1 Présentation des données

Dans le chapitre précédent, nous présentions l'instrument SDSS et le relevé eBOSS. Nous présentons ici comment les spectres sont déduits des photons acquis par les CCD, puis comment le spectre d'absorption est reconstruit à partir de chaque spectre.

1 Réduction des données

L'observation de chaque plaque fournit des données en deux dimensions : la première correspond aux différentes longueurs d'onde, la seconde aux différentes fibres (voir figure 1.1). La chaîne de réduction des données transforme ces informations en une liste de spectres. Premièrement, chaque spectre est extrait de l'image acquise par le CCD. Le spectre est étalonné en longueur d'onde à l'aide de lampes à arc. Le flux est calibré en utilisant les spectres des fibres dédiées à des étoiles standards. Puis le fond du ciel, estimé dans chaque demi-plaque grâce aux fibres dédiées, est soustrait à chaque spectre. La variance dans chaque pixel est ensuite estimée. Elle prend en compte le bruit de photon et le bruit de l'électronique. Les pixels affectés par des rayons cosmiques sont rejetés. Enfin, pour chaque objet, toutes les expositions sont ajoutées pour former un seul et même spectre, avec un plus grand rapport signal sur bruit. Le nombre typique d'expositions par objet varie entre 4 et 6.

Les données que nous utilisons dans ce manuscript ont été rendues publiques lors de la seixième publication de données SDSS (DR16 [1]). Elles sont décrites par Ahumada et al. (2019).

2 Le catalogue de quasar

Une fois les spectres extraits, il est important de les classifier afin de construire des catalogues utilisables par les différentes analyses. Pour ce faire, chaque spectre est traité par le pipeline de SDSS, décrit par Bolton et al. (2012). **Plusieurs modèles d'étoiles, de galaxies et de quasars sont ajustés à chaque spectre. Chaque modèle est construit à l'aide d'une analyse en composantes principales (PCA), propre à chaque type d'objet. Pour chaque type d'objet, l'analyse fournit des spectres typiques. Ces derniers sont de complexités croissantes : le premier donne le plus d'information comme par exemple le spectre moyen, les spectres plus complexes donnent des détails comme certaines raies d'émission ou profile d'absorption (#prov garder cette phrase ?). Ces spectres typiques sont pondérés et combinés afin**

1. https ://www.sdss.org/dr16/

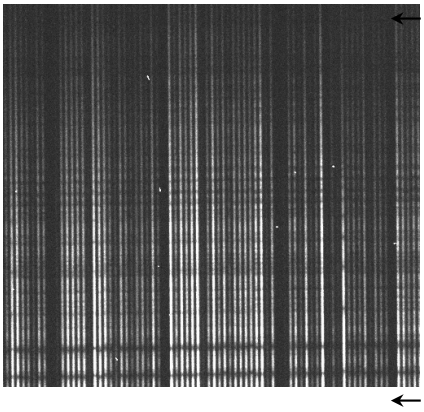

Figure 1.1 – Photographie d'un capteur CCD. L'axe des absysses indique le numéro de la fibre optique. L'axe des ordonnées donne la longueur d'onde observée.

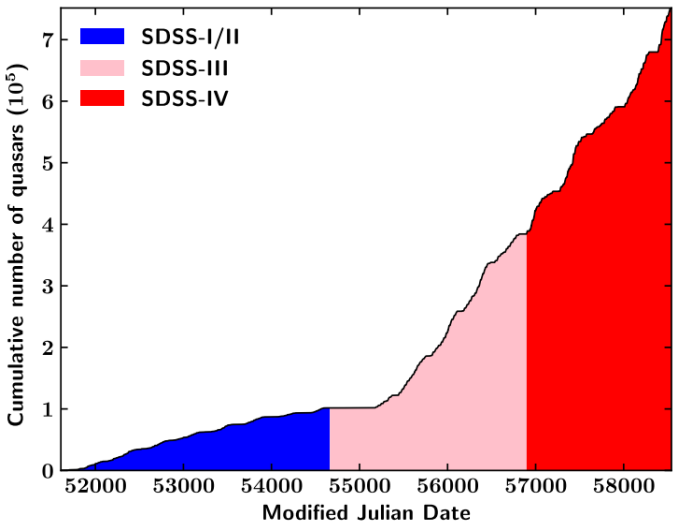

FIGURE 1.2 – Evolution du nombre de quasars observés par les différentes générations de SDSS en fonction du temps. Les générations SDSS I et II correspondent à $MJD < 54663$ (bleu), SDSS III à $54663 \leq MJD < 56898$ (rose) et SDSS IV à $56898 < MJD < 58543$ (rouge). Crédits : #prov Lyke in prep

