

# Outils pour l'analyse des données spatiales

## Sommaire

---

<b>II.1 Construction de mosaïques</b> . . . . .	<b>49</b>
II.1.1 Reprojection . . . . .	50
II.1.2 Rejet des rayons cosmiques . . . . .	51
II.1.3 Combinaison des différentes images . . . . .	53
<b>II.2 Techniques de photométrie</b> . . . . .	<b>54</b>
II.2.1 Photométrie d'ouverture . . . . .	54
II.2.2 Ajustement de PSF . . . . .	60
II.2.3 Correction de couleur . . . . .	62
<b>II.3 Le stacking</b> . . . . .	<b>63</b>
II.3.1 Principe . . . . .	63
II.3.2 Estimation des erreurs avec le bootstrap . . . . .	66

---

Ce chapitre assez technique présente différents outils que nous avons été amenés à utiliser au cours de cette thèse. Nous présenterons dans un premier temps le principe de la construction de mosaïques pour l'instrument IRAC (décrit au paragraphe 1.5). Nous exposerons et discuterons ensuite différentes méthodes de photométrie. Enfin, il sera question de la technique d'empilement (ou de *stacking*) que nous avons abondamment utilisée tout au long de ces travaux.

## II.1 Construction de mosaïques

Les détecteurs d'IRAC sont des matrices de 256x256 pixels et couvrent une région de 5'x5' sur le ciel. Or on effectue presque toujours des observations sur une région plus vaste. Il est

donc nécessaire de faire une mosaïque, c'est-à-dire d'assembler les différentes observations afin d'obtenir une grande et unique image.

Pour cela, il est commode d'utiliser le logiciel MOPEX<sup>1</sup> distribué par le SSC (Spitzer Science Center) qui présente plusieurs avantages : il interprète tout d'abord correctement les informations de distortion géométrique données dans l'en-tête des fichiers FITS. De plus, l'utilisateur a le choix entre plusieurs algorithmes pour supprimer les effets des rayons cosmiques sur les données. Enfin, c'est un logiciel très rapide.

Les méthodes de rejet des rayons cosmiques sont, de façon générale, d'autant plus performantes que le nombre de données est grand. C'est pourquoi il vaut mieux utiliser les données de type BCD (Basic Calibrated Data) distribuées par le SSC, que l'on peut par exemple télécharger sur la base de données à l'aide du logiciel Leopard<sup>2</sup>, plutôt que les données post-BCD qui contiennent déjà une mosaïque par AOR (Astronomical Observation Request).

Les ajustements de cette méthode ont été effectués en compagnie de Aurèle Duval et Julien Annaloro, deux stagiaires de L3 que j'ai encadrés, et ont été appliqués à la construction des mosaïques des champs GTO du CDF-S, HDF-N, Lockman Hole et XMM. Les étapes importantes de l'algorithme sont détaillées dans les paragraphes suivants.

### II.1.1 Reprojection

La première étape consiste à déterminer les paramètres astrométriques de la mosaïque finale étant donné un jeu d'images à mosaïquer. On obtient alors une grille sur laquelle toutes les images BCD vont être reprojctées. Il est à ce moment possible de choisir la taille des pixels de la mosaïque et il est recommandé de les prendre carrés<sup>3</sup> (ce qui n'est pas toujours le cas par défaut).

Quatre algorithmes de reprojection sont proposés par MOPEX : nous avons utilisé celui dit de «Pixel Overlap Integration», qui est celui proposé par défaut. La valeur  $M_j$  du pixel  $j$  de la mosaïque finale est la somme des pixels initiaux  $I_i$  pondérée par les surfaces communes  $a_{ij}$  entre les pixels  $i$  et  $j$  :

$$M_j = \sum_i \frac{a_{ij}}{A_j} I_i \quad \text{avec} \quad A_j = \sum_i a_{ij} \quad (\text{II.1})$$

Cette projection est appliquée aussi bien aux images qu'aux cartes d'erreurs (dans ce cas, on a une somme quadratique).

---

1. <http://ssc.spitzer.caltech.edu/postbcd/mopex.html>

2. <http://ssc.spitzer.caltech.edu/propkit/spot/index.html>

3. car les procédures IDL de la librairie ASTRON ne fonctionnent pas correctement avec des pixels rectangulaires...

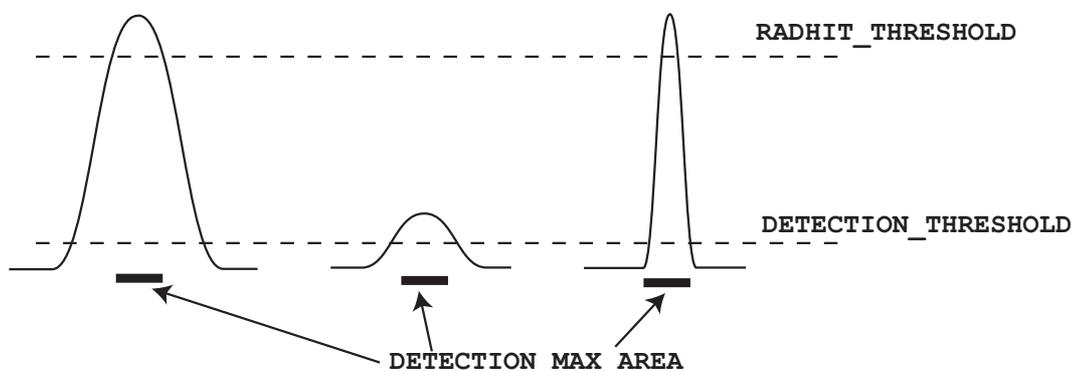
## II.1.2 Rejet des rayons cosmiques

Des particules de haute énergie, les rayons cosmiques, viennent parfois frapper le détecteur au cours d'une observation et les pixels traversés par ces rayons cosmiques voient alors leur réponse grandement modifiée. Le passage d'astéroïdes, surtout lorsque les observations se font au voisinage de l'écliptique, peuvent également altérer les observations. Tous ces phénomènes transitoires ne nous intéressent en général pas et nous ne souhaitons pas qu'ils apparaissent sur la mosaïque finale. Il sera d'autant plus facile de détecter de tels phénomènes si on dispose d'un grand nombre d'observations (supérieur à 3) de la même zone du ciel, puisque par nature, ceux-ci n'apparaîtront que sur une seule observation. Cependant, on ne dispose parfois que d'une ou deux observations, tout particulièrement sur les bords de la région observée.

MOPEX propose quatre algorithmes qui permettent d'isoler les rayons cosmiques (ou n'importe quel autre phénomène passager) :

- **Filtrage spatial**

Le premier algorithme ne nécessite pas d'avoir plusieurs images de la même région. Il repose sur l'idée que les rayons cosmiques sont de petite taille et de très grande intensité. Il est nécessaire, pour utiliser cet algorithme, d'avoir préalablement enlevé le continu des images (par exemple avec un filtre médian). Le principe est illustré figure II.1.



**Figure II.1** – Illustration du principe de rejet des rayons cosmiques selon la méthode du filtrage spatial. D'après Spitzer Mosaicker.

Les trois paramètres ajustables de cet algorithme sont : le seuil de segmentation `SEGMENTATION_THRESHOLD`, le seuil de rayons cosmiques `RADHIT_THRESHOLD` et la surface maximale de détection `DETECTION_MAX_AREA`. Il commence par sélectionner tous les groupes de pixels contigus dont la valeur est supérieure au seuil de segmentation. Ensuite deux conditions doivent être satisfaites pour qu'un groupe de pixels soit considéré comme un rayon cosmique : premièrement, le nombre de pixels doit être inférieur au nombre de pixels maximum. Ce premier critère a pour effet de rejeter les vraies sources brillantes qui sont par nature plus

étendues (même si celles-ci sont ponctuelles...). Deuxièmement, la valeur du pixel le plus brillant du groupe doit être supérieure au seuil de détection des rayons cosmiques. Cette nouvelle condition permet de séparer les sources faibles des rayons cosmiques.

Cette méthode présente plusieurs limitations : les rayons cosmiques d'intensité modérée, ou qui ont frappé le détecteur avec une incidence rasante (et qui sont donc étendus) ne seront pas détectés. Cette méthode ne permet pas non plus de repérer les astéroïdes.

#### • Filtrage temporel

Cette méthode nécessite quant à elle plusieurs observations de la même région. En pratique, il en faut au moins deux, mais plus elles seront nombreuses et meilleur sera le filtrage. Toutes ces observations ont été préalablement reprojetees sur une grille commune (II.1.1) et pour chaque pixel de la nouvelle grille, on dispose de plusieurs observations  $I_k$ . On choisit de rejeter l'observation  $k$  si l'une des deux conditions suivantes n'est pas vérifiée :

$$\text{BOTTOM\_THRESHOLD} \times \sigma < I_k - \bar{I} < \text{TOP\_THRESHOLD} \times \sigma \quad (\text{II.2})$$

où `BOTTOM_THRESHOLD` et `TOP_THRESHOLD` sont deux paramètres ajustables et  $\bar{I}$  et  $\sigma$  sont respectivement la moyenne et l'écart-type de la distribution des  $I_k$ . Cependant si le nombre d'observations est faible,  $\sigma$  et  $I_k$  seront biaisés. Dans ce cas (si le nombre d'observations est inférieur à `MIN_PIX_NUM`), on peut préférer utiliser les cartes d'incertitudes qui fournissent également une estimation de  $\sigma$  : on prend alors le minimum des incertitudes des différentes observations.

