Processus ponctuel de Gibbs et inférence Bayésienne pour réduire les biais observationnels : cas des catalogues de vitesse de galaxies

Jenny G. SORCE[1,2,3] Radu S. STOICA[4] Elmo TEMPEL[5,6]

[1]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[2]Université Paris-Saclay, CNRS, Institut d'Astrophysique Spatiale, 91405, Orsay, France
[3]Leibniz-Institut für Astrophysik, An der Sternwarte 16, 14482 Potsdam, Germany
[4]Université de Lorraine, CNRS, IECL, Inria, F-54000 Nancy, France
[5]Tartu Observatory, University of Tartu, Observatooriumi 1, 61602 Tõravere, Estonia
[6]Estonian Academy of Sciences, Kohtu 6, 10130 Tallinn, Estonia

Résumé – Les données récoltées en astrophysique sont statistiquement biaisées. Nous considérons le problème inverse qui consiste à retrouver un catalogue de vitesses particulières radiales de galaxies, aux biais statistiquement réduits, à partir de mesures de décalage spectral vers le rouge et de module de distance. Après avoir construit un processus ponctuel dont la maximisation de la densité de probabilité minimise les biais, nous utilisons un recuit simulé pour trouver les maxima correspondants aux réalisations du catalogue aux biais minimisés. La technique est calibrée, puis appliquée sur des catalogues synthétiques issus de simulations cosmologiques, avant d'être employée sur des catalogues observationnels. Ces catalogues serviront à des études cosmologiques.

Abstract – Collected data in astrophysics are statistically biased. We consider the inverse problem which consists in finding back a catalog of galaxy radial peculiar velocities, with statistically reduced biases, from redshift and distance modulus measurements. After constructing a point process whose probability density maximization minimizes biases, we use a simulated annealing to find maxima corresponding to bias-minimized realizations of the catalog. We calibrate the technique, then apply it to synthetic catalogs built from cosmological simulations before using it on observational catalogs. These catalogs will be used for cosmological studies.

1 Introduction

Les données récoltées en astrophysique sont collectivement sujettes à des biais difficiles à maîtriser, d'autant que la vérité terrain n'est pas connue. La fonction de vraisemblance ne s'écrit pas simplement. De nombreux priors sur le format des catalogues de données sont requis. Ces hypothèses sont difficiles à justifier tant les catalogues sont constitués par une collection disparate de données (différents instruments d'observation, estimateurs de propriétés, etc.).

Les catalogues de vitesse particulière radiale[1] des galaxies, utilisés pour, par exemple, dériver les paramètres cosmologiques ou construire les cartes de la distribution de matière, ne sont pas exempts de ces biais. Ces biais sont notamment dus à une probabilité croissante, avec la distance à l'observateur, de sous-estimer les distances des galaxies plutôt que de les sur-estimer, ainsi qu'à une plus grande probabilité de positionner de manière erronée une galaxie d'une zone sur-dense dans une zone sous-dense que l'inverse. Les vitesses résultant de ces mesures de distances ont une distribution totalement asymétrique et aplatie à l'encontre de la Gaussienne attendue théoriquement. Leurs catalogues constituent notre exemple.

Le but est de construire un modèle qui permette de proposer un catalogue, aux effets de biais statistiquement réduits, tout en minimisant le nombre d'hypothèses, sous la forme de priors, quant aux données constituant le catalogue initial (comme c'est le cas dans les modèles utilisés jusqu'à présent, par exemple [2, 8]). Le catalogue initial est constitué d'un ensemble de n (fini et fixé) galaxies dont une mesure du décalage spectral vers le rouge (ci-dessous 'redshift') et du module de distance (différence entre magnitudes mesurée et absolue) avec leur incertitude sont disponibles.