de produire un spectre simulé qui ajuste aux mieux le spectre à classifier. Une fois tous les modèles ajustés, ils sont triés selon leur χ^2 réduit. Le spectre est alors classifié selon le modèle possédant le plus faible χ^2 réduit. Tous les spectres mesurés par SDSS et classifiés comme quasar constituent l'échantillon *superset*. Il contient 1 440 627 objets.

Certains quasars peuvent être difficiles à classifier à cause d'absorptions intenses comme les DLA ou les BAL [1]. Tous les quasars du relevé BOSS ont été inspectés visuellement (PÂRIS et al. 2016) afin de confirmer la classification du pipeline de SDSS, et d'estimer leur redshifts. Cet échantillon représente 297 301 objets. Cependant, à cause du grand nombre de quasars observés par eBOSS (voir figure 1.2), l'observation visuelle n'a pas pu être effectuée pour tous ces objets. Afin de vérifier la première identification faite par le pipeline de SDSS, une seconde classification est alors effectuée (**CITE:Lyke in prep**). Elle est complétée par l'algorithme `QuasarNET` (BUSCA et BALLAND 2018). Ceci réduit le nombre d'identifications douteuses et donc le nombre d'inspections visuelles requises. A la fin, 0,6 % des spectres, soit 8581 spectres, requièrent une inspection visuelle.

Une fois les spectres classifiés, un sous-échantillon du superset est construit. Il contient 750 426 objets, confirmés comme quasar par la chaîne de traitement précédente. Pour chaque objet, le catalogue fournit plusieurs estimations de redshift. Le pipeline de SDSS produit une première estimation. L'algorithme `QuasarNET` en fournit une seconde. Les spectres inspectés visuellement possèdent une autre estimation. Enfin, l'algorithme `redvsblue` [2] produit plusieurs estimations de redshift, parmi lesquelles figurent `Z_PCA` et `Z_LYAWG`.

Différents algorithmes sont alors appliqués au catalogue, afin d'identifier les DLA et les BAL présents. L'identification utilise l'addition des différentes expositions pour chaque objet. Concernant les DLA, l'algorithme de détection est décrit dans PARKS et al. (2017). Il est appliqué sur les quasars à un redshift $2 \leq$ `Z_PCA` ≤ 6, afin d'avoir suffisamment de pixels dans la zone $900 < \lambda_{\mathrm{RF}} < 1346$ Å. Parmi les 270 315 spectres inspectés, 39 514 DLA ont été identifiés, distribués dans 35 686 spectres. Concernant les BAL, l'algorithme utilisé est très similaire à celui décrit dans GUO et MARTINI (2019).

1. Les BAL (Broad Absorption Line) sont des quasars qui présentent des absorptions intenses au voisinage de leurs raies d'émission. Cette absorption est interprétée comme étant due à un absorbeur dense situé juste devant le quasar. La figure 1.3 présente un spectre de quasar avec un BAL.
2. https ://github.com/londumas/redvsblue