#### • Filtrage spatio-temporel

Cette méthode plus complexe est une combinaison des deux précédentes. On commence par créer, pour chaque observation, une carte de détection comme dans l'algorithme de filtrage spatial. On compare ensuite ces différentes cartes de détection : tout pixel qui n'appartient à un objet détecté que sur une fraction d'images inférieure à `MAX_OUTL_FRAC` et sur un nombre d'images inférieur ou égal à `MAX_OUTL_IMAGE` est considéré comme un *dual-outlier*. Enfin, l'algorithme examine chaque groupe de pixels constituant une détection et le considère comme un rayon cosmique si plus de la moitié des pixels sont des dual-outliers. Cette dernière étape permet d'éliminer les faux pixels outliers. En effet, si dans un groupe de pixels, seuls quelques uns sont marqués comme des duals outliers, il y a de fortes chances pour que ça n'en soit pas et qu'ils aient été marqués comme tels parce qu'ils étaient sur le bord d'une source.

#### • Filtrage temporel par boîte

Cette méthode est très voisine du filtrage temporel présenté ci-dessus, mais elle permet de mieux gérer les cas où le nombre d'observations est faible. Avec cet algorithme,  $\sigma$  et  $\bar{I}$  ne

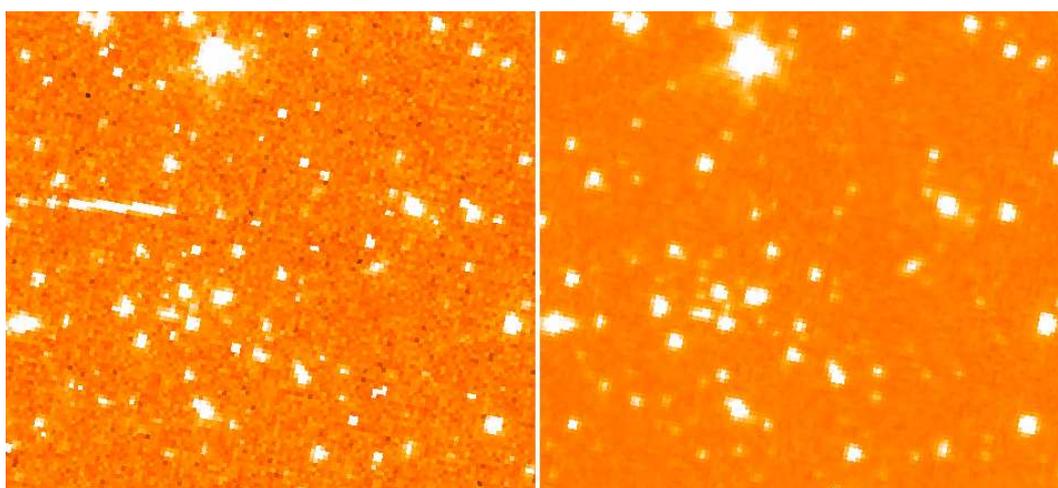
sont plus calculés sur l'échantillon des  $N$  observations du même pixel, mais sur l'échantillon composé des  $N$  observations des pixels d'une région de largeur `BOX_X` et de hauteur `BOX_Y` centré sur le pixel, ce qui permet d'augmenter la statistique.

On peut choisir d'appliquer plusieurs types de filtrages, ce qui est souvent un choix judicieux, car les différentes méthodes sont complémentaires. Pour notre part, nous avons choisi de combiner les résultats des trois premiers algorithmes (le dernier n'existait pas lorsque nous avons construit les mosaïques) En combinant toutes ces informations, on obtient finalement un *masque-R*. Dans ce masque, on attribue quatre bits à chaque pixel, chacun des bits correspondant à un des quatre algorithmes.

### II.1.3 Combinaison des différentes images

Les pixels qui ont été marqués comme étant contaminés par un rayon cosmique par un des quatre algorithmes précédents sont tout d'abord masqués et doivent donc être reprojétés sur la nouvelle grille. En effet chaque pixel de la grille originale affecté par un rayon cosmique correspond en réalité à plusieurs pixels sur la nouvelle grille et le fait de le masquer affecte la valeur de ces pixels qui doivent donc être recalculés. On ne reprojette pas l'intégralité des images, mais seulement les pixels masqués afin d'économiser du temps de calcul.

Une fois cette opération effectuée, il ne reste plus qu'à combiner les différentes observations. On effectue une moyenne pondérée par les cartes d'erreur, dans laquelle les pixels masqués précédemment ne sont pas pris en compte. On obtient finalement la mosaïque, ainsi qu'une mosaïque d'erreurs et une mosaïque de couverture (nombre d'observations par pixel).



**Figure II.2** – Comparaison d'une image BCD (à gauche) et de la mosaïque finale de cinq BCD (à droite) pour des données IRAC 3.6  $\mu\text{m}$ . La taille du champ est environ 2.5'x2.5'. Les niveaux de couleurs sont les mêmes pour les deux images.

Une image BCD et le résultat de la mosaïque sont montrés figure II.2 : on remarque que de nombreux rayons cosmiques ont été supprimés. On remarque aussi que la mosaïque est moins bruitée : en effet, le rapport signal sur bruit augmente avec le temps d'intégration, ie. avec la combinaison d'observations indépendantes.

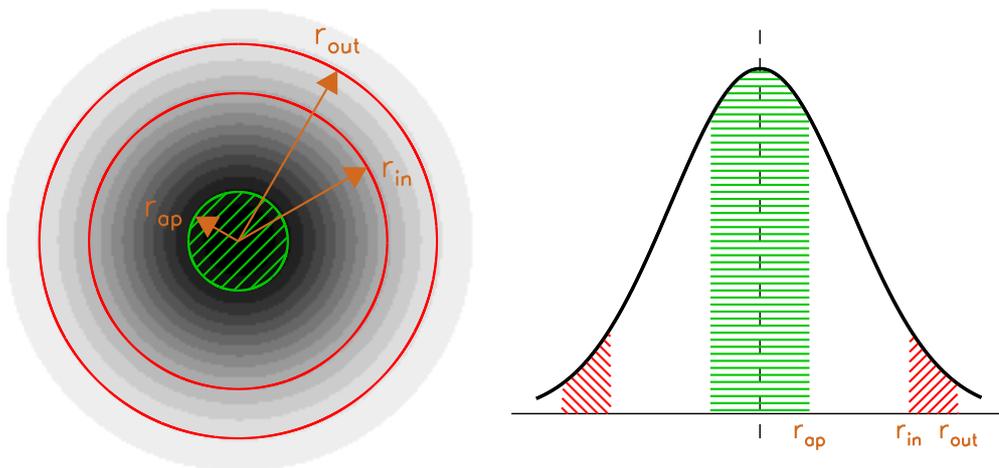
La table II.1 récapitule la valeur des différents paramètres que nous choisissons pour la construction des mosaïques IRAC des champs GTO.

## II.2 Techniques de photométrie

Nous présentons ici deux méthodes qui permettent de mesurer le flux de sources ponctuelles, c'est-à-dire non résolues par l'instrument et dont l'image est donc la figure de diffraction de tout le système optique. On appellera PSF (pour Point Spread Function) l'image d'une source ponctuelle à travers tout le système. Dans le cas des données Spitzer il est recommandé d'extraire soi-même la PSF sur ses propres données : pour cela, il suffit de sélectionner une dizaine de sources brillantes, non résolues et isolées et de les additionner.

### II.2.1 Photométrie d'ouverture

La photométrie d'ouverture consiste à intégrer l'énergie d'une source à travers une ouverture circulaire de rayon  $r_{ap}$  et centrée sur la source, et à lui soustraire la valeur du fond mesuré dans une couronne autour de la source, de rayon intérieur  $r_{in}$  et de rayon extérieur  $r_{out}$ .



**Figure II.3** – Illustration du principe de photométrie d'ouverture. La PSF est représentée en niveau de gris à gauche et en coupe à droite.

Supposons que l'on souhaite mesurer le flux  $S$  d'une source posée sur un fond de valeur

Module	Paramètre	Valeur
medfilter	WINDOW_X	40
	WINDOW_Y	40
	N_OUTLIERS_PER_WINDOW	1000
detect_radhit	DETECTION_MAX_AREA	7
	SEGMENTATION_THRESHOLD	2
	RADHIT_THRESHOLD	15
mosaic_int	INTERP_METHOD	1
detect	DETECTION_MAX_AREA	50000
	DETECTION_MIN_AREA	0
	DETECTION_THRESHOLD	3
mosaic_covg	TILEMAX_X	2000
	TILEMAX_Y	2000
mosaic_dual_outlier	MAX_OUTL_IMAGE	1
	MAX_OUTL_FRAC	0.5
level	THRESHOLD_RATIO	0.5
mosaic_outlier	THRESH_OPTION	1
	BOTTOM_THRESHOLD	3
	TOP_THRESHOLD	20
	MIN_PIX_NUM	4
	TILE_XSIZ	1000
	TILE_YSIZ	1000
	mosaic_rmask	RM_THRES
	MIN_COVERAGE	3
	MAX_COVERAGE	100
mots-clés divers	have_uncertainties	1
	compute_uncertainties_internally	0
	use_outlier_for_rmask	1
	use_dual_outlier_for_rmask	1
	use_box_outlier_for_rmask	0
	sigma_weighted_coadd	1
	create_unc_mosaic	1

**Table II.1** – Récapitulatif des différents paramètres utilisés pour la construction des mosaïques GTO de IRAC à toutes les longueurs d'onde.

