Rapport de stage - LPSC

(Laboratoire de Physique Subatomique et de Cosmologie)

Sujet : *étude d'un outil de qualité (BDT) permettant de rejeter les redshifts photométriques mal reconstruits dans LSST pour améliorer les performances photo-z.*

Tuteur : Jean-Stéphane RICOL

Mickaël LALANDE

L3 Physique / Magistère

Université Joseph Fourier

(durée : 9 semaines)

Juin/Juillet 2014

Introduction

Une des plus grande incertitude systématique dans la mesure des paramètres cosmologiques vient des erreurs que l'on commet dans l'estimation du redshift photométrique (photo-z). La spectroscopie reste plus précise mais nécessite un temps d'observation trop grand pour pouvoir détecter les 10 milliards de galaxies que verra LSST.

La photométrie est sensible aux changements globaux de la distribution du spectre d'énergie des galaxies dans chaque bande photométrique. Dans notre étude, chaque galaxie est caractérisée par 3 paramètres : redshift (z), type et taux de poussière (ebv). Le problème c'est que le système (z, type, ebv) est dégénéré et qu'on a également des erreurs photométriques provenant de l'instrument, l'atmosphère, etc. Du coup un lot de (zp, typep, ebvp) [1] peut avoir un meilleur χ^2 que le système vrai.

Ces mauvaises reconstructions restent marginales, cependant elles ont un léger impact sur le biais de notre distribution totale et ne sont donc pas souhaitables pour contraindre les paramètres cosmologiques. Nous utiliserons un outil d'analyse multivariable TMVA utilisant la méthode BDT afin d'essayer d'éliminer ces mauvaises reconstructions tout en conservant une bonne efficacité.

Il existe aujourd'hui 2 techniques pour calculer le photo-z :

– méthode empirique (efficace seulement à petit redshift)
– méthode d'ajustement de templates (moins précise mais permet d'aller à plus haut redshift)

Notre étude est basée sur la deuxième méthode, à laquelle on rajoute une coupure de qualité basée sur les caractéristiques des densités de probabilités du système (z, type, ebv).

Dans un premier temps je vais présenter rapidement le laboratoire dans lequel j'ai fait mon stage, le projet LSST, ainsi que mes objectifs. La partie 2 présente la méthode de simulation pour la reconstruction photométrique LSST actuellement utilisée. Les trois parties suivantes présenterons le travail que j'ai effectué : test de la méthode et recherche d'améliorations. Enfin je finirai par faire une synthèse globale avant de conclure sur l'étude effectuée ainsi que sur ce que ce stage m'a personnellement apporté.

1. L'indice p correspond aux valeurs photométriques.

Table des matières

1 Présentation du stage

1.1 Le LPSC : Laboratoire de Physique Subatomique et de Cosmologie

Le LPSC est une unité mixte de recherche du CNRS et de l'Université de Grenoble. Avec un effectif de plus de 200 personnes regroupant des chercheurs, ingénieurs, et autres, il représente un acteur majeur de la recherche grenobloise. Impliqué dans plusieurs grands projets scientifiques, comme le LHC ou encore le LSST, il est également présent sur la scène internationale.

La recherche fondamentale est le moteur principal du laboratoire qui se divise en plusieurs domaines d'études :

- Quarks & Leptons : Symétries fondamentales
- Astroparticules et Cosmologie
- Hadrons et noyaux : Énergie nucléaire
- Théorie : Interdisciplinaire
- Pôle accélérateurs et sources d'ions

Chacun de ces domaines regroupe plusieurs équipes de chercheurs. J'ai été affecté au groupe DARK qui se situe à la frontière entre l'astrophysique, la physique des particules et la cosmologie.

1.2 Le LSST : Large Synoptic Survey Telescope

Le LSST va être un des plus grand télescope encore jamais réalisé. Perché en haut du site de Cerro Pachon, une montagne chililienne située à 2680m, il pourra observer des objets de petite luminosité avec une très courte durée d'exposition. Son champ de vue très large et sa rapidité lui permettront de couvrir le ciel 2 fois par semaine !

Ce grand projet scientifique servira dans un champ large de l'astrophysique, de l'étude du système solaire à la cosmologie. Il servira principalement pour 4 grands domaines d'études : la matière noire et l'énergie noire, le système solaire, les occultations et transitoires, et enfin, la structure de la Voie Lactée.

Les premières images sont prévues à l'horizon 2020 et il sera en service pour une dizaine d'année.

1.3 Objectifs du stage

L'objectif principal est tourné sur l'amélioration de la méthode d'élimination des mauvaises reconstructions photométriques, tout en respectant les critères LSST, en particulier le biais. Pour cela nous avons à disposition une simulation qui reproduit les conditions que l'on observera avec LSST et un outil (TMVA) d'analyse multivariables pour dissocier au maximum les bonnes reconstructions du photo-z, des mauvaises. La partie suivante présente justement succinctement le travail effectué au préalable qui est détaillé dans la thèse [2] et l'article [1].

2 Méthode de reconstruction Photo-Z

2.1 La librairie de galaxies

Nous avons à disposition une librairie de 51 types de galaxies interpolés linéairement à partir de 6 échantillons : 0 (El), 10 (Sbc), 20 (Sbd), 30 (Irr), 40 (SB3) et 50 (SB2) [2]. À chaque galaxie simulée correspond un spectre d'énergie associé à trois variables : le redshift (z), le type (Type) et la concentration en poussière (ebv). Il est ensuite simulé ce que LSST verrait dans ces 6 bandes photométriques.