Après avoir rappelé les équations reliant les propriétés des galaxies, la deuxième partie présente la construction du modèle proposé. Tandis que la troisième partie décrit la technique employée pour maximiser la densité de probabilité du modèle, la dernière en illustre les résultats en l'appliquant aux catalogues synthétiques de galaxies, construits à partir de simulations cosmologiques, et aux catalogues observationnels. Avant de conclure, une reconstruction des champs de vitesse et de sur-densité de l'Univers local[2] est proposée comme illustration de l'utilité de ces catalogues aux biais minimisés.

2 Processus ponctuel de Gibbs pour minimiser les biais

2.1 Propriétés des galaxies

Dans le modèle cosmologique standard, dit ΛCDM, les relations entre les propriétés des galaxies sont les suivantes. Il est possible de mesurer pour chaque galaxie :

[1]Vitesse de la galaxie dans l'Univers, due uniquement à la gravitation, c'est-à-dire expansion soustraite. Elle est radiale quand elle est mesurée, selon la ligne de visée, par nous, observateur.

[2]Volume s'étendant sur quelques centaines de Mégaparsecs (Mpc) - unité de mesure de distance - centré sur nous par définition.

• **Le module de distance,** μ,

• **Le redshift observationnel,** $\mathbf{z}_{obs}$

Nous cherchons :

• **La distance de luminosité,** d_{lum} qui s'obtient à partir de μ :

$$\mu = 5\log_{10}(d_{lum}\ (\mathrm{Mpc})) + 25 \tag{1}$$

• **Le redshift cosmologique,** $\mathbf{z}_{cos}$ qui se dérive de la distance :

$$d_{lum} = (1 + z_{cos}) \int_0^{z_{cos}} \frac{c_l dz}{H_0 \sqrt{(1+z)^3 \Omega_m + \Omega_\Lambda}} \tag{2}$$

où H_0 est la constante de Hubble, c_l la vitesse de la lumière, Ω_m et Ω_Λ sont les paramètres cosmologiques correspondant à la matière (noire incluse) et à l'énergie sombre respectivement.

• **La vitesse particulière radiale,** v_{pec} qui combine les redshifts :

$$v_{pec} = c_l \frac{z_{obs} - z_{cos}}{1 + z_{cos}} \tag{3}$$

2.2 Densité de probabilité

Nous dénotons par la suite z_{obs}=$\{z_{obs,\,i}\}_{1\leq i\leq n}$, μ=$\{\mu_{\,i}\}_{1\leq i\leq n}$, et σ=$\{\sigma_{\,i}\}_{1\leq i\leq n}$, l'ensemble des redshifts observationnels, des modules de distance et de leurs incertitudes de mesure[3] où n est le nombre de galaxies-points. Considérant l'Univers local comme un volume fini, l'algorithme, prenant en entrée un ensemble de n (z_{obs},μ_{obs},σ_{obs}), n fixé, doit fournir, en sortie, un nouvel ensemble de n ($\tilde{\mu}_i$,$\tilde{\sigma}_i$) où $\tilde{\sigma}_i$ est obtenu avec une fonction explicitée ci-dessous. Le catalogue est assimilé à un processus ponctuel de Gibbs. Nous construisons la densité de probabilité, p associée au processus, qui prend en compte les effets des biais susmentionnés. Une réalisation du processus, qui maximise p, correspond à un catalogue aux effets de biais statistiquement minimisés :

$$p(\mu|\mathbf{d}, c) \propto \mathrm{e}^{-U(\mu|\mathbf{d},c)} = \mathrm{e}^{-U_1(\mu|\mathbf{d},c) - U_2(\mu|\mathbf{d},c)} \tag{4}$$

où **d** est le jeu de données de n galaxies-points (z_{obs},μ_{obs},σ_{obs}), c=$\{c_i\}_{i\in\mathbb{N}}$ un ensemble de paramètres constants positifs et U la fonction énergie. Cette fonction dépend des données observées (principalement U_1) et prend en compte le positionnement relatif des galaxies (terme U_2) [5, 1]. Notre construction s'apparente au produit d'une vraisemblance et d'un prior pour déterminer une loi a posteriori.