$B$  que l'on supposera uniforme. Le profil de cette source est  $P(r, \theta)$  où  $r$  et  $\theta$  sont analogues aux coordonnées polaires dont l'origine est prise au centre de la source mais  $r$  est ici un angle :

$$P(r, \theta) = S \times \text{PSF}(r, \theta) + B \quad (\text{II.3})$$

PSF( $r, \theta$ ) est le profil de la PSF. Celui-ci est normalisé de façon à avoir  $\iint \text{PSF}(r, \theta) r dr d\theta = 1$ . La mesure du fond dans la couronne située entre les rayons  $r_{\text{in}}$  et  $r_{\text{out}}$  donne :

$$B_{\text{mes}} = \frac{1}{A_{\text{couronne}}} \int_0^{2\pi} \int_{r_{\text{in}}}^{r_{\text{out}}} P(r, \theta) r dr d\theta \quad (11.4)$$

où  $A_{\text{couronne}}$  est l'aire de la couronne :  $A_{\text{couronne}} = \pi(r_{\text{out}}^2 - r_{\text{in}}^2)$ . On obtient alors :

$$B_{\text{mes}} = B + \frac{S}{A_{\text{couronne}}} \int_0^{2\pi} \int_{r_{\text{in}}}^{r_{\text{out}}} \text{PSF}(r, \theta) r dr d\theta \quad (11.5)$$

On mesure finalement le flux  $S_{\text{mes}}$  dans un ouverture de rayon  $r_{\text{ap}}$  en ayant pris soin de préalablement soustraire la valeur du continu estimée précédemment. On a :

$$S_{\text{mes}} = \int_0^{2\pi} \int_0^{r_{\text{ap}}} (P(r, \theta) - B_{\text{mes}}) r dr d\theta \quad (11.6)$$

qui conduit à :

$$S_{\text{mes}} = S \times \left[ \int_0^{2\pi} \int_0^{r_{\text{ap}}} \text{PSF}(r, \theta) r dr d\theta - \frac{A_{\text{aperture}}}{A_{\text{couronne}}} \int_0^{2\pi} \int_{r_{\text{in}}}^{r_{\text{out}}} \text{PSF}(r, \theta) r dr d\theta \right] \quad (11.7)$$

où  $A_{\text{aperture}} = \pi r_{\text{ap}}^2$  est l'aire de l'ouverture.

Ainsi le flux mesuré  $S_{\text{mes}}$  par photométrie d'ouverture est proportionnel au vrai flux  $S$  de la source. Le facteur de proportionnalité est appelé «correction d'ouverture». La valeur de cette correction ne dépend que de la PSF et des paramètres de photométrie (à savoir la valeur de  $r_{\text{ap}}$ ,  $r_{\text{in}}$  et  $r_{\text{out}}$ ) et peut être facilement calculée. Ce facteur tient compte du fait que seule une partie du flux de la source a été intégrée, et que le continu est légèrement surestimé car mesuré trop près de la source. En pratique, on choisit les rayons  $r_{\text{ap}}$ ,  $r_{\text{in}}$  et  $r_{\text{out}}$  de façon à minimiser cette correction, ce facteur étant généralement compris entre 1 et 2.

Dans les faits, on choisira un rayon d'ouverture  $r_{\text{ap}}$  ni trop petit (car la statistique est moins bonne et car on est plus sensible à un éventuel défaut de centrage) ni trop grand (car on ne gagne plus en rapport signal sur bruit et on risque d'intégrer une partie du flux d'une source voisine). Une taille de l'ordre de la FWHM de la PSF semble optimum. Les rayons de la couronne  $r_{\text{in}}$  et  $r_{\text{out}}$  seront choisis un peu plus grands, de l'ordre de quelques fois la FWHM. Le choix de ces valeurs est aussi le fruit d'un compromis : d'une part le rayon de la couronne doit être inférieur à la taille des éventuelles inhomogénéités du continu afin de mesurer le plus précisément possible la valeur du continu sous la source, et d'autre part la largeur de cette couronne doit être assez grande pour avoir un nombre de pixels suffisants dans la couronne pour estimer le continu le plus précisément possible.

L'estimation de la correction d'ouverture est un point délicat, même si le principe est simple. Il suffit en effet de mesurer le flux d'une PSF avec les paramètres décrits plus haut et de comparer au flux absolu de cette PSF. C'est ce dernier point (la normalisation de la PSF)

qui pose problème car il est difficile de connaître le «flux d'une PSF» lorsque celle-ci a été extraite sur les données car on ne connaît pas la valeur absolue du continu. Autrement dit il est difficile d'estimer précisément l'énergie contenue dans les ailes de la PSF et cette énergie n'est en général pas négligeable. Par exemple, pour les PSF de l'instrument MIPS, environ 15% de l'énergie est contenue dans les pixels situés à des distances supérieures à  $4 \times \text{HWHM}$ <sup>4</sup> du centre de la PSF.

Nous avons adopté cette méthode de photométrie pour les données provenant de l'instrument IRAC et également pour les données provenant de MIPS 70 et MIPS 160. A  $24 \mu\text{m}$  nous avons préféré la méthode décrite au paragraphe suivant. Les valeurs des paramètres utilisés pour la photométrie sont récapitulées dans le tableau II.2.

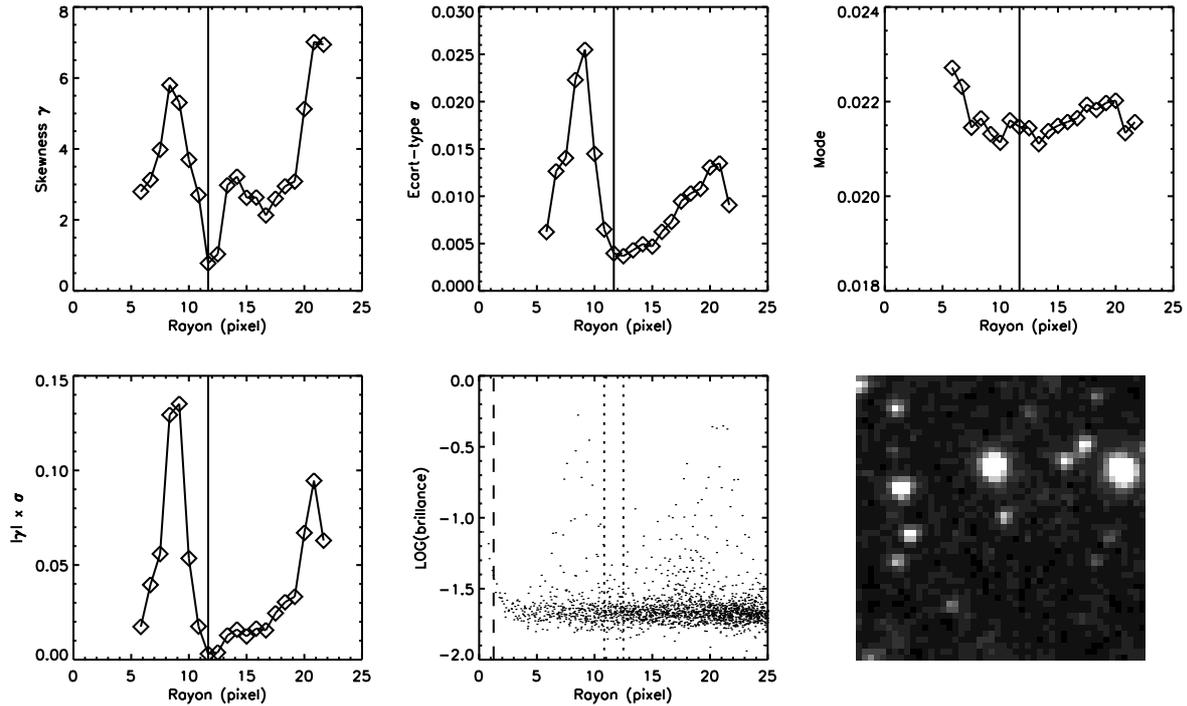
Instrument	$r_{\text{ap}}$	$[r_{\text{in}} ; r_{\text{out}}]$	Correction d'ouverture
IRAC $3.6 \mu\text{m}$	$1.5''$	adaptatif	1.74
IRAC $4.5 \mu\text{m}$	$1.5''$	adaptatif	1.83
IRAC $5.8 \mu\text{m}$	$1.5''$	adaptatif	2.18
IRAC $8.0 \mu\text{m}$	$1.5''$	adaptatif	2.44
MIPS $24 \mu\text{m}$	$6.6''$	$[20'' ; 32'']$	1.65
MIPS $70 \mu\text{m}$	$18''$	$[50'' ; 70'']$	1.68
MIPS $160 \mu\text{m}$	$25''$	$[80'' ; 110'']$	2.19

**Table II.2** – Récapitulatif des différents paramètres utilisés pour la photométrie d'ouverture pour les différentes longueurs d'onde Spitzer. A  $24 \mu\text{m}$ , où nous avons très peu utilisé cette méthode, nous avons reporté dans cette table les valeurs données dans le Handbook.