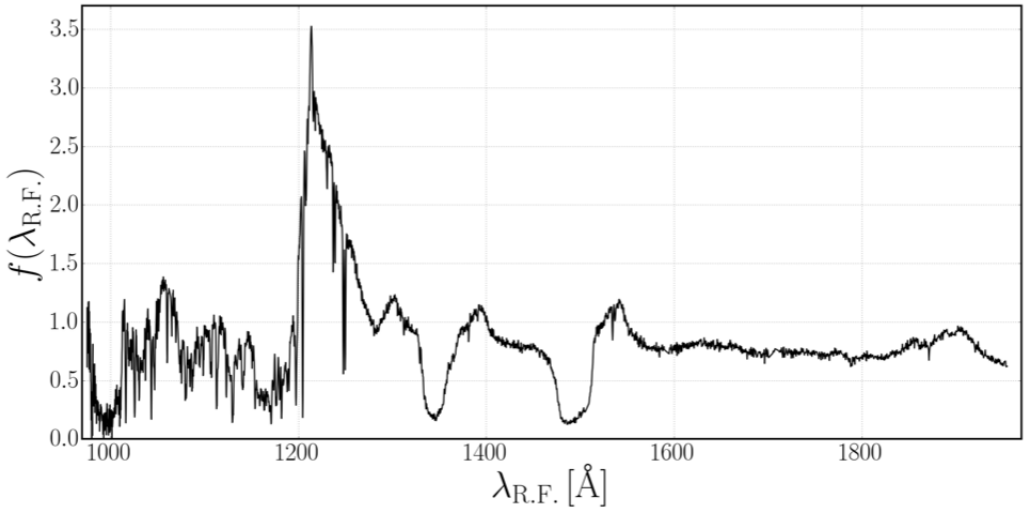

FIGURE 1.3 – Spectre d'un quasar pris par SDSS présentant un BAL. Le spectre est très fortement absorbé pour les longueurs d'onde légèrement inférieures à chaque raie d'émission.

Les BAL sont recherchés dans les spectres ayant un redshift entre 1,57 et 5,6. L'algorithme fournit la probabilité qu'un spectre possède un BAL. Le champ `BAL_PROB` du catalogue indique cette probabilité.

A la fin, le catalogue ainsi construit (DR16Q dans la suite de ce manuscrit) contient 750 426 quasars confirmés. Nous référons le lecteur à l'article **CITE:Lyke in prep** pour davantage d'informations.

L'analyse Lyα des données complètes d'eBOSS (**CITE:dr16**) utilise le catalogue DR16Q. Le redshift des quasars est choisi comme étant `Z_LYAWG`. Les quasars sont sélectionnés avec un redshift $1{,}77 < z \leq 4$. L'échantillon correspondant représente alors 341 468 objets. La distribution en redshift de ces quasars traceurs (ceux utilisés comme traceurs pour la fonction de corrélation croisée Lyα-QSO) est présentée dans le graphique de gauche de la figure 1.4.

3 La sélection des forêts

L'analyse Lyα des données complètes d'eBOSS, dont nous nous servons dans ce manuscrit, est présentée dans l'article **CITE:dr16** Les informations que nous donnons dans la suite de ce chapitre sont tirés de cette article. Nous y référons le lecteur pour davantage d'informations.

Les spectres produits par la chaîne de réduction des données SDSS sont rebinnés : 3 pixels du spectre original, d'une taille $\Delta log_{10}(\lambda) \sim 10^{-4}$, sont combinés en 1 seul pixel d'analyse, d'une taille $\Delta log_{10}(\lambda) \sim 3 \times 10^{-4}$. Ceci est fait afin de réduire le temps de calcul nécessaire pour estimer les fonctions de corrélation. Dans la suite, l'utilisation de "pixel" réfère à ces pixels d'analyse.

L'absorption Lyα est mesurée dans deux régions disctinctes du spectre. La première, dénommée région Lyα, correspond aux longueurs d'onde $1040 \leq \lambda_{\mathrm{RF}} \leq 1200\,$Å, c'est à dire entre les raies d'émission Lyβ et Lyα. La seconde, dénommée région Lyβ, correspond aux longueurs d'onde $920 \leq \lambda_{\mathrm{RF}} \leq 1020\,$Å, c'est à dire entre la limite de la série de Lyman et la raie d'émission Lyβ. Les pixels d'absorption Lyα dans la région Lyα sont dénommés pixels Lyα(Lyα), et ceux dans la région Lyβ sont dénommés pixels Lyα(Lyβ). De plus, l'analyse se limite aux pixels dont la longueur d'onde observée est comprise entre $3600 \leq \lambda_{obs} \leq 6000\,$Å. La limite inférieure provient de la sensibilité des CCD et de l'augmentation des absorptions atmosphériques intenses dans l'UV. La limite supérieure provient des raies d'émission du ciel dans le proche infrarouge qui bruitent le signal. Ces limites sur λ_{obs} correspondent à un redshift minimal $z_{\mathrm{QSO}} = 2$ pour les quasars utilisés pour leurs pixels dans la région Lyα (quasars Lyα), et