2.2.1 Terme U_1

Le terme U_1 s'assure de la plausibilité du module de distance d'une galaxie, μ_i, étant donné la valeur initialement mesurée, $\mu_{obs,\,i}$, et son incertitude, $\sigma_{obs,\,i}$, ainsi que la valeur de la vitesse particulière radiale qui lui est associée, $v_{pec,i}(\mu_i, z_{obs,\,i})$:

$$U_1(\mu|\mathbf{d}, c_0, c_1) = \sum_{i=1}^{n} c_0 \frac{(\mu_i - \mu_{obs,\,i})^2}{2\sigma^2_{obs,\,i}} + c_1 \frac{|v_{pec,i}|}{v_{ref}} \tag{5}$$

où c_0 et c_1 sont des constantes dont les valeurs sont données dans le tableau 1, v_{ref} est une valeur de réference fixée à 10,000 km s^{-1}. Les $\sigma_{obs,\,i}$s ne pouvant être nuls que théoriquement, ce cas n'est pas détaillé. Il est à noter que le second terme force la parcimonie et permet ainsi de converger vers la distribution de vitesse Gaussienne théorique attendue.

[3] L'ordre de grandeur de l'incertitude de mesure sur le module de distance est supérieur à celui associé au redshift observationnel, l'incertitude sur ce dernier est négligeable au premier ordre.

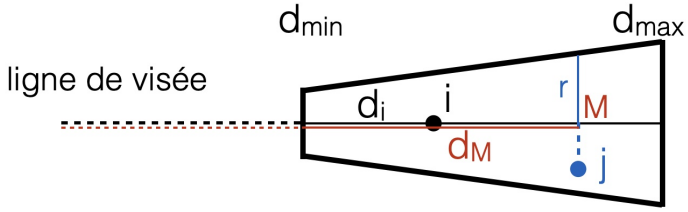

FIGURE 1 : Volume d'interaction en projection 2D pour le point i. Le point j appartient au volume du point i.

2.2.2 Terme U_2

Le terme U_2 favorise les configurations de galaxies avec des valeurs de vitesses particulières plausibles en considérant le catalogue dans son ensemble. Une distribution des galaxies à petites échelles en accord avec la théorie est retrouvée. Pour cela, les vitesses des galaxies considérées comme étant en interaction sont comparées. Il faut introduire :

• **Le volume d'interaction,** V : il dépend de la distance de la galaxie et de son incertitude. La projection 2D de V est montrée sur la Fig. 1 avec $\mathrm{d}_{min/max}$ les distances obtenues avec $\mu_i \pm n_\sigma \sigma_{obs,\,i}$, M la projection du point j sur la ligne de visée du point perturbé i, r est défini par $r = \alpha_{pc} \times d_M$ où d_M est la distance au point M. Les valeurs de n_σ et α_{pc} sont données dans le tableau 1.

• **La fonction de relation entre les vitesses aux petites échelles**, σ_v, est un ajustement polynomial de la relation déterminée à partir de simulations cosmologiques :

$$\sigma_v = \mathrm{a} + \mathrm{b} \times \sigma + \mathrm{c} \times \sigma^2 \tag{6}$$

où a=64, b=70 et c=22 km s^{-1}. Cette fonction détermine la différence moyenne autorisée, σ_v, entre vitesses de galaxies appartenant au même volume d'interaction en fonction de l'incertitude sur le module de distance, σ (reliée au volume par définition). Etant donné la variance cosmique de la relation mesurée entre différentes régions simulées, l'incertitude sur les valeurs des paramètres de l'ajustement est insignifiante.