Aux longueurs d'onde IRAC, la densité de sources est très grande et l'estimation du continu peut-être biaisée par la présence d'une ou plusieurs sources dans la couronne. Nous avons donc mis au point une méthode pour optimiser le rayon de la couronne afin d'éviter la présence de sources dans cette couronne. L'épaisseur de la couronne est fixe et vaut  $2''$ , mais le rayon peut quant à lui varier de  $7$  à  $26''$  par pas de  $1''$ . Pour chacune de ces couronnes, on mesure  $|\gamma|\sigma$  où  $\sigma$  et  $\gamma$  sont respectivement l'écart-type et l'asymétrie (*skewness*) de la distribution des pixels contenus dans la couronne. On cherche ensuite le minimum de cette fonction. En effet, si aucune source n'est présente dans la couronne, la distribution des valeurs des pixels est gaussienne ( $\gamma = 0$ ) et  $\sigma$  mesure le bruit de photons. En revanche si des sources sont présentes dans la couronne, celles-ci vont modifier la distribution des pixels en ajoutant des pixels plus brillants : l'asymétrie ne sera donc plus nulle et l'écart-type  $\sigma$  sera plus grand. Minimiser  $|\gamma|\sigma$  revient donc à chercher des couronnes non contaminées par des sources. Un exemple du résultat de cette optimisation est donné figure II.4.

Nous avons utilisé la procédure `aper` du package DAOPHOT pour faire de la photométrie

4. Half Width Half Maximum

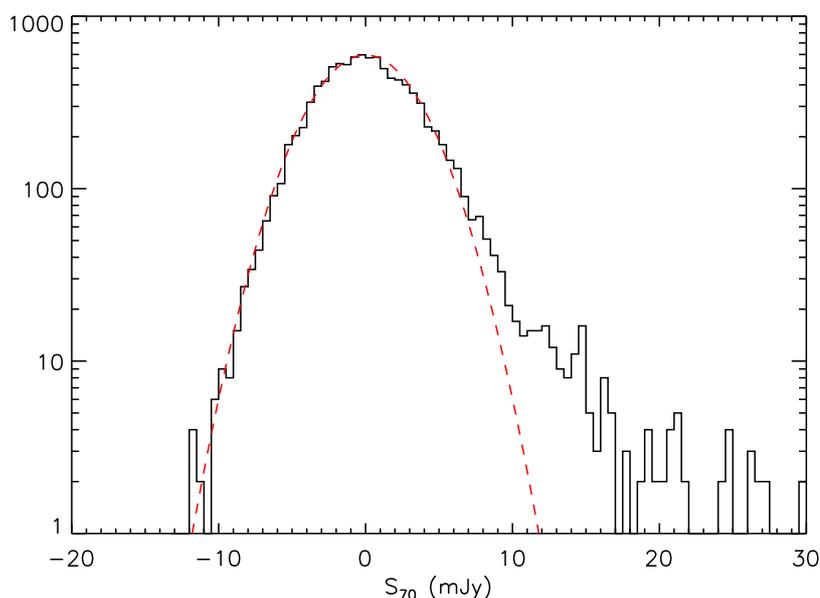


**Figure II.4** – Illustration de la méthode d'optimisation du rayon de la couronne servant à mesurer le continu dans le cas de la photométrie IRAC (ici à  $3.6 \mu\text{m}$ ). De haut en bas et de gauche à droite sont tracées, l'asymétrie  $\gamma$ , l'écart-type  $\sigma$ , le mode et la quantité  $|\gamma|\sigma$  de la distribution des pixels situés dans une couronne de rayon donné en abscisse et de largeur  $2''$ . Le cinquième graphe représente le profil radial de la source dont on souhaite mesurer le flux. Enfin le dernier graphe est une image de  $\sim 1 \times 1'$  centrée sur la source. Dans les quatre premiers graphes, le trait vertical correspond au résultat de l'optimisation (ie. au minimum de  $|\gamma|\sigma$ ). Dans le cinquième graphe, le trait vertical en tirets longs représente le rayon d'ouverture et les traits verticaux en tirets courts délimitent la région finalement retenue pour estimer la valeur du continu. Le rayon est donné en pixel, la taille d'un pixel étant de  $1.22''$ .

d'ouverture. Cette fonction retourne également les erreurs sur le flux qui sont exprimées de la façon suivante :  $\sigma_{\text{tot}} = \sqrt{\sigma_1^2 + \sigma_2^2 + \sigma_3^2}$  où :

- $\sigma_1$  est l'erreur due au bruit de photons. On peut l'estimer à partir de la carte d'erreur en sommant quadratiquement les erreurs des pixels contenus dans l'ouverture de rayon  $r_{\text{ap}}$ .
- $\sigma_2$  est l'erreur induite par la non uniformité du fond. En notant  $\sigma$  l'écart-type de la distribution des pixels de la couronne, on a  $\sigma_2 = \sqrt{A_{\text{ouverture}}/A_{\text{pixel}}}$   $\sigma$  où  $A_{\text{pixel}}$  est la surface d'un pixel.
- $\sigma_3$  est l'erreur sur l'estimation de la valeur du fond :  $\sigma_3 = \sqrt{\frac{A_{\text{ouverture}}^2}{A_{\text{couronne}}}}$   $\sigma$ . En pratique cette erreur est négligeable devant  $\sigma_2$  si on estime le fond sur un nombre de pixels suffisant (c'est-à-dire plus grand que le nombre de pixels à l'intérieur de l'ouverture).

Enfin, il est possible, étant donné une carte, de déterminer son seuil de détection. Pour cela, on mesure le flux par photométrie d'ouverture à un grand nombre de positions aléatoires. L'écart-type de la distribution des flux obtenus donne le seuil de détection à  $1-\sigma$ . Comme on tombe parfois sur des sources, l'histogramme des flux mesurés est biaisé vers les grands flux et il convient de rejeter les outliers avant d'estimer l'écart-type. Un exemple est donné figure II.5. Cette méthode présente l'avantage de tenir compte de toutes les sources d'incertitude (bruit de photons, bruit de confusion, erreur de photométrie...). On considèrera, en général, les sources détectées si leur flux est supérieur à 3 fois le seuil de détection précédemment estimé.



**Figure II.5** – Histogramme des flux mesurés à 10000 positions aléatoires dans le champ CDF-S sur la carte MIPS 70  $\mu$ m. La ligne en tirets rouge est une gaussienne d'écart-type  $\sigma = 3.3$  mJy.

Le principal avantage de cette technique de photométrie est qu'elle est très robuste et permet de mesurer le flux de sources faibles, voire à la limite de détection ( $2-3\sigma$ ). Cependant, elle n'est pas adaptée aux cas où la confusion due à la forte densité de sources domine et où certaines sources peuvent alors être très proches. En effet, si le rayon d'ouverture est trop grand, on intégrera une partie du flux de la source voisine, ce qui n'est pas souhaitable. La technique d'ajustement de PSF présentée au paragraphe suivant permet de se sortir de ce cas de figure.

#### • Calibration en flux

Les cartes IRAC provenant de l'archive Spitzer sont déjà calibrées et elles sont en MJy/sr. En revanche, les cartes MIPS réduites avec le DAT (Gordon et al. 2005) ne le sont pas : elles sont en unités «instrumentales». Pour obtenir des cartes en MJy/sr, il faut multiplier

ces dernières par 0.047, 702 et 44.6 à 24, 70 et 160  $\mu\text{m}$  respectivement <sup>5</sup>.

Ensuite, pour obtenir des flux en mJy, il faut multiplier le résultat retourné par la procédure de photométrie d'ouverture par  $10^{-6} \times \left( \frac{\text{pixsize}[\text{''}]}{3600} \frac{\pi}{180} \right)^2 \times 10^{-3}$  où pixsize est la taille d'un pixel.

Enfin, si l'on souhaite obtenir des magnitudes, il suffit d'appliquer la relation :

$$m_\lambda = -2.5 \log \left( \frac{S_\nu}{S_{\nu,\text{ref}}} \right) \quad (\text{II.8})$$

où  $S_{\nu,\text{ref}}$  est le flux de référence à la longueur d'onde considérée et pour le système de magnitude considéré. Dans le système de magnitude AB, le flux de référence ne dépend pas de la longueur d'onde et vaut :  $S_{\nu,\text{ref(AB)}} = 3631$  Jy. En revanche dans le système de magnitude Véga, le flux de référence est égal au flux de l'étoile Véga à la longueur d'onde considérée. Ces flux de référence sont donnés dans la table II.3 pour les sept longueurs d'onde de Spitzer.

Instrument	$S_{\nu,\text{ref(Vega)}} \text{ (Jy)}$
IRAC 3.6 $\mu\text{m}$	280.9
IRAC 4.5 $\mu\text{m}$	179.7
IRAC 5.8 $\mu\text{m}$	115.0
IRAC 8.0 $\mu\text{m}$	64.1
MIPS 24 $\mu\text{m}$	7.14
MIPS 70 $\mu\text{m}$	0.775
MIPS 160 $\mu\text{m}$	0.159

**Table II.3** – Flux de référence pour le système de magnitude Véga pour les sept longueurs d'onde Spitzer (d'après les Handbook d'IRAC et de MIPS).