Quand les distributions spectroscopiques sont caractérisées par des effets locaux (comme par exemple un pic à une certaine longueur d'onde), les distributions photométrique sont quand à elles caractérisées par des effets plus globaux sur les magnitudes observées dans chaque bande. À partir de cela, en est déduit des densités de probabilités pour chacun des 3 paramètres : z, type et ebv. Or comme nous l'avons dit, il y a toujours quelques mauvaises reconstructions que l'on essayera justement d'éliminer au maximum avec l'outil TMVA explicité dans la section suivante.

Nous avons au total 1 728 784 galaxies. La figure 1 représente la distribution des galaxies en fonction du redshift spectroscopique.

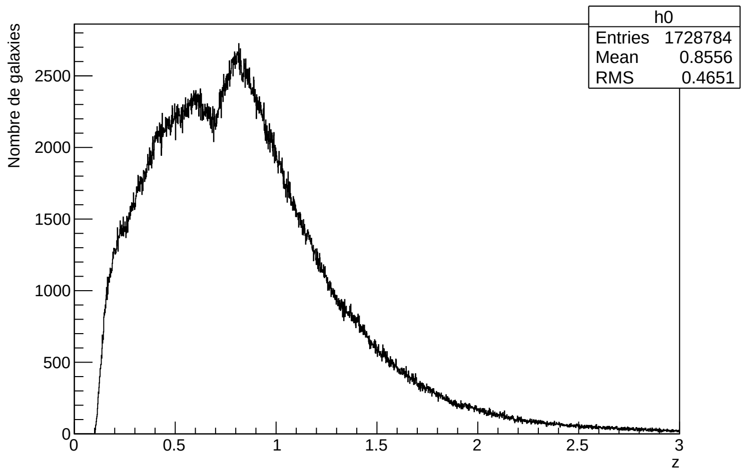


FIGURE 1 – Distribution des galaxies en fonction du redshift z

2.2 La reconstruction photométrique

Le LSST sera équipé de 6 bandes photométriques : UGRIZ, couvrant le domaine du visible entre 300 et 1100 nm. Chaque bande mesura un flux et on en déduira une magnitude. À partir de ces données il faudra reconstruire le photo-z. La méthode utilisée dans notre étude consiste à minimiser un χ^2 à partir des magnitudes pour obtenir la meilleure valeur possible pour nos paramètres. On obtient ensuite les densités de probabilités (PDF) en marginalisant chaque variable sur les deux autres.

À partir de toutes ces distributions, on crée plusieurs variables (18 actuellement) permettant de caractériser la forme de ces densités de probabilités (par exemple le nombre de pics, le rapport des intégrales des pics, etc.), mais aussi la différence de magnitude dans chaque bande photométrique.

2. El : elliptique, S : spirale, Irr : irrégulière, SB : Starbust.

Une méthode intuitive consisterait à éliminer toutes les mauvaises reconstructions directement à partir de ces variables une par une. Mais le problèmes c'est qu'à force de jouer sur chacune des variables indépendamment on finit par perdre presque toutes les galaxies. C'est pourquoi on utilise l'outil de calcul multivariable TMVA [3] qui permet de combiner toutes ces variables.

2.3 Création du BDT avec TMVA

L'outil TMVA nous permet de passer des 18 variables [3] caractérisant les densités de probabilité sur z, type et ebv, ainsi que les magnitudes, à une seule variable. Dans notre cas nous utilisons la méthode *Boosted Decision Tree* et notre variable se nommera BDT. Cette méthode nécessite un échantillon d'entrainement pour lequel on connait les redshift spectroscopiques. Cet entrainement se fait sur 10% du total du nombre de galaxies et on doit définir une limite correspondant aux bonnes et aux mauvaises reconstructions. La méthode va d'elle même jouer sur les 18 variables pour dissocier au maximum ces galaxies, comme nous pouvons le voir sur un exemple figure 2. Cette limite est imposée sur l'erreur entre le redshift "vrai", c'est à dire le redshift spectroscopique z_s et le redshift photométrique z_p, que l'on définit suivant la formule (1).

$$\Delta z = \left| \frac{z_p - z_s}{1 + z_s} \right| \tag{1}$$

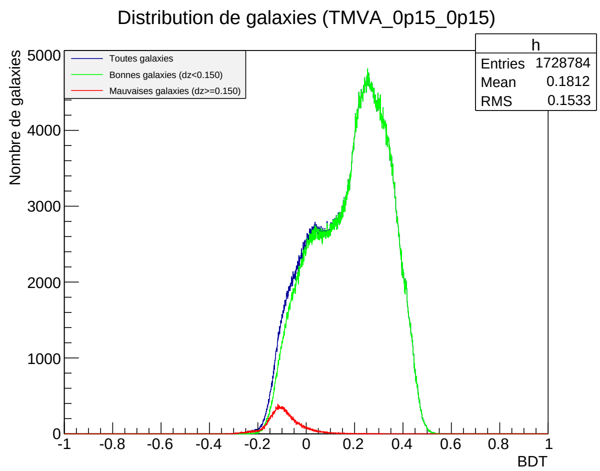


FIGURE 2 – Distribution des galaxies en fonction du BDT (bleu : toutes les galaxies, verte : bonnes galaxies ($\Delta z < 0{,}15$), rouge : mauvaises galaxies ($\Delta z \geqslant 0{,}15$)).

Dans la prochaine partie je vais présenter le travail que j'ai effectué en commençant par tester les différentes options de la méthode BDT afin d'essayer d'améliorer les performances au niveau de l'efficacité tout en respectant les spécifications LSST.

3. Ces 18 variables sont détaillées dans l'annexe B

3 Amélioration de la méthode BDT

Dans cette partie nous allons essayer d'améliorer la méthode BDT à partir d'un même fichier test que l'on prendra comme référence : TMVA_0p15_0p15.root. Celui ci correspond à un entrainement où l'on a définit la limite des bonnes et des mauvaises galaxies à Δz = 0,15. Pour cela, nous allons jouer sur plusieurs options et variables qui contribuent à la création du BDT. Notre objectif principal sera de conserver le maximum de bonnes reconstructions ($\Delta z < 0{,}15$) tout en éliminant le maximum de mauvaises ($\Delta z \geqslant 0{,}15$).