•**Les fonctions d'interaction positive** h**, totale** f **et absente** q **entre les points** i **et** j :

$$\begin{aligned} h_i &= \sum_{j=1, j\neq i}^{n} \mathbb{1}\{i \sim_s j\} \times \mathbb{1}\{||v_{pec,i}| - |v_{pec,j}|| > \sigma_v\} \times \frac{1}{\sigma_j^2} \\ f_i &= \sum_{j=1, j\neq i}^{n} \mathbb{1}\{i \sim_s j\} \times \frac{1}{\sigma_j^2} \\ q_i &= \mathbb{1}\{\forall j, j \neq i, \ \ i \nsim_s j\} \end{aligned} \tag{7}$$

où ($i \sim_s j$) indique que l'interaction entre i et j existe : ($i \in S_j$ & $j \notin S_i$) ou ($j \in S_i$ & $i \notin S_j$) ou ($j \in S_i$ & $i \in S_j$). Inversement ($i \nsim_s j$) dénote l'absence d'interaction. $\mathbb{1}$ est une fonction indicatrice égale à 1 si la condition est réalisée et valant 0 si elle ne l'est pas. Il est à noter que h_i et f_i dépendent de σ_j avec $j \neq i$.

Finalement, le terme U_2 s'écrit :

$$U_2(\mu|\mathbf{d}, c_2) = \sum_{i=1}^{n} c_2 \left(\frac{h_i}{f_i} + q_i \right) \tag{8}$$

où la valeur de la constante c_2 est donnée dans le tableau 1 et avec la convention $\frac{h_i}{f_i} = 0$ quand $f_i = 0$.

TABLE 1 : Liste des paramètres calibrés avec un catalogue synthétique issu d'une simulation (vérité terrain).

Paramètre	valeur	rôle
n_{sa}	10	nb. itérations du recuit simulé
n_{mh}	4000	nb. itérations de l'échantillonnage
n_σ	1	volume de l'interaction
T_0	1	température initiale
α_{pc}	0.05	élongation du volume
$\sigma_{v'}$	300^2 km^2 s^{-2}	variance de la distribution de v_{pec}
c_0, c_1, c_2	1	constantes

3 Echantillonnage et recuit simulé

Pour trouver une réalisation minimisant la fonction énergie, nous utilisons un algorithme de Metropolis-Hastings, incorporé dans un recuit simulé, comme détaillé ci-dessous.

Algorithme 1 : Minimisation de la fonction énergie

pour $t = 1...n_{sa}$ **faire**
  $T(t) = \frac{T_0}{1+\ln(t)}$
  pour $m = 1...n_{mh}$ **faire**
    pour $i = 1...n$ **faire**
      $\tilde{\mu}_i \Leftarrow$ eq. 11
      $d_{lum,i}$, $v_{pec,i}$, $\tilde{d}_{lum,i}$, $\tilde{v}_{pec,i} \Leftarrow$ eq. 1 to 3
      $\tilde{\sigma}_i \Leftarrow$ eq. 10
      $U_D(..., (\mu_i, \sigma_i), ...)$, $U_I(..., (\mu_i, \sigma_i), ...)$, $U_D(..., (\tilde{\mu}_i, \tilde{\sigma}_i), ...)$, $U_I(..., (\tilde{\mu}_i, \tilde{\sigma}_i), ...)$ $\Leftarrow$ eq. 5 & 8
      $p_{\tilde{\mu}_i,\tilde{\sigma}_i}$, $p_{\mu_i,\sigma_i} \Leftarrow$ eq. 4
      α=min$\{1, (p_{\tilde{\mu}_i,\tilde{\sigma}_i}/p_{\mu_i,\sigma_i})^{1/T(t)}\sigma_i/\tilde{\sigma}_i\}$
      $R = U_n[0, 1[$
      if $\alpha \geq R$, $\mu_i = \tilde{\mu}_i$ & $\sigma_i = \tilde{\sigma}_i$
    fin
  fin
fin

Une minimisation de l'énergie correspond à un maximum de la densité de probabilité donc à une réalisation du catalogue aux biais statistiquement réduits, $(\tilde{\mu}, \tilde{\sigma})$:

$$(\tilde{\mu}) = \arg\min\{U(\mu|\mathbf{d}, c)\} \quad (9)$$

avec :