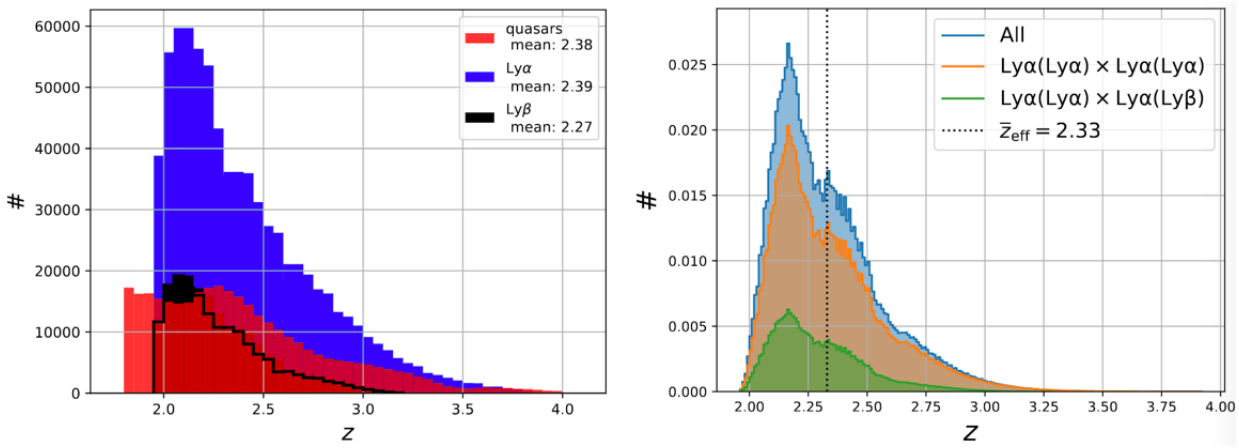

FIGURE 1.4 – Gauche : distribution en redshift des pixels Lyα(Lyα) (bleu) et des pixelsLyα(Lyβ) (noir) (nombres divisés par 50), ainsi que celle des quasars traceurs (rouge). Droite : distribution pondérée du redshift efficace de chaque paire de pixels utilisée pour le calcul de la fonction corrélation du Lyα. La distibution orange présente les paires de pixels Lyα(Lyα)-Lyα(Lyα), la verte les paires de pixels Lyα(Lyα)-Lyα(Lyβ), et la bleu la somme des deux. Les paires représentés sont celles nécessaires pour la mesure de la position du pic BAO, c'est à dire avec une séparation $r \in [80\,;120]h^{-1}$ Mpc. La ligne en pointillés indique le redshift effectif de la mesure, c'est à dire la moyenne de la distribution bleue. Il vaut $z_{eff} = 2{,}334$.

$z_{\rm QSO} = 2{,}53$ pour les quasars utilisés pour leurs pixels dans la région Lyβ (quasars Lyβ). Parmi les 341 468 quasars traceurs, 256 328 sont des quasars Lyα, et 103 080 sont des quasars Lyβ.

D'autres sélections sont aussi appliquées. Les quasars pour lesquels la probabilité d'avoir un BAL est supérieure à 0.9 sont écartés. Les mauvaises observations durant BOSS ou eBOSS sont mises de côté. De plus, chaque région nécessite au moins 50 pixels. Cette sélection décale le redshift minimal à $z_{\rm QSO} = 2{,}10$ pour les quasars Lyα, et à $z = 2{,}65$ pour les quasars Lyβ. Le pixel d'absorption Lyα de plus bas redshift se trouve à $z = 1{,}96$. Enfin, l'ajustement du continuum (voir section suivante) échoue pour environ 2 % des spectres). Ces spectres sont aussi écartés. Après toutes ces sélections, l'échantillon final contient 210 005 quasars Lyα et 69 656 quasars Lyβ.