## II.2.2 Ajustement de PSF

Cette technique, comme son nom l'indique, consiste à ajuster une (ou plusieurs) PSF sur les données. Elle est plus communément appelée *PSF fitting*. Considérons un champ contenant  $N$  sources de flux  $S_i$ . On ajustera alors la fonction suivante aux données (en minimisant les résidus) :

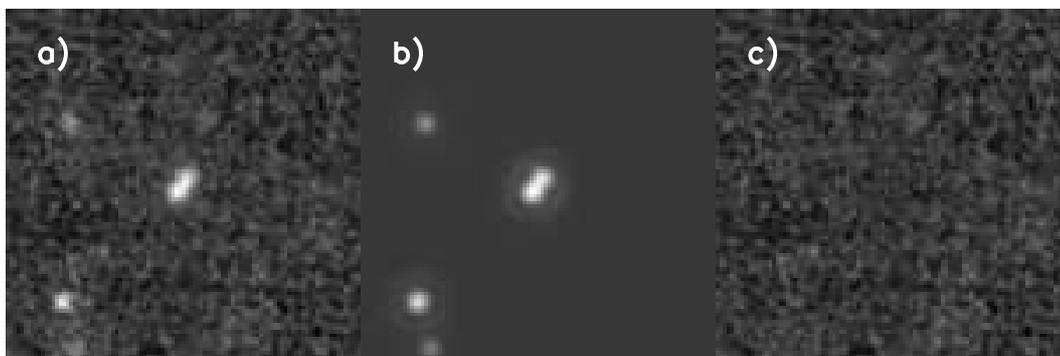
$$f(x, y) = \sum_{i=1}^N S_i \times PSF(x - x_i, y - y_i) + B_i \quad (\text{II.9})$$

où les  $x_i$  et  $y_i$  sont les positions des différentes sources et les  $B_i$  sont le continu sous chaque source. Il y a donc 4 paramètres libres par source à ajuster.

5. <http://ssc.spitzer.caltech.edu/mips/calib/conversion.html>

Comme précédemment, cette méthode requiert la connaissance de la PSF et le point délicat est la normalisation de la PSF qu'on utilise : il faut en effet que celle-ci vérifie  $\iint \text{PSF}(x, y) dx dy = 1$ .

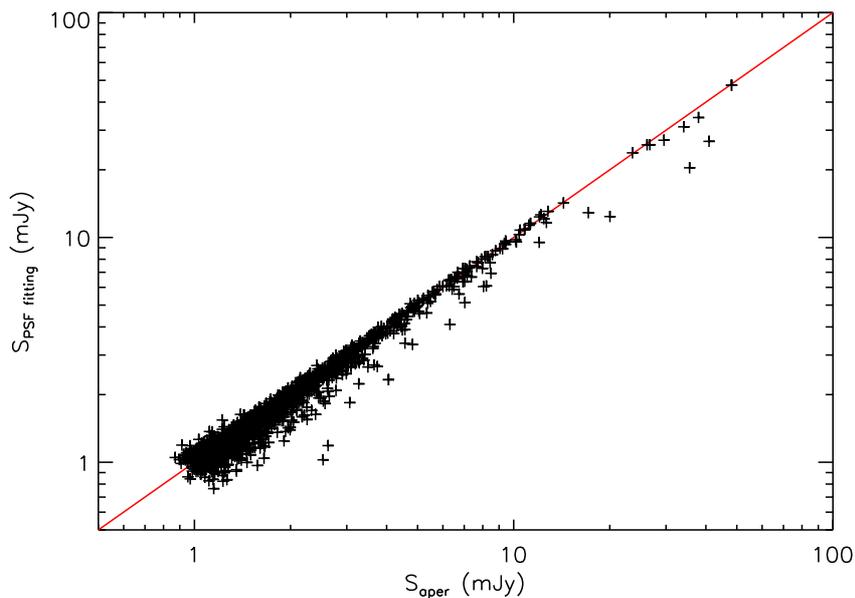
On opère généralement de façon itérative. L'algorithme commence par détecter les sources les plus brillantes dans la carte et ajuste des PSF sur celles-ci et soustrait le meilleur ajustement. Il continue ensuite, en détectant les sources plus faibles et en les soustrayant à leur tour et détecte ainsi de suite des sources de plus en plus faibles. Ceci permet de détecter des sources faibles qui auraient été cachées derrière des sources brillantes. Enfin un ajustement final, reprenant toutes les sources préalablement détectées est effectué. Nous avons utilisé l'algorithme *starfinder*. Un exemple est donné figure II.6. On remarquera en particulier que les deux sources voisines au centre ont été correctement séparées.



**Figure II.6** – Illustration de la méthode d'ajustement de PSF sur une région de  $2' \times 2'$  du champ Boötes à  $24 \mu\text{m}$ . a) Image originale. b) Meilleur ajustement trouvé par l'algorithme *starfinder*. c) Résidus.

Enfin, nous avons comparé les résultats obtenus avec cette méthode de photométrie à ceux obtenus avec la technique de photométrie d'ouverture (figure II.7). L'accord est généralement très bon (13% rms) excepté pour quelques sources qui sont résolues. En effet la méthode d'ajustement de PSF n'est pas du tout adaptée au cas des sources résolues, contrairement à la photométrie d'ouverture qui peut donner un résultat correct, à condition d'intégrer dans un grand rayon (et de rendre ainsi la correction d'ouverture faible).

Nous avons choisi de ne pas utiliser cette méthode de photométrie à  $70$  et  $160 \mu\text{m}$  car les résultats dépendaient énormément de la PSF choisie. De plus, comme le nombre de sources brillantes à ces longueurs d'onde est faible, l'extraction d'une PSF propre sur les données est très délicate. La photométrie d'ouverture dépend elle aussi de la PSF (via la correction d'ouverture) mais cette dépendance nous a semblée plus faible.



**Figure II.7** – Comparaison des flux obtenus avec les deux méthodes de photométrie (photométrie d'ouverture en abscisse, ajustement de PSF en ordonnée) pour un échantillon de 1356 sources  $24 \mu\text{m}$  dans le champ Boötes. La droite rouge est la droite de pente 1 passant par l'origine.

### II.2.3 Correction de couleur

Supposons que la source observée ait la distribution spectrale d'énergie donnée par  $S_\nu(\lambda)$  où  $S_\nu$  est en  $\text{WHz}^{-1} \text{m}^{-2}$ . On observe cette source à travers un filtre  $\mathcal{F}$  de réponse spectrale  $\mathcal{R}(\lambda)$  (proportionnel au nombre d'électrons produits par la détection d'un seul photon de longueur d'onde  $\lambda$ )<sup>6</sup>.

Le signal détecté quand on observe cette source dans le filtre  $\mathcal{F}$  est proportionnel au nombre d'électrons créés, c'est-à-dire à  $\int_0^\infty (\lambda \mathcal{R}(\lambda)) S_\nu(\lambda) d\nu$ . Le coefficient de proportionnalité est le coefficient de calibration qui a été obtenu en mesurant des sources de flux connu. Ce coefficient de calibration dépend donc de la forme du spectre des sources de référence. On utilise généralement la convention suivante : les coefficients de calibration sont donnés pour un spectre de référence de la forme  $\nu S_{\nu, \text{eff}} = \text{cte}$ . Cette convention est celle utilisée pour IRAC ; pour MIPS on utilise comme spectre de référence celui d'un corps noir à 10000 K. Dès lors, comme les sources observées n'ont généralement pas le même spectre que celui de

6. Les transmissions des filtres MIPS sont données en électrons par unité d'énergie sur le site du SSC. En divisant ces courbes par  $\lambda$ , on se ramène à des transmissions en électrons/photon. Les transmissions des filtres IRAC sont déjà en électrons/photon.

référence, il est nécessaire d'appliquer une correction  $K$  dite de couleur dont l'expression est :

$$K = \frac{\int_0^\infty (\lambda \mathcal{R}(\lambda)) \frac{S_\nu(\lambda)}{S_\nu(\lambda_{\text{eff}})} d\nu}{\int_0^\infty (\lambda \mathcal{R}(\lambda)) \frac{S_{\nu,\text{ref}}(\lambda)}{S_{\nu,\text{ref}}(\lambda_{\text{eff}})} d\nu} \quad (\text{II.10})$$

où  $\lambda_{\text{eff}}$  est la longueur d'onde effective du filtre  $\mathcal{F}$ . Cette longueur d'onde effective est définie différemment pour IRAC et pour MIPS :

$$\lambda_{\text{eff, IRAC}} = \frac{\int_0^\infty \mathcal{R}(\lambda) d\lambda}{\int_0^\infty \lambda^{-1} \mathcal{R}(\lambda) d\lambda} \quad \text{et} \quad \lambda_{\text{eff, MIPS}} = \frac{\int_0^\infty \lambda \mathcal{R}(\lambda) d\lambda}{\int_0^\infty \mathcal{R}(\lambda) d\lambda} \quad (\text{II.11})$$

On pourra trouver les valeurs des corrections de couleurs tabulées pour quelques formes de spectre (corps noir de différentes températures et powerlaw avec différents exposants) dans les Handbook des instruments IRAC<sup>7</sup> et MIPS<sup>8</sup>. Ces corrections sont généralement faibles (inférieures à 10%). On pourra enfin consulter le site du SSC ([http://ssc.spitzer.caltech.edu/postbcd/cookbooks/synthetic\\_photometry.html](http://ssc.spitzer.caltech.edu/postbcd/cookbooks/synthetic_photometry.html)) qui a développé une librairie (*Spitzer Synthetic Photometry*) qui permet de calculer ses propres corrections de couleur.

Nous n'avons pas appliqué de correction de couleur sur les flux que nous avons mesurés.

## II.3 Le stacking

Le *stacking* (ou empilement) est une technique très puissante qui consiste à combiner les observations de différentes sources (mais présentant des propriétés physiques semblables) afin d'augmenter le rapport signal sur bruit et d'obtenir des informations moyennes. Nous présenterons dans cette partie, le principe de cette méthode que nous avons beaucoup utilisée, ainsi qu'une méthode pour estimer les erreurs sur le résultat obtenu et enfin le lien entre le stacking et la fonction de corrélation angulaire.