• **les nouvelles incertitudes,** $\tilde{\sigma}_i$**s** telles que :

$$\tilde{\sigma}_i = \sigma_i[(1 - p_{v,i})^2 + p_{v,i}\,(\frac{h_i}{f_i} + q_i))] \quad (10)$$

où p_v est la fonction de répartition donnée par la théorie [4] : $p_v = \frac{1}{\sigma_{v'}\sqrt{2\pi}} \int_{-\infty}^{-|v_{pec}|} \exp(-\frac{v^2}{2\sigma_{v'}^2})dv$ avec $\sigma_{v'}$ obtenue avec des simulations répondant au modèle cosmologique et dont la valeur est donnée dans le tableau 1.

• **les nouveaux modules de distance,** $\tilde{\mu}_i$**s** tirés aléatoirement selon la loi de proposition :

$$\tilde{\mu}_i = \mu_i + U_n[-0.5, +0.5] \times \gamma\ \sigma_i \quad (11)$$

où $U_n[-0.5, +0.5]$ définit un nombre aléatoire tiré uniformément entre -0.5 et 0.5. γ prend des valeurs entre 1 et 3 suivant l'avancée de l'algorithme pour accélérer la convergence.

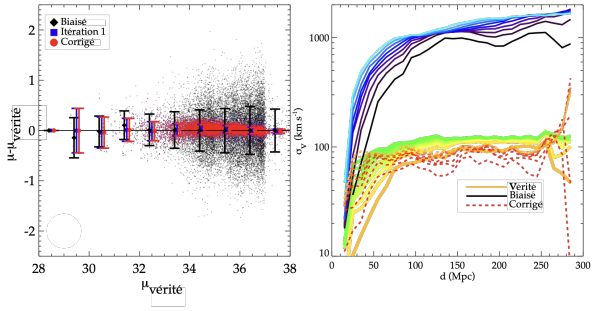

FIGURE 2 : *Gauche* : Différence entre modules de distance vrais et i) biaisés (losanges noirs), ii) après une itération du recuit simulé (carrés bleus), iii) corrigés (cercles rouges) en fonction des vrais. *Droite* : Variance des vitesses aux petites échelles dans des volumes, V (une ligne par élongation) pour un catalogue sans erreur (vérité, lignes continues aux couleurs chaudes), avec des erreurs (biaisé, lignes continues aux couleurs froides), après application de l'algorithme (corrigé, lignes pointillées). La variance est celle des vitesses des galaxies appartenant à un même volume V, d'élongation proportionnelle à une incertitude sur le module de distance de valeur 0,2 (orange, rouge, noir) à 1,8 (vert, orange, bleu) mag.

Le code a, entre autres, été parallélisé pour calculer les termes d'interaction avec openMP et pour perturber, simultanément, plusieurs points sans interaction avérée (éloignés dans l'espace réel) avec MPI. Comme le recuit simulé résulte en *une* réalisation maximisant la densité de probabilité, moyenner plusieurs réalisations permet de réduire les effets stochastiques inhérents à chaque solution individuelle. Des tests montrent que moyenner 5 à 10 réalisations donne des résultats stables. Il faut ∼7500 à 15000 cpu.heures pour obtenir une réalisation moyennée d'un catalogue de 15000+ points.

4 Applications aux catalogues

4.1 Catalogue synthétique

Nous appliquons l'algorithme à des catalogues synthétiques, construits, à partir de simulations cosmologiques, pour reproduire les catalogues observationnels. Les résultats étant similaires pour tous les catalogues testés, ils sont montrés pour l'un d'eux. La Fig. 2 gauche montre que les modules de distances corrigés (cercles rouges et carrés bleus) diffèrent de ceux sans erreur (vérité) d'au maximum ∼0.5 mag contre près de quatre fois plus pour ceux biaisés (losanges noirs). La Fig. 2 droite montre la variance des vitesses aux petites échelles dans des volumes (représentant différentes incertitudes selon le gradient de couleur) à des distances croissantes de l'observateur. La variance est quasiment retrouvée : lignes pointillées (corrigé) vs. continues de couleurs froides (biaisé) par rapport à celles continues (vérité) de couleurs chaudes.