Il faut savoir que dans la méthode BDT, l'algorithme classe les variables en 2 tableaux distincts : Separation et Variable Importance. Le premier tableau montre le pouvoir de séparation de chaque variable, c'est à dire que plus la valeur est grande et plus cette variable permet de séparer les bonnes des mauvaises reconstructions [4]. L'importance d'une variable correspond au nombre de fois où la variable est appelée dans l'algorithme. Cette dernière est plus difficile à s'imager.

La modification des options ne joue seulement sur le rang Variable Importance des 18 variables. Si l'on veut modifier le pouvoir de séparation il faut alors tester différents entrainement, comme nous le verrons dans la partie 4.1.

3.1 Étude des options BDT

3.1.1 NTrees, MaxDepth, nCuts

Le BDT : Boosted Decision Tree, est basé sur la construction d'arbres. Les 3 options NTrees, MaxDepth et nCuts permettent de jouer sur la manière dont ces arbres vont être créés. Voici, les valeurs de références de notre fichier test :

- NTrees : 850 (400, 1200)
- MaxDepth : 3 (2, 4)
- nCuts : 20 (10, 30)

Afin de tester l'influence de ces valeurs, nous avons fait varier ces options d'environ ±50% (valeurs indiquées entre parenthèses). On remarque que lorsque l'on baisse la valeur, quelque soit l'option, on obtient une moins bonne efficacité/réjection. Tandis que lorsque l'on augmente ces valeurs on l'améliore très légèrement. Pour observer un effet plus conséquent, nous avons combinés, sur la figure 3, les 3 valeurs supérieures et inférieures.

Nous remarquons qu'au niveau de la distribution, le fait d'augmenter les valeurs des options a tendance à compacter la distribution et tirer légèrement les mauvaises reconstructions un peu plus vers la gauche (cf. reconstruction de référence figure 7 en haut à gauche). Au contraire, le fait de baisser les valeurs des options, entraine un étalement des mauvaises galaxies (attention bien noter l'échelle différente). La courbe d'efficacité/réjection permet d'affirmer cette observation. En effet, la courbe verte (valeurs inférieures) est bien en dessous de la courbe de référence (en noir) et on note un légèrement amélioration pour la courbe rouge (valeurs supérieures). Le gros avantage d'augmenter ces options est également d'améliorer le biais à haut redshift pour un nombre d'outliers sensiblement identique. On remarque que la courbe verte est légèrement meilleure autour du redshift égal à 2, mais cet effet est dû au fait que l'on a plus d'outliers, ce qui entraine un décalage dans le biais.

4. Separation $\in [0; 1]$. Separation = 1 correspond à une séparation complète des bonnes et des mauvaises reconstructions (il n'y a pas de recouvrement). Separation = 0 correspond à une distribution des bonnes et des mauvaises valeurs identique.

