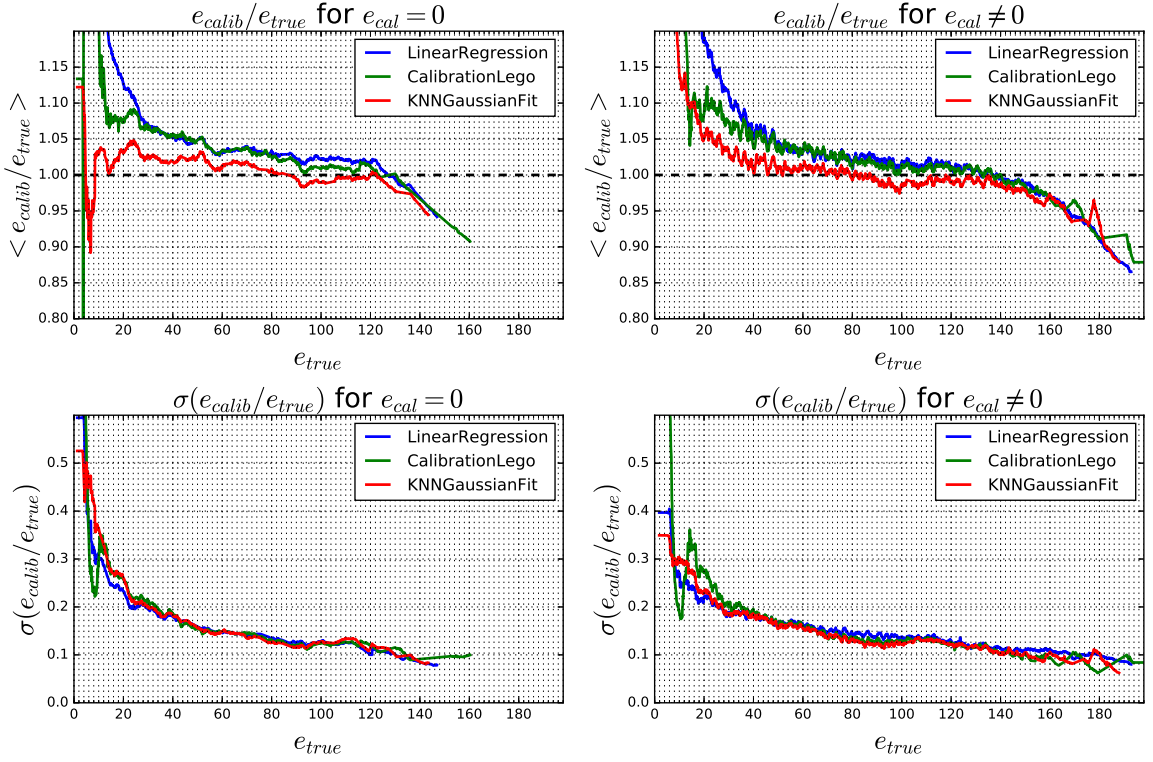
FIGURE 22 – Courbe de calibration pour  $e_{cal} = 0$ .FIGURE 23 –  $e_{calib}/e_{true}$  en fonction de  $e_{cal}$  et  $h_{cal}$ .

### 3 Comparaison des méthodes

#### 3.1 Méthodes basées sur KNN

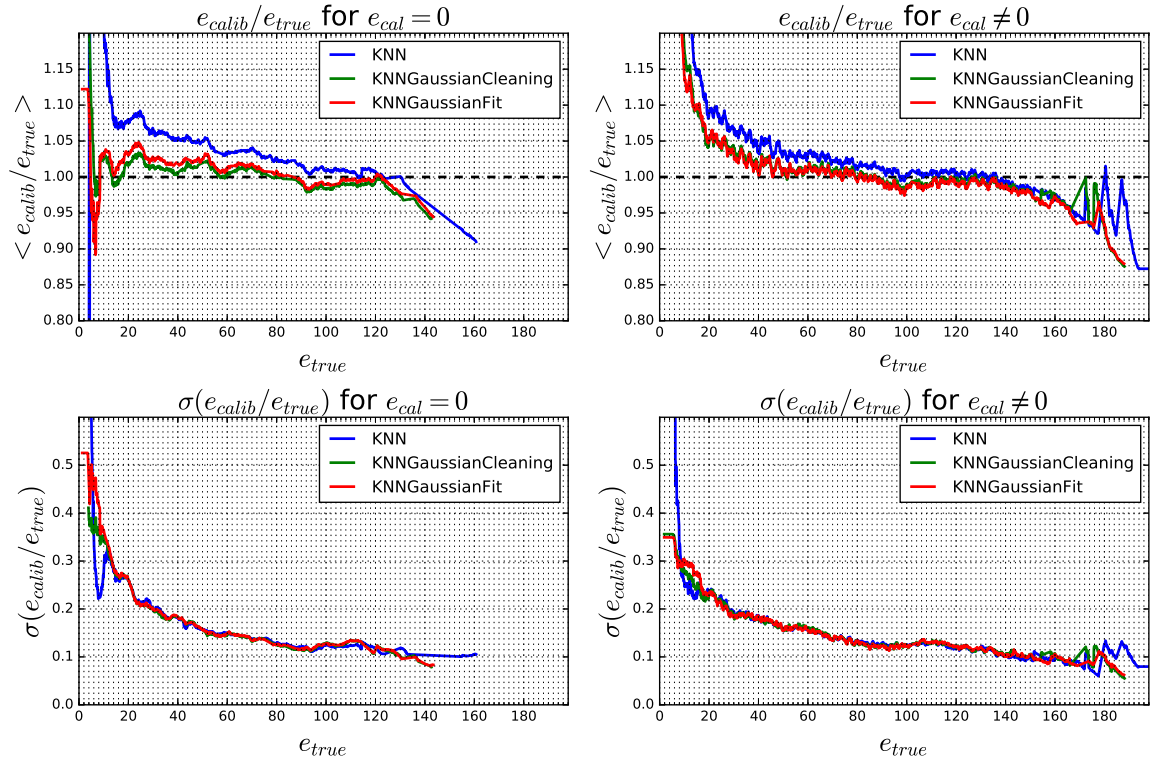
#### 3.2 Meilleure méthode



FIGURE 24 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .FIGURE 25 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .

## 4 Partage du programme



FIGURE 26 –  $e_{calib}/e_{true}$  en fonction de  $e_{true}$ .

## 5 Annexes

### 5.1 Comment créer une calibration ?

### 5.2 Fonctions utiles du programme