J'ai développé au cours de ma thèse une librairie de fonctions IDL qui code tous les outils présentés dans cette partie et qui sera rendue publique prochainement.

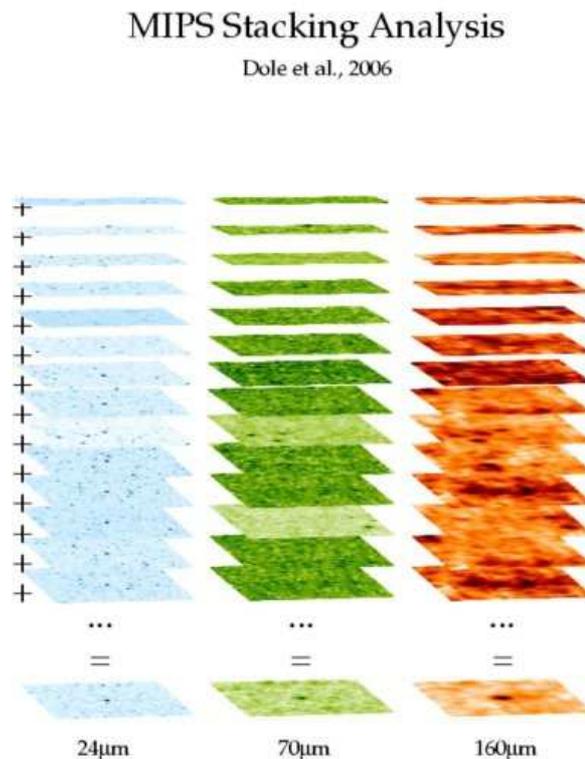
### II.3.1 Principe

L'observation d'un objet à différentes longueurs d'onde permet d'obtenir des informations sur la nature de cet objet et des processus physiques qui y ont lieu. Il est donc très intéressant

7. <http://ssc.spitzer.caltech.edu/irac/dh/iracdatahandbook3.0.pdf>

8. <http://ssc.spitzer.caltech.edu/mips/dh/mipsdatahandbook3.3.pdf>

de combiner les données provenant de longueur d'ondes différentes. Malheureusement, malgré de récentes avancées technologiques, les détecteurs dans certaines gammes de longueur d'ondes restent moins sensibles. C'est en particulier le cas dans l'infrarouge lointain et le domaine sub-millimétrique où viennent en plus s'ajouter des problèmes de confusion. Ainsi, certains objets faibles ne sont détectés qu'à une seule longueur d'onde (par exemple  $24 \mu\text{m}$ ) et pas aux autres ( $70$  ou  $160 \mu\text{m}$ ).

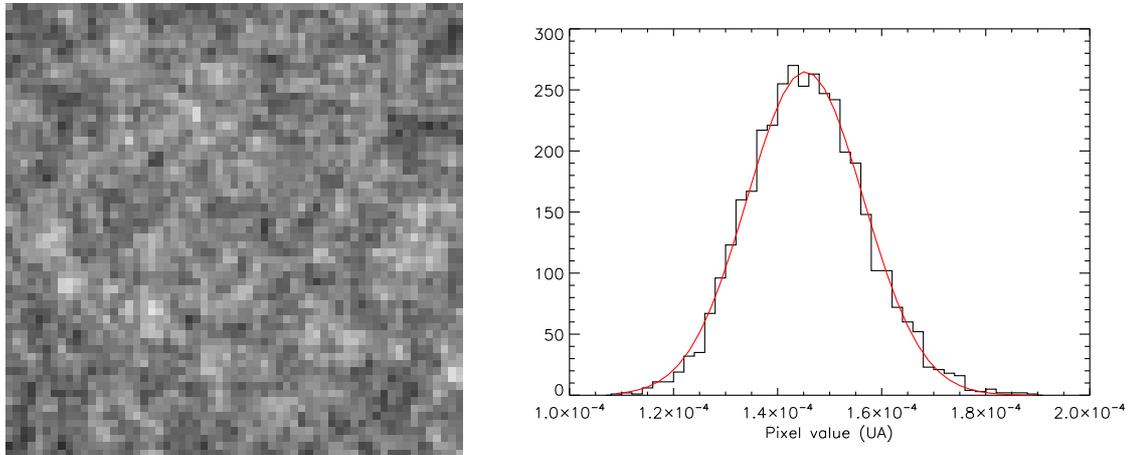


**Figure II.8** – Illustration de la méthode de stacking où on stacke à  $70$  et  $160 \mu\text{m}$  des sources détectées à  $24 \mu\text{m}$ .

Comme le montre la figure II.8, le stacking consiste à extraire des petites images centrées sur les sources aux longueurs d'onde où on ne les détecte pas et ensuite à sommer (ou plutôt faire la moyenne) de toutes ces petites images. Comme le bruit, qui empêche la détection individuelle des sources, est poissonien, le rapport signal sur bruit de la moyenne va augmenter comme  $\sqrt{N}$  où  $N$  est le nombre d'images combinées. On observe finalement la détection d'une source moyenne là où on ne détectait aucun signal. Il est important de noter que cette méthode ne permet pas d'obtenir d'informations individuelles sur les sources. Il convient alors de sélectionner correctement l'échantillon initial à stacker afin que les sources soient physiquement semblables et que le stacking ait un sens.

On peut noter qu'il est envisageable d'effectuer le stacking sur un catalogue de flux plutôt que sur les images elles-mêmes, ce qui serait beaucoup plus rapide. Pour cela, il suffirait de mesurer le flux des sources à stacker (par photométrie d'ouverture) bien que celles-ci

ne soient pas détectées individuellement. L'inconvénient de cette méthode est que les effets d'environnements que nous allons présenter au paragraphe III.4 ne sont pas bien contrôlés et seront probablement plus difficiles à corriger. Nous n'avons pas utilisé cette technique.



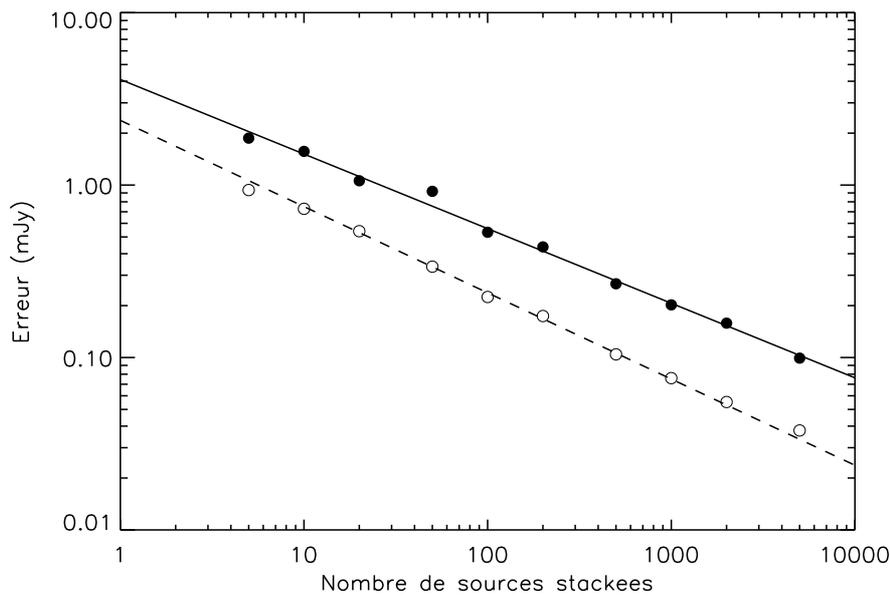
**Figure II.9** – Résultat du stacking à  $70 \mu\text{m}$  dans le champ CDF-S de 1000 images centrées à des positions aléatoires. A gauche : image stackée. A droite : histogramme des valeurs des pixels (en unité arbitraire) de l'image obtenue ; la ligne rouge est une gaussienne.

Un des premiers tests que l'on peut faire est de vérifier qu'on n'a pas de signal lorsqu'on stacke à des positions aléatoires sur la carte : le résultat est montré figure II.9. On remarque en particulier que la distribution des pixels est bien gaussienne.

On peut également vérifier que le bruit diminue bien comme  $1/\sqrt{N}$ . Pour cela nous avons stacké  $N$  images à des positions aléatoires et nous avons mesuré d'une part le bruit sur l'image résultante (en mesurant le flux à des positions aléatoires sur la petite carte et en calculant ensuite l'écart-type des valeurs obtenues) et d'autre part l'écart-type de la distribution des pixels de l'image stackée. Le résultat est montré figure II.10. L'écart-type des pixels diminue bien comme  $1/\sqrt{N}$  alors que l'erreur photométrique diminue quant à elle légèrement moins rapidement (la pente est de  $-0.43$ ) car elle tient compte à la fois du bruit de photons (qui diminue bien comme  $N^{-1/2}$ ) mais aussi de l'erreur sur l'estimation de la valeur du ciel.