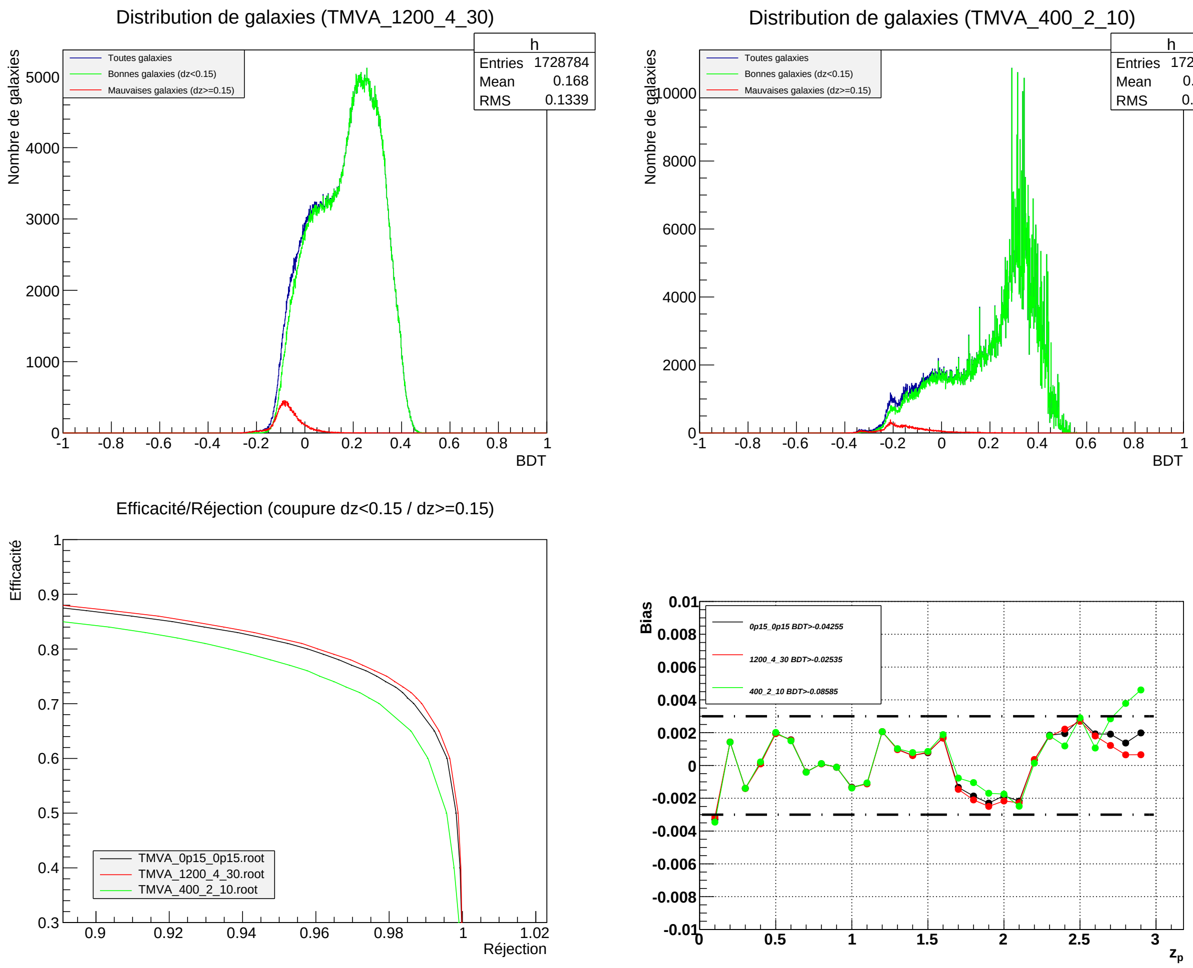


FIGURE 3 – En haut : distribution des bonnes et des mauvaises galaxies en fonction du BDT (à gauche : valeurs supérieures ; à droite : valeurs inférieures) ; en bas à gauche : efficacité en fonction de la réjection ; en bas à droite : biais sur Δz en fonction du redshift avec coupure BDT conservant 90% du nombre de galaxies totales.

Remarque : *L'option MaxDepth à tendance à augmenter légèrement le temps de calcul de quelques minutes, étant sur un temps de base d'une dizaine de minutes.*

3.1.2 BoosType

Dans notre cas, l'option BoosType peut prendre 3 valeurs : AdaBoost (par défault), RealAdaBoost ou Bagging. Sachant que pour les deux premières on peut également jouer sur le paramètre AdaBoostBeta qui est fixé à 0,5 par défaut.

Le type RealAdaBoost donne des résultats similaire (au niveau de la distribution et de l'efficacité/réjection) à la section précédente lorsque l'on diminue les valeurs des options NTrees, MaxDepth et nCuts. Ce type est donc moins bon que celui de référence AdaBoost. Au niveau du type Bagging, on obtient quelque chose de complètement différent. La distribution est coupée en deux et il faudrait faire une étude plus fine pour voir si ce type à un réel intérêt ou non.

Dans mon cas je me suis alors principalement intéressé au type de base AdaBoost, en jouant sur le paramètre AdaBoostBeta. J'ai alors essayé une valeur supérieure (0,7) et inférieure (0,3) à la valeur de référence fixé à 0,5. Les résultats sont présentés sur la figure 4.