Enfin, des corrections sur le flux et la variance du flux sont appliquées afin de corriger les imperfections de la chaîne de réduction. De plus, certaines régions spectrales en longueur d'onde observée sont masquées, à cause de l'augmentation de la variance du flux causée par les raies du ciel. Tout ceci est décrit en détail dans **CITE:dr16**

4 Définition du champ d'absorption

Dans cette section, nous décrivons comment, du flux mesuré $f_q(\lambda)$ du quasar q à la longueur d'onde observée λ, nous pouvons déduire les fluctuations du champ d'absorption Lyα. La démarche que nous décrivons ici est très bien expliquée dans la thèse **CITE:Victoria** Nous y référons le lecteur pour plus de détails.

4.1 Calcul des δ

Comme présenté dans la section **??**, le contraste de l'absorption Lyα $\delta_q(\lambda)$ est défini comme

$$\delta_q(\lambda) = \frac{f_q(\lambda)}{\overline{F}(\lambda) C_q(\lambda)} - 1\,, \tag{1.1}$$

où $\overline{F}(\lambda)$ est la transmission moyenne au redshift $z = \lambda/\lambda_{\mathrm{Ly}\alpha} - 1$, et $C_q(\lambda)$ donne le flux du spectre sans absorption. Il est dénommé *continuum* et est différent pour chaque quasar. Le produit $\overline{F}(\lambda)C_q(\lambda)$ représente donc le flux moyen attendu du quasar q. La figure **??** représente ces quantités pour les régions Lyα et Lyβ. Du fait de la faible résolution et du faible rapport signal sur bruit des données, $\overline{F}(\lambda)$ n'est pas mesurable. Pour ne pas dépendre d'analyses externes, on détermine directement le produit $\overline{F}(\lambda)C_q(\lambda)$ pour chaque spectre. Afin de prendre en compte la variabilité d'un spectre à un autre, le produit $\overline{F}(\lambda)C_q(\lambda)$ est modélisé par une relation linéaire en $\log_{10}\lambda$:

$$\overline{F}(\lambda)C_q(\lambda) = \overline{C}(\lambda_{\mathrm{RF}})(a_q + b_q \log_{10}(\lambda)) \,, \tag{1.2}$$

où $\overline{C}(\lambda_{\mathrm{RF}})$ donne le continuum moyenné sur tous les spectres, en fonction de la longueur d'onde dans le référentiel du quasar. a_q et b_q sont propres à chaque quasar, ils modélisent la diversité des quasars. Les termes a_q, b_q et $\overline{C}(\lambda_{\mathrm{RF}})$ sont déterminés en minimisant la fonction de vraissemblance L, qui s'exprime comme

$$-2\ln L = \sum_i \frac{\left[f_i - \overline{F}C_q(\lambda_i, a_q, b_q)\right]^2}{\sigma_q^2(\lambda_i)} + \ln[\sigma_q^2(\lambda_i)] \,. \tag{1.3}$$

La variance $\sigma_q^2(\lambda)$ dans chaque pixel est donnée par

$$\sigma_q^2(\lambda) = \eta(\lambda)\sigma^2_{\mathrm{instru},q}(\lambda) + \sigma^2_{\mathrm{cosmo}}(\lambda)(\overline{F}C_q(\lambda))^2 + \frac{\epsilon(\lambda)(\overline{F}C(\lambda))^4}{\sigma^2_{\mathrm{instru},q}(\lambda)} \,. \tag{1.4}$$

Le terme $\eta(\lambda)\sigma^2_{\mathrm{instru},q}(\lambda)$ rend compte de la variance qui est provoquée par l'instrument et liée à l'incertitude sur la mesure des flux. Le terme $\sigma_{\mathrm{cosmo}}(\lambda)$ est purement cosmologique et donne la variance propre au champ d'absorption du Lyα. Il traduit le fait que certaines régions de l'univers peuvent absorber plus ou moins que la moyenne. Du fait de l'isotropie, σ_{cosmo} ne dépend que de z. Le terme ad hoc $\epsilon(\lambda)(\overline{F}C(\lambda))^4/\sigma^2_{\mathrm{instru},q}(\lambda)$ rend compte du fait que la variance augmente avec le rapport signal sur bruit, probablement à cause de la diversité des quasars. **Enfin, le terme** $\ln[\sigma_q^2]$ **dans l'expression de la vraissemblance provient de la normalisation** $\sqrt{2\pi\sigma^2}$ **au dénominateur des distributions gaussiennes. Ce terme est généralement omis car constant. Mais dans notre cas,** σ_q **dépend de** a_q**,** b_q **et** $\overline{C}(\lambda_{\mathrm{RF}})$**.**