Nous avons également testé l'effet d'un rééchantillonnage avant de faire le stacking. Celui-ci consiste à réinterpoler (de façon bilinéaire) chaque petite image extraite sur une grille décalée d'une fraction de pixel par rapport à la grille initiale de façon à ce que la source soit au centre du pixel central<sup>9</sup>. Pour faire ce test, nous avons distribué aléatoirement 1000 sources sur une carte. Nous avons pris des PSF de type *top-hat* de rayon  $r_0$ , c'est-à-dire que le signal vaut 1 à l'intérieur de ce rayon et est nul à l'extérieur. Nous avons ensuite stacké toutes ces sources, avec et sans rééchantillonnage et nous avons comparé les profils obtenus à ceux de la PSF. Nous avons fait ceci pour différentes valeurs du rayon  $r_0$  (1.5, 2, 2.5, 3, 5 et 8 pixels) (voir figure II.11). De façon assez surprenante, on constate que le ré-échantillonnage

9. Les images extraites sont carrées et ont un nombre de pixels impair



**Figure II.10** – Variation de l'erreur du stacking en fonction du nombre d'images stackées. Les points pleins correspondent à l'erreur photométrique (en mJy) et les points vides à la dispersion des pixels de l'image stackée (en unité arbitraire et décalé de façon arbitraire). La droite en trait continu est le meilleur ajustement des points pleins : on trouve une pente de  $-0.43$ . La droite en tiret est le meilleur ajustement des points vides et la pente a été fixée à  $-0.5$ . Cette simulation a été effectuée dans le CDF-S à  $70 \mu\text{m}$ .

n'aide pas à reconstituer le profil de la PSF mais a tendance à diluer celui-ci. En effet, le ré-échantillonnage par interpolation bilinéaire (ou linéaire) lisse le signal. Il n'est donc pas utile de ré-échantillonner l'image avant de faire le stacking.

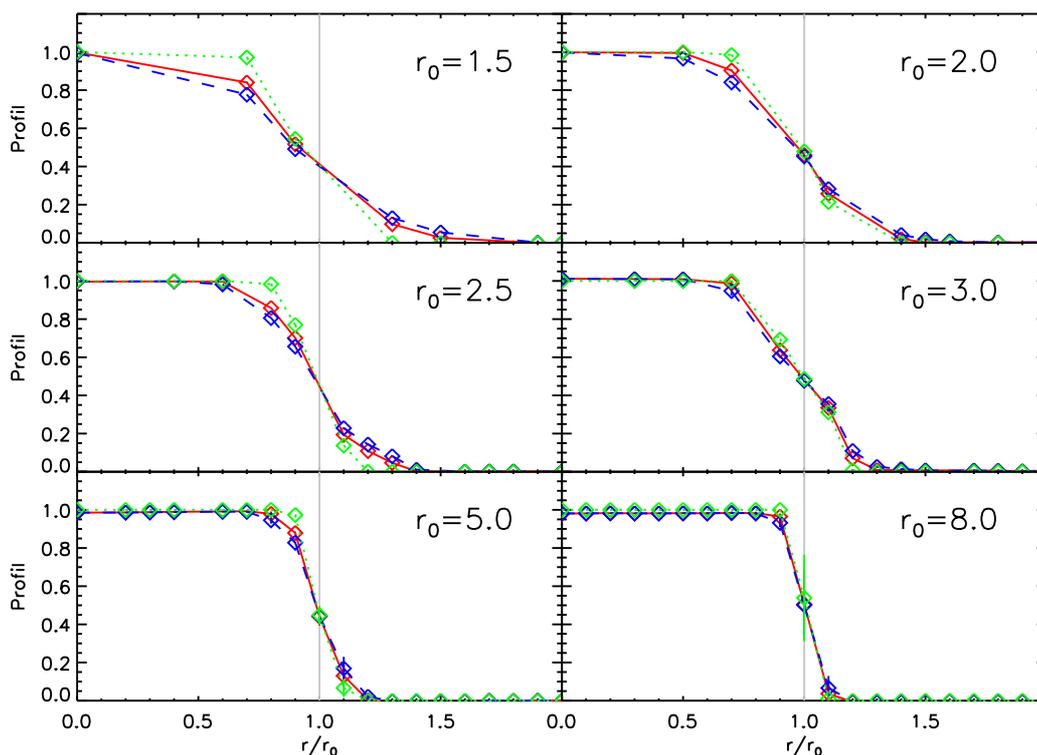
Enfin il arrive parfois que le bruit ne soit pas uniforme sur la carte, les bords du champ ayant été observés moins longtemps par exemple. On peut alors souhaiter pondérer le stacking par une carte de poids  $w = 1/\sigma^2$  où  $\sigma$  est la carte de bruit. La valeur du pixel repéré par les coordonnées  $i$  et  $j$  de l'image stackée  $S$  est alors donnée par :

$$S_{ij} = \frac{\sum_k w_{k,ij} M_{k,ij}}{\sum_k w_{k,ij}} \quad (\text{II.12})$$

où les  $M_k$  sont les petites images que l'on stacke et les  $w_k$  les cartes de poids associées.

### II.3.2 Estimation des erreurs avec le bootstrap

Le terme *bootstrap* fait référence à une légende allemande dans laquelle le baron Münchhausen racontait qu'il s'était tiré lui-même d'un marécage en se tirant par la boucle (*strap*)



**Figure II.11** – Effet du re-échantillonnage lors du stacking avec différents rayons  $r_0$  de PSF. On compare le profil des images stackées obtenues (avec re-échantillonnage en tiret bleu ; sans re-échantillonnage avec la ligne continue rouge) au profil attendu (ligne pointillée verte).

de ses bottes. Le bootstrap est une méthode statistique qui permet d'obtenir l'incertitude sur n'importe quelle grandeur estimée sur une distribution en n'utilisant rien d'autre que la distribution elle-même. Cette méthode, très gourmande en temps de calcul, a été introduite par Efron au début des années 80 (Efron 1982, par exemple).

Supposons que l'on ait un échantillon  $\mathbf{x}$  de  $N_s$  éléments que l'on notera  $\mathbf{x} = (x_1, x_2, \dots, x_{N_s})$ . On calcule sur cet échantillon un estimateur  $\mathcal{E}(\mathbf{x})$  et on souhaite déterminer l'incertitude  $\sigma_{\mathcal{E}(\mathbf{x})}$  sur cet estimateur. Pour cela, on construit  $N_b$  échantillons de bootstrap  $\mathbf{x}_k^*$  de la façon suivante : chaque échantillon est constitué de  $N_s$  éléments tirés *avec remise* parmi les éléments de  $\mathbf{x}$ . C'est à dire qu'un échantillon de bootstrap contient certains éléments en double, d'autres en triple... mais ne contient pas tous les éléments de  $\mathbf{x}$ . On calcule alors l'estimateur pour chacun des  $\mathbf{x}_k^*$  :  $\mathcal{E}(\mathbf{x}_k^*)$ . On calcule enfin l'écart-type  $\sigma_b$  :

$$\sigma_b = \frac{1}{N_b - 1} \sum_{k=1}^{N_b} (\mathcal{E}(\mathbf{x}_k^*) - \bar{\mathcal{E}})^2 \quad \text{où} \quad \bar{\mathcal{E}} = \frac{1}{N_b} \sum_{k=1}^{N_b} \mathcal{E}(\mathbf{x}_k^*) \quad (\text{II.13})$$

qui converge vers  $\sigma_{\mathcal{E}(\mathbf{x})}$  quand  $N_b \rightarrow \infty$ .

Dans le cas du stacking, l'échantillon est composé des  $N_s$  sources que l'on veut stacker et l'estimateur est la photométrie sur l'image stackée. Le bootstrap revient donc à reprendre tout

le processus du stacking avec un échantillon initial légèrement modifié et étudier comment le résultat final (c'est-à-dire le flux moyen mesuré) s'en trouve modifié. Ceci est fait  $N_b$  fois. Cette estimation de l'incertitude contient donc des informations à la fois sur la distribution de flux sous-jacentes des sources stackées, sur le bruit instrumental mais aussi sur toutes les erreurs introduites au cours des processus de stacking (ré-échantillonnage par exemple) et de photométrie.

Notons  $S_k$  les flux des  $N_s$  sources que l'on stacke et supposons que ceux-ci sont issus d'une loi de probabilité  $\mathcal{P}(S)$  caractérisée par ses moments  $\mu_i$  d'ordre  $i$  :

$$\mu_1 = \int S d\mathcal{P} \quad \text{et} \quad \mu_i = \int (S - \mu_1)^i d\mathcal{P}, \quad i \geq 2 \quad (\text{II.14})$$

On note  $M_i$  les moments empiriques d'ordre  $i$  :

$$M_1 = \frac{1}{N_s} \sum_{k=1}^{N_s} S_k \quad \text{et} \quad M_i = \frac{1}{N_s} \sum_{k=1}^{N_s} (S_k - M_1)^i, \quad i \geq 2 \quad (\text{II.15})$$

qui convergent vers les moments  $\mu_i$  de la distribution de probabilité quand  $N_s \rightarrow \infty$ . Chaque flux  $S_k$  est mesuré avec une erreur  $\epsilon_k$ . Cette erreur peut avoir de nombreuses origines (bruit de photons, erreur lors de la photométrie ou lors du stacking...). On obtient alors les flux  $S'_k = S_k + \epsilon_k$ . On supposera que les  $\epsilon_k$  suivent une loi de probabilité gaussienne de moyenne nulle et d'écart-type  $\sigma_{\text{noise}}$ . On note également  $\sigma_\epsilon$  l'écart-type des  $\epsilon_k$  qui converge vers  $\sigma_{\text{noise}}$  quand  $N_s \rightarrow \infty$ . On note  $M'_k$  les moments empiriques de l'échantillon des flux mesurés :

$$M'_1 = \frac{1}{N_s} \sum_{k=1}^{N_s} S'_k \quad \text{et} \quad M'_i = \frac{1}{N_s - 1} \sum_{k=1}^{N_s} (S'_k - M'_1)^i, \quad i \geq 2 \quad (\text{II.16})$$