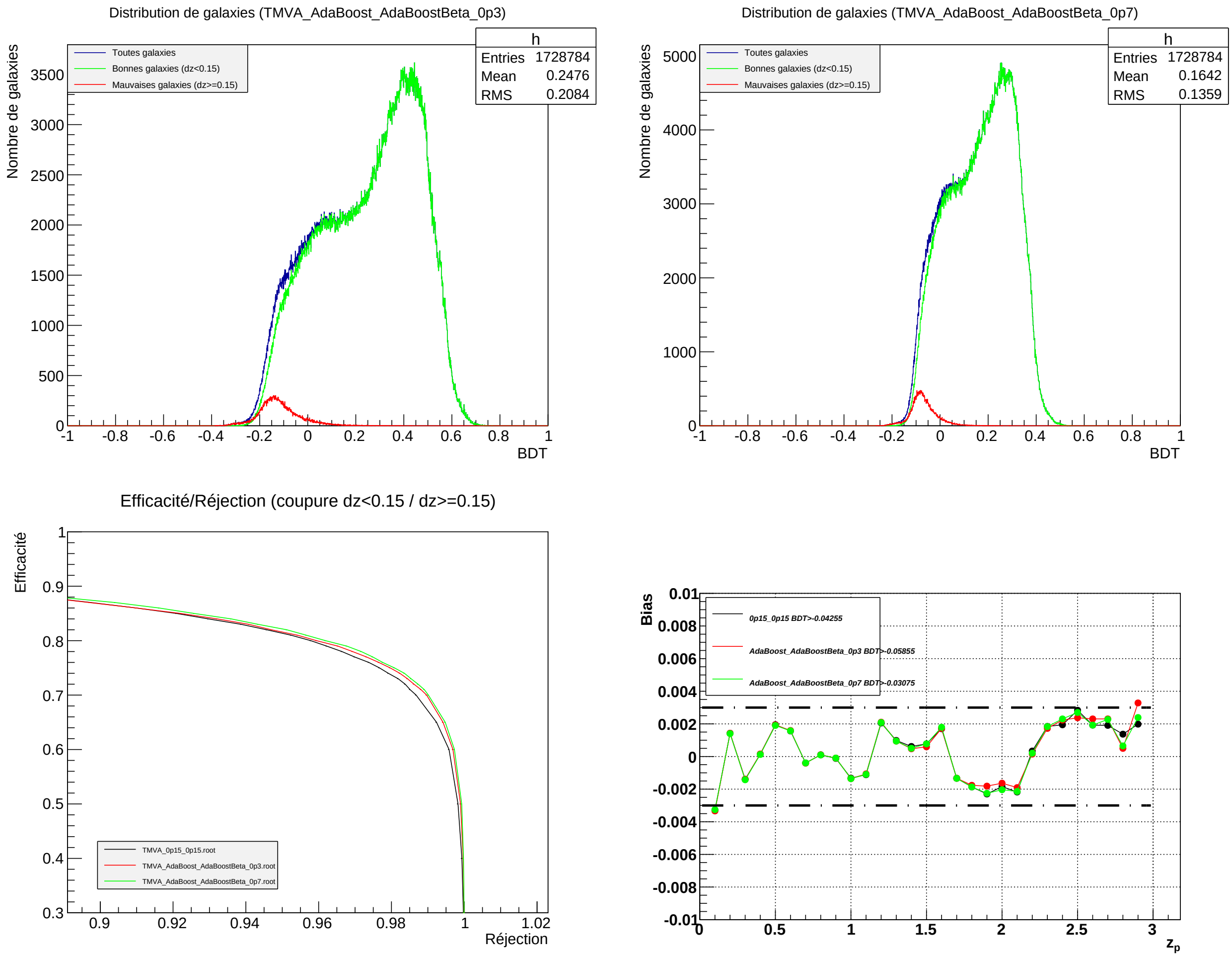


FIGURE 4 – En haut : distribution des bonnes et des mauvaises galaxies en fonction du BDT (à gauche : AdaBoostBeta = 0,3 ; à droite : AdaBoostBeta = 0,7) ; en bas à gauche : efficacité en fonction de la réjection ; en bas à droite : biais sur Δz en fonction du redshift avec coupure BDT conservant 90% du nombre de galaxies totales.

Cette fois-ci le fait d'augmenter et de diminuer la valeur du paramètre AdaBoostBeta améliore dans les 2 cas l'efficacité en fonction de la réjection. Les distributions sont quant à elles plus étalé dans le cas d'un AdaBoostBeta plus faible et plus piquée pour la valeur plus élevée, mais sans grande influence sur la position relative des bonnes et des mauvaises distributions. Au niveau du biais, la valeur à 0,3 (courbe en rouge) semble légèrement meilleure, malgré un petit pic dépassant le critère à haut redshift.

3.1.3 SeparationType

Dans le cas de notre étude il y a 5 valeurs possible pour l'option SeparationType : GiniIndex (par défaut), CrossEntropy, GiniIndexWithLaplace, MisClassificationError, SDivSqrtSPlusB.

Après avoir testé chaque valeur, j'ai remarqué qu'il n y avait pas énormément de différence dans la distribution des galaxies en fonction du BDT, tout comme il n'y a pas d'effet visible sur l'efficacité en fonction de la réjection. Sauf pour la valeur : SDivSqrtSPlusB où la distribution est complètement différente et la courbe efficacité/réjection est très mauvaise. De plus lors de ce test, j'ai eu une erreur stoppant la méthode de Boost à cause d'une erreur trop importante sur le BDT. Cette dernière valeur n'est donc pas à retenir.

3.2 Étude du nombre de variables utilisées pour la création du BDT

La méthode dispose actuellement de 18 variables caractérisant les densités de probabilité des 3 paramètres : redshift, type et taux de poussière, ainsi que les magnitudes observées pour chaque bande photométrique. Nous avons par exemple le nombre de pics, le rapport des intégrales des pics, la différence des magnitudes, etc. L'annexe B détaille chaque variable ainsi que leur pouvoir de séparation dans le cas de notre test de référence.

Pour amélioré la méthode nous avons commencé par essayer de retirer les variables qui ont le pouvoir de séparation le plus faible. Cependant, cela diminue petit à petit l'efficacité en fonction de la réjection. L'effet devient notoire à partir de 6 variables retirées, on peut voir sur la figure 5 à gauche, l'effet d'un retrait de 10 variables.

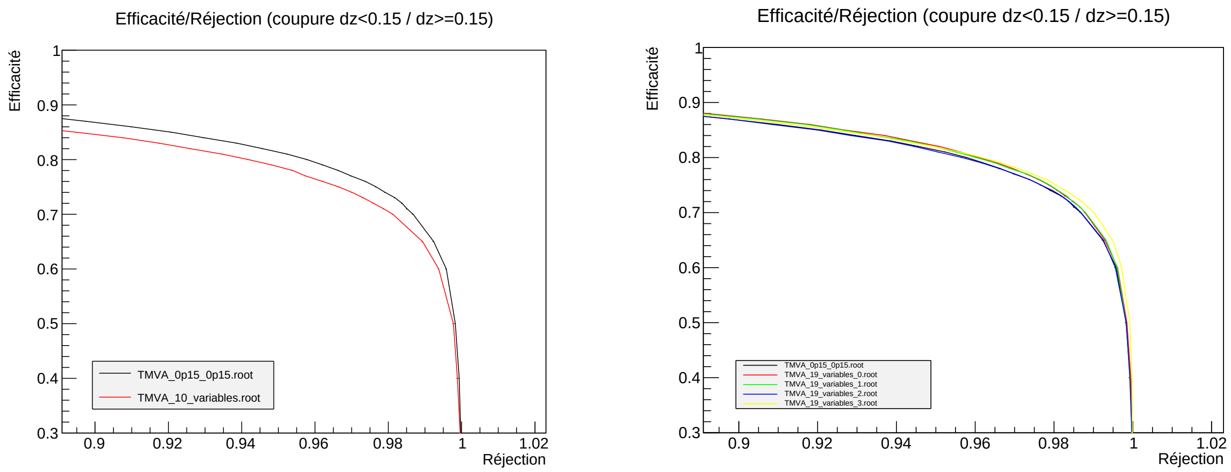


FIGURE 5 – À gauche : efficacité en fonction de la réjection pour les 10 variables les mieux classées (en rouge), courbe de référence des 18 variables (en noir). À droite : efficacité/réjection pour 19 variables, courbe de référence (noir), varp_marg1[0] (rouge), varp_marg1[1] (vert), varp_marg1[2] (bleu), mag[3] (jaune).