Afin d'obtenir le produit $\overline{F}C_q$ pour chaque quasar, il faut maximiser la fonction de vraissemblance L (équation 1.3), et donc ajuster les paramètres a_q, b_q, $\eta(\lambda)$, $\sigma_{\mathrm{cosmo}}(\lambda)$ et $\epsilon(\lambda)$. Cependant, celle fonction dépend elle même du produit $\overline{F}C_q$. Pour résoudre ce problème, l'ajustement est fait de manière itérative. Premièrement, les 5 paramètres à ajuster, ainsi que $\overline{C}(\lambda_{\mathrm{RF}})$, sont initialisés. Ceci permet d'ajuser $\overline{F}C_q$ sur $f_q(\lambda)$ en maximisant la fonction de vraissemblance L. Dans cet ajustement, chaque pixel observé à une longueur d'onde λ est pondéré par un poids

$$w_q(\lambda) = \frac{1}{\sigma_q^2(\lambda)} \,. \tag{1.5}$$

Une fois cette ajustement fait, $\overline{C}(\lambda_{\mathrm{RF}})$ est de nouveau calculé. Puis, les paramètres $\eta(\lambda)$, $\sigma_{\mathrm{cosmo}}(\lambda)$ et $\epsilon(\lambda)$ sont ajustés de manière à ce que, à lambda fixé, $\sigma_q(\lambda)$ reproduise la variance estimée à partir des flux $f_q(\lambda)$ de tous les quasars. Ce nouvel ajustement permet une nouvelle fois d'ajuster $\overline{F}C_q$, etc. Cette procédure est répétée jusqu'à obtenir des valeurs stables pour les paramètres ajustés. En pratique, 5 itérations sont nécessaires. Cette procédure permet donc d'estimer, pour chaque quasar, le produit $\overline{F}(\lambda)C_q(\lambda)$, et ainsi d'estimer les $\delta_q(\lambda)$ (équation 1.1) indépendamment dans chaque région.

Lors du calcul des δ_q**, les DLA identifiés dans chaque forêt sont masqués. Pour ce faire, un profil de Voigt, dépendant de la densité de colonne mesurée par l'algorithme de détection, est ajusté à l'endroit du spectre où le DLA a été identifié. Les pixels pour lesquels l'absorption est plus grande que 20 % ne sont pas utilisés. Les autres sont corrigés**

en utilisant le profile de Voigt ajusté. La figure ?? montre deux spectres présentant un DLA. Le profile de Voigt est montré en dessous de la forêt, en rouge. Les bandes rouges indiquent les pixels non utilisés.

4.2 Prise en compte de l'ajustement du continuum

La procédure décrite dans la section précédente permet d'ajuster le continuum de chaque quasar afin d'estimer le champ δ dans chaque région d'absorption Lyα. Cependant, cet ajustement biaise le champ δ mesuré, et introduit de fausses corrélations lors du calcul des fonctions de corrélation. Il est donc important de prendre en compte cet effet.

En redéfinissant les paramètres a_q et b_q, l'équation 1.2 peut être mise sous la forme

$$\overline{F}(\lambda)C_q(\lambda) = \overline{C}(\lambda_{\rm RF})(a_q + b_q(\Lambda - \overline{\Lambda}_q)) \,, \tag{1.6}$$

où $\Lambda = \log_{10}\lambda$, et $\overline{\Lambda}_q$ est la moyenne de Λ dans chaque forêt :

$$\overline{\Lambda}_q = \frac{\sum\limits_{\lambda} w_q(\lambda)\Lambda}{\sum\limits_{\lambda} w_q(\lambda)} \,. \tag{1.7}$$

Le champ δ_q mesuré est donc donné par

$$\delta_q(\lambda) = \frac{f_q(\lambda)}{\overline{C}(\lambda_{\rm RF})(a_q + b_q(\Lambda - \overline{\Lambda}_q))} - 1 \,. \tag{1.8}$$