On peut en particulier montrer que :

$$M'_2 = M_2 + \sigma_\epsilon^2 + 2 \text{cov}(S_k, \epsilon_k) \quad (\text{II.17})$$

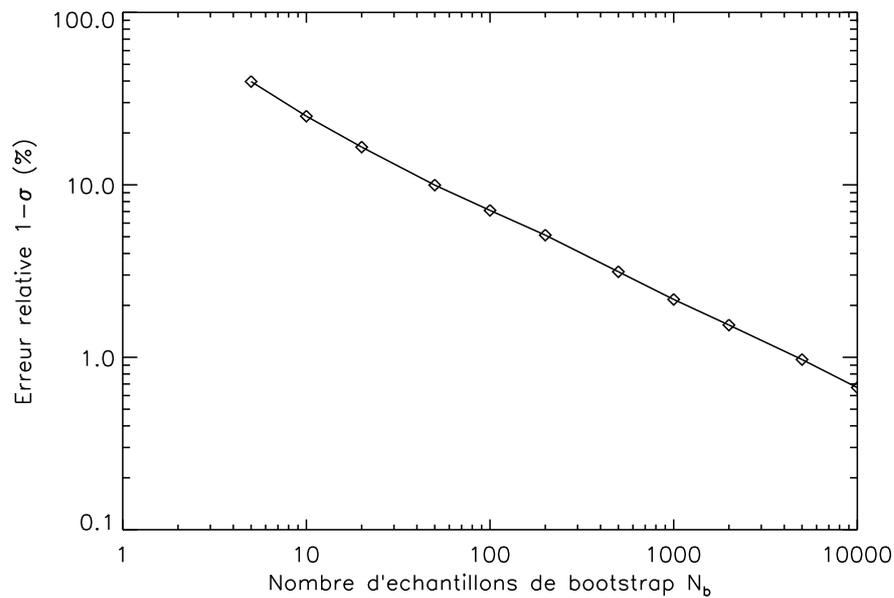
où  $\text{cov}(S_k, \epsilon_k)$  est la covariance des échantillons  $(S_k)_k$  et  $(\epsilon_k)_k$ . La variance empirique de l'échantillon est donc due aussi bien à la distribution intrinsèque des sources qu'aux erreurs de mesure. L'erreur  $\sigma_b$  mesurée via le bootstrap converge vers  $\sqrt{\frac{M'_2}{N_s}}$  quand  $N_b \rightarrow \infty$  et la variance empirique  $M'_2$  converge elle-même vers  $\mu_2 + \sigma_{\text{noise}}^2$  quand  $N_s \rightarrow \infty$ . Ceci peut-être résumé de la façon schématique suivante :

$$\sqrt{N_s} \sigma_b \xrightarrow{N_b \rightarrow \infty} \sqrt{M'_2} \xrightarrow{N_s \rightarrow \infty} \sqrt{\mu_2 + \sigma_{\text{noise}}^2} \quad (\text{II.18})$$

En plus d'estimer l'incertitude sur le flux moyen, le bootstrap permet donc, à condition que l'échantillon contienne suffisamment de sources et que l'on soit capable d'estimer de façon suffisamment précise  $\sigma_{\text{noise}}$ , de remonter à la distribution intrinsèque des sources.

Une question qui se pose alors naturellement est de savoir comment choisir le nombre d'échantillons de bootstrap  $N_b$ . Nous avons étudié cette question avec Mathilde Jauzac au

cours de son stage de M2. Pour cela nous avons généré un échantillon gaussien<sup>10</sup> de  $N_s$  éléments de variance empirique  $M$ . Nous avons ensuite déterminé l'erreur sur la moyenne donnée par le bootstrap avec un nombre d'échantillon de bootstrap  $N_b$ . On obtient alors l'erreur relative  $\epsilon = \frac{\sqrt{N_s}\sigma_b - M}{M}$ . Nous avons effectué 1000 réalisations de cette expérience et déterminé l'écart-type des erreurs relatives. Ceci a été fait pour différentes valeurs de  $N_s$  et de  $N_b$ . Cette erreur relative ne dépend pas du nombre de sources  $N_s$  et vaut approximativement  $\frac{1}{\sqrt{2N_b}}$  (voir figure II.12). Ainsi  $N_b = 5000$  permet d'obtenir l'incertitude sur la moyenne avec une précision de 1% ( $1-\sigma$ ).



**Figure II.12** – Erreur relative de l'incertitude  $\sigma_b$  déterminée avec le bootstrap en fonction du nombre d'échantillons de bootstrap  $N_b$  ( $\sigma_b$  est comparé à  $\sqrt{M'_2}$ ). Cette erreur ne dépend pas du nombre de sources  $N_s$ .

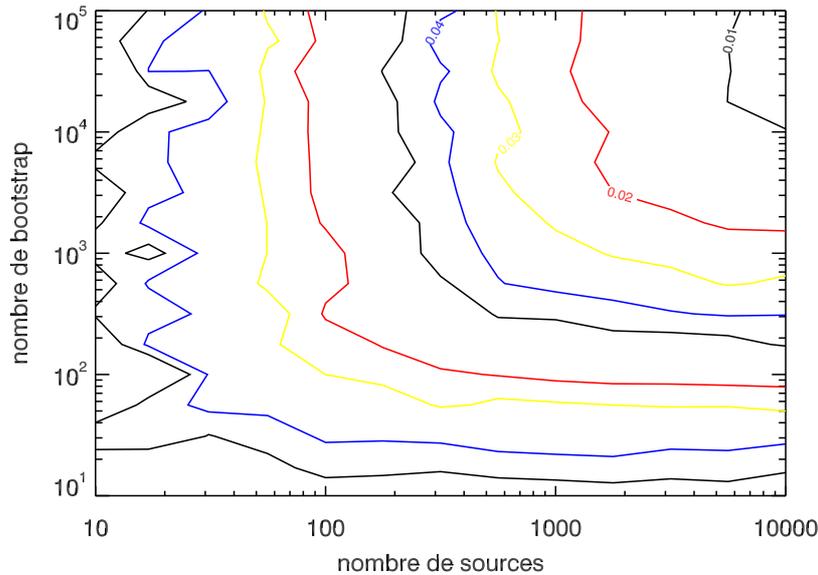
Cependant, si on cherche à déterminer la dispersion intrinsèque de la population  $\sigma_{\text{pop}} = \sqrt{\mu_2}$ , on va être limité par le nombre de sources  $N_s$  de notre échantillon. En effet, l'écart-type empirique  $\sqrt{M'_2}$  d'un échantillon de  $N_s$  sources, est une bonne approximation de  $\sigma_{\text{pop}}$  à un certain pourcentage près donné par

$$\eta = \frac{1}{2} \sqrt{\frac{\alpha - 1}{N_s} - \frac{2(\alpha - 2)}{N_s^2} + \frac{\alpha - 3}{N_s^3}} \quad \text{où} \quad \alpha = \frac{\mu_4}{\mu_2^2} \quad (\text{II.19})$$

Il ne servira donc plus à rien d'augmenter indéfiniment  $N_b$  pour obtenir  $\sqrt{M'_2}$  avec une très bonne précision puisque cette grandeur ne restera qu'une valeur approchée de  $\sigma_{\text{pop}}$  à cause du nombre limité de sources dans l'échantillon. On pourra dans ce cas considérer qu'il ne

10. Nous avons vérifié que les résultats qui suivent ne sont pas modifiés si l'échantillon n'est pas gaussien.

sert à rien d'estimer l'écart-type empirique à mieux que  $\sim \eta/3\%$  puisque celui-ci n'est qu'une approximation de  $\sigma_{\text{pop}}$  à  $\eta\%$  près. On peut alors en déduire un nombre optimal d'échantillons de bootstrap  $N_{b,\text{optimal}} \sim \frac{5}{\eta^2}$ , si on est intéressé par la mesure de  $\sigma_{\text{pop}}$ . Dans le cas où la distribution intrinsèque des sources est gaussienne, on a  $\alpha = 3$  et le nombre de bootstraps optimal est  $\sim 10N_s$ . Ceci est illustré figure II.13.



**Figure II.13** – Erreur relative de l'incertitude  $\sigma_b$ , déterminée avec le bootstrap en fonction du nombre d'échantillons de bootstrap  $N_b$  ( $\sigma_b$  est comparé à  $\mu_2 + \sigma_{\text{noise}}^2$ ). Les courbes de niveau représentent respectivement 1, 2, 3, 4, 5, 8, 10, 15 et 20% d'erreur. Extrait du rapport de stage de M2 de Mathilde Jauzac.

Enfin, on peut noter qu'on utilise parfois une méthode voisine du bootstrap : le *jackknife*. Celle-ci consiste à rejeter une source de l'échantillon  $\mathbf{x} = (x_1, x_2, \dots, x_{N_s})$  pour obtenir un échantillon  $\mathbf{x}_j$  de  $N_s - 1$  sources. On calcule alors les  $N_s$  valeurs  $\mathcal{E}(\mathbf{x}_j)$  de l'estimateur  $\mathcal{E}$  puis l'écart-type  $\sigma_{\text{jackknife}}$  des valeurs obtenues. Cet écart-type est un estimateur de l'incertitude sur  $\mathcal{E}(\mathbf{x})$ . Cette méthode est en général moins précise (le biais et la variance sont plus grands) que le bootstrap.