Nous avons donc ensuite essayé de rajouter des variables pour voir si cela améliore la courbe efficacité/réjection :

- varp_marg1[0], Separation Rank : 5
- varp_marg1[1], Separation Rank : 8
- varp_marg1[2], Separation Rank : 18
- mag[3], Separation Rank : 4

Nous pouvons voir sur la figure 5 la courbe efficacité/réjection pour l'ajout du chacune des variables indépendamment. On remarque que la meilleure courbe est la jaune, celle qui correspond à l'ajout de la variable mag[3] et l'on remarque que c'est la mieux classée. Les courbes verte et rouge sont semblables (correspondant aux rang 5 et 8), tandis que la courbe bleue n'apporte pas d'amélioration par rapport à la courbe de référence (en effet, la distribution ebv est plate, donc on attendait pas d'impact notable).

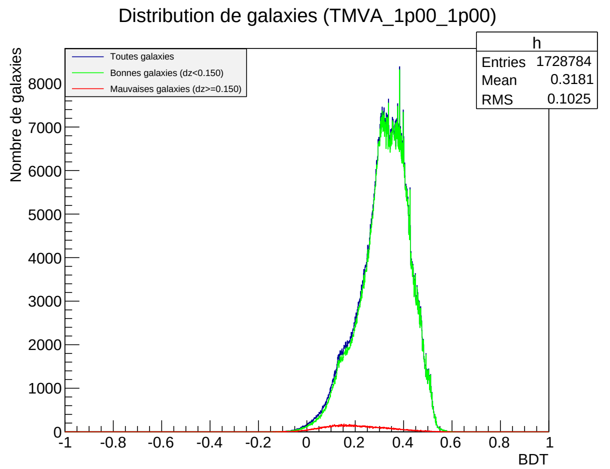
On peut en conclure, comme on l'attendait, que la classement des variables selon leur pouvoir de séparation joue directement sur la courbe efficacité/réjection en fonction du BDT. Plus on rajoute de variables bien placées, et meilleur sera la dissociation des bonnes et des mauvaises valeurs en fonction du BDT. Au niveau du temps de calcul on ne remarque pas une grande différence, on a donc tout intérêt à mettre le maximum de variable ayant un bon pouvoir de séparation.

4 Étude de différents entrainements (à modifier !)

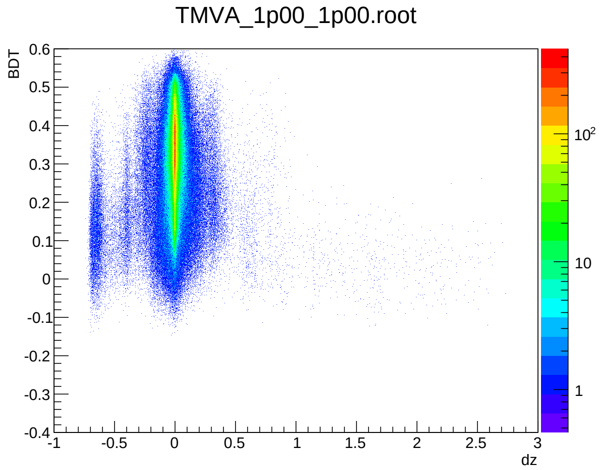
4.1 Création de plusieurs fichiers tests avec différents entrainements

Pour notre entrainement, nous devons définir les bonnes et les mauvaises galaxies par rapport à l'erreur Δz. J'ai donc essayé plusieurs coupures : 1.00, 0.20, 0.15 (par défaut), 0.10, 0.05, 0.02 et 0.003. Sachant que pour LSST la limite bonnes/mauvaises galaxies est définit à $\Delta z = 0{,}15$. C'est cette limite qui nous servira de référence pour comparer les tests après entrainement (quelque soit l'entrainement).

Nous allons voir quelle est l'influence de cet entrainement sur la distribution des galaxies en fonction de la variable BDT. Pour commencer, regardons la distribution avec la limite $\Delta z = 1{,}00$.



(a) Distribution des galaxies en fonction de la variable BDT. En bleu : toutes les galaxies, en vert : bonnes galaxies ($\Delta z < 0{,}15$) et en rouge : mauvaises galaxies ($\Delta z \geqslant 0{,}15$)



(b) Variable BDT en fonction de l'erreur Δz. L'échelle de couleur représente la concentration en galaxies

FIGURE 6 – Entrainement avec bonnes galaxies : $\Delta z < 1{,}00$ et mauvaises galaxies $\Delta z \geqslant 1{,}00$

Nous voyons sur la figure 6 que les distributions des bonnes et de mauvaises galaxies ne sont quasiment pas dissociées par rapport à la variable BDT. Nous allons voir maintenant l'effet qu'a un entrainement avec une définition des bonnes et des mauvaises galaxies à 0,15 et 0,003 dans le but de dissocier au maximum les bonnes des mauvaises reconstructions photo-z.

Nous pouvons voir sur la figure 7 que ces 2 entrainements on observe une dissociation beaucoup plus nette entre les bonnes et les mauvaises reconstructions. En effet, les galaxies ayant un bon photo-z se concentrent pour des BDT élevés tandis que les mauvaises se trouvent à des BDT plus faible. Le fait d'augmenter le critère à 0,003 n'est pas forcément meilleur car une grande partie des bonnes galaxies commence également à descendre avec les mauvaises (en fonction de la variable BDT).