4.2 Catalogue observationnel

L'algorithme est appliqué aux catalogues observationnels. Les résultats sont montrés pour le troisième catalogue du projet Cosmicflows [7]. Ce catalogue contient 15000+ modules de distance. La variance des vitesses aux petites échelles est réduite d'un ordre de grandeur en accord avec les attentes du modèle (cf. Fig. 3).

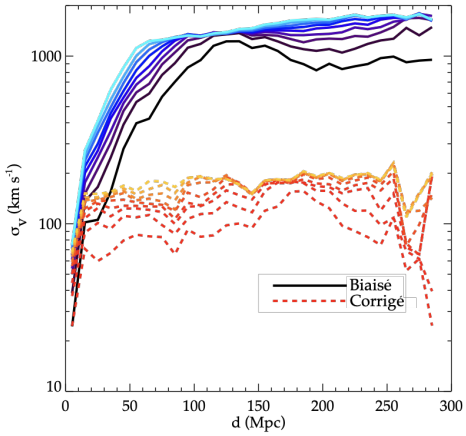

FIGURE 3 : Comme Fig. 2 droite mais pour le catalogue observationnel.

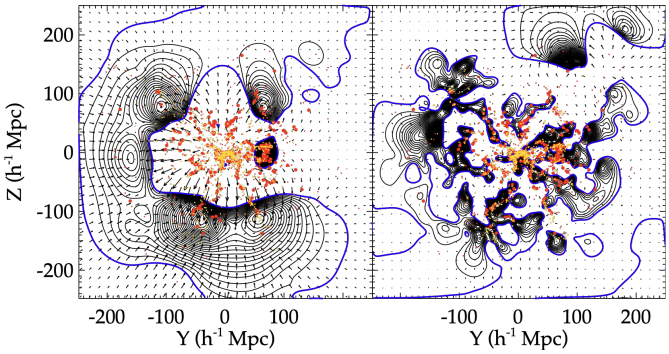

FIGURE 4 : Plan supergalactique YZ d'une reconstruction des champs de surdensité (contours) et de vitesse (flèche) de l'Univers local. La couleur bleue délimite les zones surdenses de celles sous-denses. Les points rouges sont les galaxies isolées du catalogue de redshift 2MRS [3] et les amas et groupes de galaxies de [6]. Les points jaunes correspondent aux galaxies des catalogues biaisé (gauche) et corrigé (droite).

Pour illustrer l'utilité de ce catalogue aux biais minimisés, il est combiné avec une régression au sens des moindres carrés [9] afin de reconstruire les champs de surdensité et de vitesse de l'Univers local. Cette technique est choisie pour son incapacité à gérer les biais. Le plan supergalactique YZ de la reconstruction des champs de surdensité (contours) et de vitesses (flèches) de l'Univers local est montré sur la Fig. 4. Le panneau de gauche (droite) est la reconstruction obtenue avec le catalogue biaisé (corrigé). La couleur bleue dénote la séparation entre zones surdenses et sous-denses. Les points jaunes sont les données des catalogues. Les points rouges correspondent à un catalogue d'amas/groupes de galaxies [6] et de galaxies isolées [3] pour comparaisons. Les différences entre structures reconstruites sont notables : la reconstruction basée sur le catalogue biaisé montre des structures arrondies et le champ des vitesses converge uniformément. Ce sont des effets de biais. Au contraire, la reconstruction obtenue avec le catalogue corrigé donne des structures plus définies et de multiples points de convergence vers des îlots de surdensité.