Du fait que a_q et b_q sont ajustés sur les $f_q(\lambda)$, l'utilisation de l'équation 1.6 pour déterminer le continuum force la moyenne et la pente de chaque région spectrale à être nulle. Ainsi le champ δ mesuré est biaisé, et de fausses corrélations sont induites dans le calcul de la fonction de corrélation. Afin de prendre en compte cet effet, le champ δ est transformé selon la relation

$$\tilde{\delta}_q(\lambda_i) = \sum_j \eta_{ij}^q \delta_q(\lambda_j) \,, \tag{1.9}$$

avec

$$\eta_{ij}^q = \delta_{ij}^K - \frac{w_j}{\sum\limits_k w_k} - \frac{w_j\left(\Lambda_i - \overline{\Lambda}_q\right)\left(\Lambda_j - \overline{\Lambda}_q\right)}{\sum\limits_k w_k\left(\Lambda_k - \overline{\Lambda_q}\right)^2} \,, \tag{1.10}$$

où δ_{ij}^K est le symbole de Kronecker. Dans ce calcul, les poids $w_q(\lambda)$ sont corrigés de la dépendance en redshift du biais du Lyα **et du facteur de croissance** :

$$w_i = \sigma_q^{-2}(\lambda_i)\left(\frac{1+z_i}{1+2.25}\right)^{\gamma_{{\rm Ly}\alpha}-1} \,, \tag{1.11}$$

avec $\gamma_{{\rm Ly}\alpha} = 2.9$ (McDonald et al. 2004). Grâce à cette transformation, l'effet sur la fonction de corrélation peut ainsi être modélisé et pris en compte. Sa modélisation passe par le calcul de la *matrice de distorsion*, détaillée dans la section **??**.

Enfin, l'utilisation de $\overline{C}(\lambda_{\rm RF})$ dans l'équation 1.2 force la moyenne des δ_q dans chaque bin de longueur d'onde observée à zéro. Cependant, la transformation 1.9 modifie légèrement la moyenne des δ_q dans chaque bin. Afin que la fonction de corrélation croisée Lyα-QSO tende vers zéro à grand séparation, l'effet est corrigé en soustrayant explicitement la moyenne dans chaque bin :

$$\hat{\delta}_q(\lambda_i) = \tilde{\delta}_q(\lambda_i) - \overline{\tilde{\delta}(\lambda)} \,. \tag{1.12}$$

Dans la suite de ce manuscript, lorsque nous parlons du champ δ, nous référons à ces $\hat{\delta}_q(\lambda_i)$. Le graphique de gauche de la figure 1.4 montre la distribution en redshift des δ dans chaque région d'absortion. Le graphique de droite présente la distribution pondérée par les poids w_i (équation 1.11) du redshift effectif de chaque paire, utilisé dans le calcul de la fonction de corrélation.

Bibliographie

Ahumada, Romina et al. (2019). « The Sixteenth Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra ». In : arXiv : 1912.02905.

Bolton, Adam S. et al. (2012). « Spectral Classification and Redshift Measurement for the SDSS-III Baryon Oscillation Spectroscopic Survey ». In : doi : 10.1088/0004-6256/144/5/144. arXiv : 1207.7326.

Busca, Nicolas et Christophe Balland (2018). « QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks ». In : arXiv : 1808.09955.

Guo, Zhiyuan et Paul Martini (2019). « Classification of Broad Absorption Line Quasars with a Convolutional Neural Network ». In : doi : 10.3847/1538-4357/ab2590. arXiv : 1901.04506.

McDonald, Patrick et al. (2004). « The Lyman-alpha Forest Power Spectrum from the Sloan Digital Sky Survey ». In : doi : 10.1086/444361. arXiv : 0405013 [astro-ph].

Pâris, Isabelle et al. (2016). « The Sloan Digital Sky Survey Quasar Catalog: twelfth data release ». In : doi : 10.1051/0004-6361/201527999. arXiv : 1608.06483.

Parks, David et al. (2017). « Deep Learning of Quasar Spectra to Discover and Characterize Damped Lya Systems ». In : doi : 10.1093/mnras/sty196. arXiv : 1709.04962.