Afin de pouvoir comparer nous avons placé une barre horizontale en pointillée pour repérer la perte de 10% [5] du nombre de galaxies total. L'entrainement ayant comme limite bonnes/mauvaise galaxies 0,15 permet à l'aide d'une coupure sur la variable BDT de conserver une grande partie des bonnes reconstructions tout en éliminant la plupart des très mauvaises reconstructions. La question est de

5. Pour déterminer les valeurs de BDT correspondantes, j'ai réalisé un code C++ qui permet d'évaluer l'intégrale de la distribution des galaxies total en fonction du BDT. L'annexe A donne les coupures BDT correspondant à une efficacité de 100% à 50% avec un pas de 10%

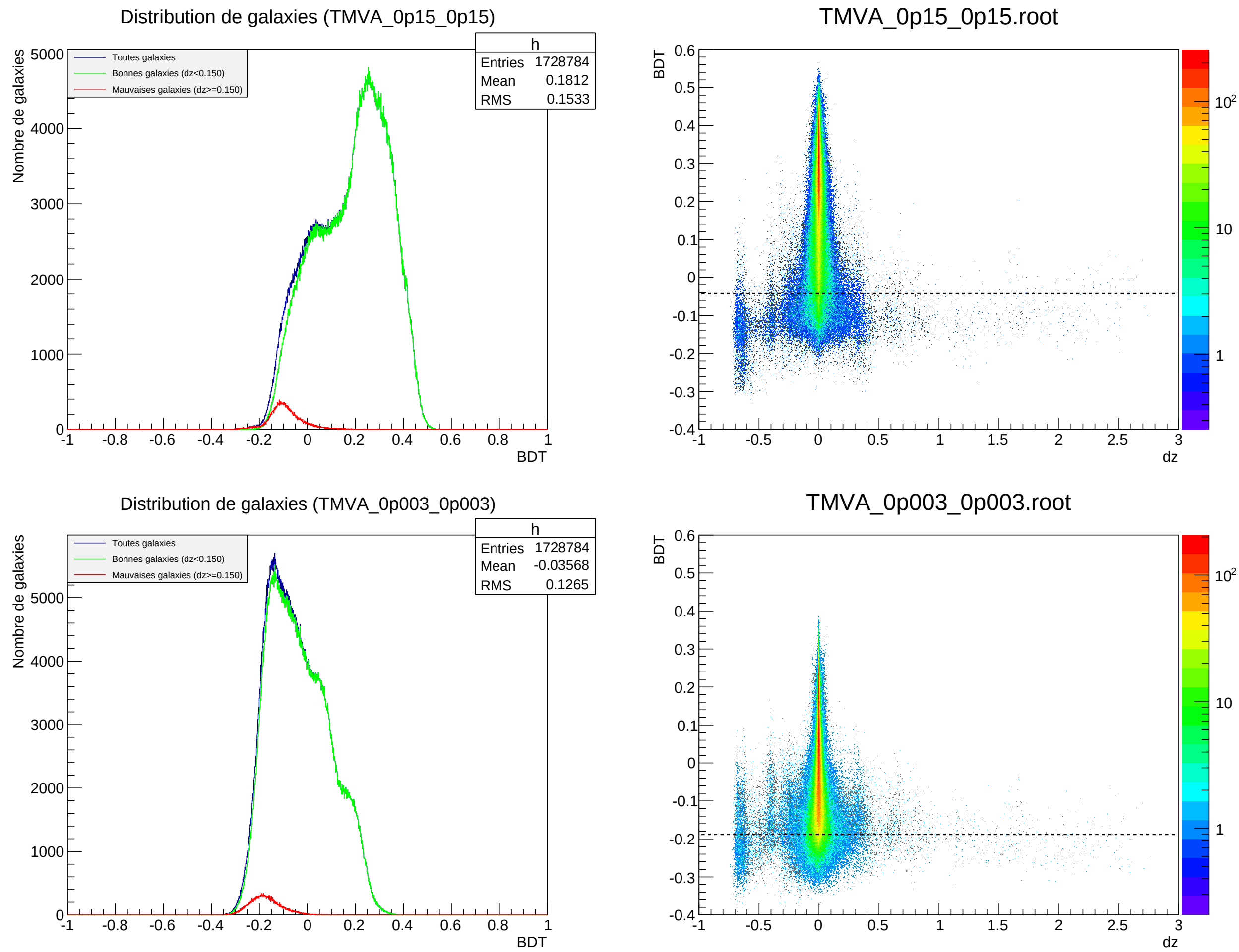


FIGURE 7 – Entrainement avec bonnes galaxies : $\Delta z < 0{,}15; 0{,}003$ et mauvaises galaxies $\Delta z \geqslant 0{,}15; 0{,}003$

savoir quelle influence cela aura sur les critères imposés par LSST.

Remarque : *Nous avons également testé des entrainements avec “trou”, c'est à dire, en définissant les bonnes galaxies avec $\Delta z < 0.05$ et les mauvaise pour $\Delta z \geqslant 0.15$, cependant le nombre de galaxie ne reste pas constant, et cela est compliqué à comparer. Nous avons donc laissé tombé cette idée.*

4.2 Vérification des critères LSST

Il y a trois critères imposés par LSST :

- RMS : $\dfrac{\sigma_z}{1+z} < 0{,}05$ (avec objectif 0,02)
- Outliers : $\eta < 10\%$
- Bias : $\Delta z = \left|\dfrac{z_p - z_s}{1+z_s}\right| < 0{,}003$

Le premier critère RMS, indique l'étalement des données. Le pourcentage d'outliers correspond au nombre de galaxies ayant une reconstruction du photo-z catastrophique (ceci étant défini pour $\Delta z \geqslant 0{,}15$) et le biais représente l'écart entre $\Delta z = 0$ et la moyenne de toute la distribution à un redshift donné. Afin de tester ces critères mon tuteur m'a mis à disposition un code C++ déjà tout fait.

Remarque : *Il n'y a pas de critère sur l'efficacité totale (ie. le nombre total de galaxies à conserver), on peut donc se permettre de retirer 10% des galaxies par rapport à notre librairie de référence. Sachant que de base, aucune distribution ne respecte ces critères une coupure sur le BDT s'impose afin d'éliminer une partie des mauvaises reconstructions.*

J'ai alors testé plusieurs coupures BDT sur chacun des entrainements en conservant de 100% à 50% avec un pas de 10% afin de voir l'influence sur les critères LSST. Le critère le plus sensible est celui du biais. En conservant toutes les reconstructions nous ne respectons pas ces critères. Par contre dès 90% d'efficacité nous avons des résultats assez bon comme nous pouvons le voir sur la figure 8.