5 Conclusion

Ce travail propose d'utiliser un processus ponctuel de Gibbs et l'inférence Bayésienne pour minimiser les effets des biais observationnels dans les catalogues de données en astrophysique. Les catalogues de vitesse particulière de galaxies constitue notre cas école. Après avoir construit une densité de probabilité dont la maximisation correspond à des réalisations des catalogues dans lesquels les effets des biais sont minimisés, cette dernière est maximisée par échantillonnage inséré dans un recuit simulé. L'algorithme est testé sur des catalogues synthétiques issus de simulations cosmologiques (vérité terrain connue). Les catalogues de sortie de l'algorithme ont statistiquement des erreurs réduites par rapport à ceux d'entrée. L'algorithme est appliqué sur des catalogues observationnels. Une reconstruction des champs de surdensité et de vitesse de l'Univers local, effectuée avec une régression au sens des moindres carrés, constitue une illustration de l'intérêt d'un tel catalogue aux biais minimisés. L'estimateur ne gérant pas les biais, il est immédiat de constater que les reconstructions, obtenues avec les catalogues corrigés, sont plus fidèles aux attentes que celles basées sur les catalogues biaisés. L'algorithme est adapté au modèle cosmologique standard avec des paramètres cosmologiques fixés. Dans un travail futur, nous nous intéresserons à la variation des paramètres voire à un changement du paradigme cosmologique. Ce type d'algorithme peut aussi être appliqué à tout catalogue de données ponctuelles en astrophysique.

Remerciements

Les auteurs remercient le 'Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu)' et le GENCI (https ://www.genci.fr/) pour avoir financé ce projet à travers la dotation de temps de calcul. Les auteurs remercient les rapporteurs qui ont contribué à améliorer la qualité de cet article. JS est soutenue par l'ANR pour le projet LOCALIZATION (ANR-21-CE31-0019). ET est soutenu par l'ETAg (PRG1006) et l'Europe à travers l'ERDF CoE (TK133). Ce travail a été soutenu par le Programme National de Cosmologie et Galaxies (PNCG) du CNRS/INSU avec l'INP et l'IN2P3, co-financé par le CEA et le CNES.

Références

[1] S. N. CHIU, D. STOYAN, W. S. KENDALL et J. MECKE : Stochastic Geometry and its Applications. *Wiley series in probability and statistics*, 2013.

[2] R. GRAZIANI, H. M. COURTOIS, G. LAVAUX, Y. HOFFMAN, R. B. TULLY, Y. COPIN et D. POMARÈDE : The peculiar velocity field by forward-modelling Cosmicflows-3 data. *MNRAS*, 488(4):5438–5451, Oct 2019.

[3] J. P. HUCHRA, L. M. MACRI, K. L. MASTERS, T. H. JARRETT, P. BERLIND, M. CALKINS, A. C. CROOK et AL. : The 2MASS Redshift Survey : Description and Data Release. *ApJs*, 199:26, avril 2012.

[4] R. K. SHETH et A. DIAFERIO : Peculiar velocities of galaxies and clusters. *MNRAS*, 322:901–917, avril 2001.

[5] R. S. STOICA : Marked point processes for statistical and morphological analysis of astronomical data. *The European Physical Journal Special Topics*, 186:123–165, septembre 2010.

[6] E. TEMPEL, M. KRUUSE, R. KIPPER, T. TUVIKENE, J. G. SORCE et R. S. STOICA : Bayesian group finder based on marked point processes. Method and feasibility study using the 2MRS data set. *A&A*, 618:A81, octobre 2018.

[7] R. B. TULLY, H. M. COURTOIS et J. G. SORCE : Cosmicflows-3. *AJ*, 152:50, août 2016.

[8] A. VALADE, Y. HOFFMAN, N. I. LIBESKIND et R. GRAZIANI : HMC reconstruction from peculiar velocities. *MNRAS*, 513(4):5148–5161, juillet 2022.

[9] S. ZAROUBI, Y. HOFFMAN et A. DEKEL : Wiener Reconstruction of Large-Scale Structure from Peculiar Velocities. *ApJ*, 520:413–425, août 1999.