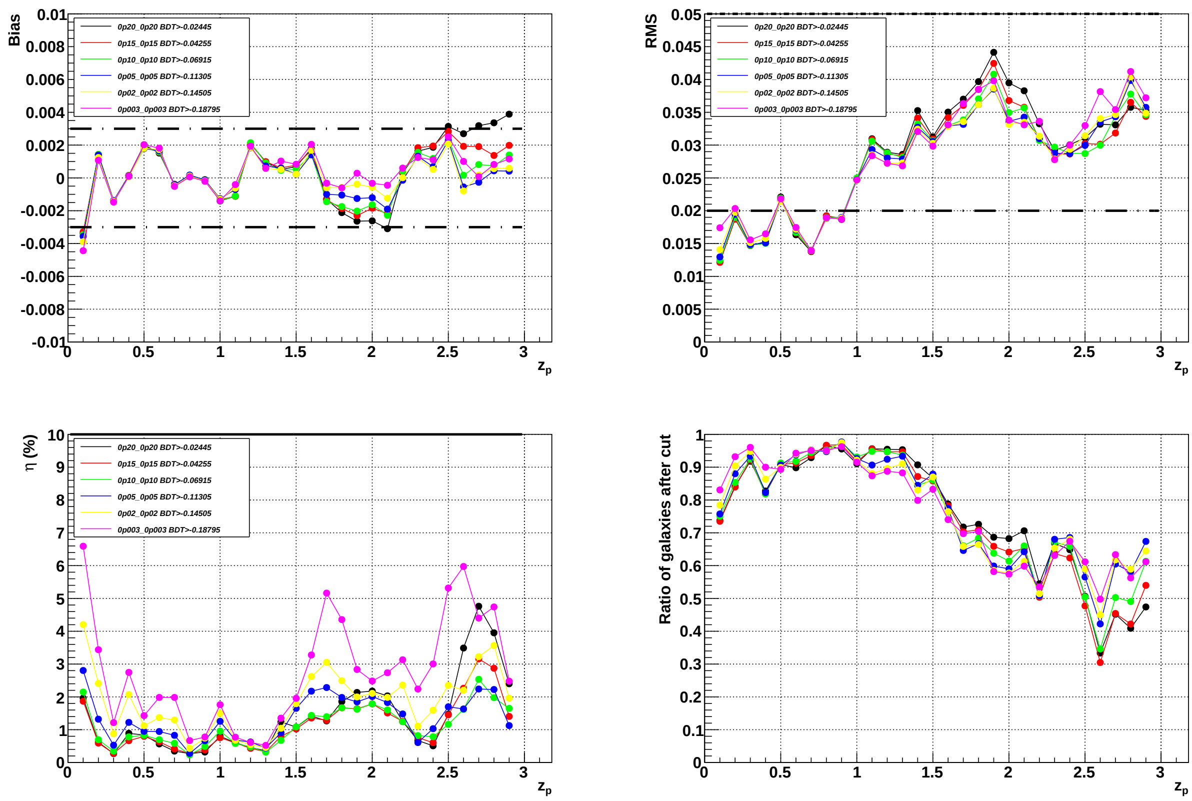


FIGURE 8 – Comparaison des différents entrainements avec une coupure BDT conservant 90% des galaxies, en fonction du redshift photométrique z_p

Au niveau du biais, le graphe montre qu'il est meilleur plus l'entrainement est effectué avec une limite bonnes/mauvaise galaxies petite. On remarque 3 zones qui approchent le critère imposé ou le dépasse même : à petits redshift, autour de $z_p = 2$, et pour $z_p > 2.5$. La première peut s'expliquer par un problème de condition aux limites proche du $z = 0$. La dernière est principalement due au fait que l'on possède très peu de galaxies à grands redshift (cf. figure 1). Par contre la zone autour du redshift égal à deux est plus intéressante et nous allons essayer de comprendre pourquoi il y a une différence aussi importante entre les différents entrainements dans le section suivante.

Au niveau du RMS on remarque qu'il est également meilleur pour les entrainements ayant une limite bonnes/mauvaises galaxies faibles même s'ils possèdent plus d'ouliers (ie. de galaxies très mal reconstruites pour $\Delta z > 0{,}15$). On vérifie bien que le ratio de galaxies après la coupure est semblable pour tous les tests, ayant exprès fait différentes coupures sur la variable BDT pour avoir le même nombre de galaxies. On remarque tout de même un baisse plus importante de ce ratio à haut redshift ce qui est dû au nombre restreint de galaxies lointaines que l'on a à disposition dans notre librairie.

Si l'on baisse encore l'efficacité à 80% et plus, le biais devient encore meilleur, tout comme le RMS où l'on arrive presque à tomber en dessous la limite 0,02 qui est l'objectif. De même, le nombre d'outliers diminue encore plus cependant on commence à perdre une part non négligeable de bonnes reconstructions, ce qui n'est pas souhaitable.

4.3 Étude autour du redshift $z = 2$

Pour comprendre ce qui se passe autour de du redshift égal à 2, là où il y a beaucoup de variations entre les différents tests d'entrainements, nous avons pris les deux cas extrêmes pour bien distinguer ce qu'il se passe. J'ai donc sélectionné les tests pour lesquels on a définit le limite bonnes/mauvaises galaxies à $\Delta z = 0{,}20$ (courbe noire) et $\Delta z = 0{,}003$ (courbe rose).

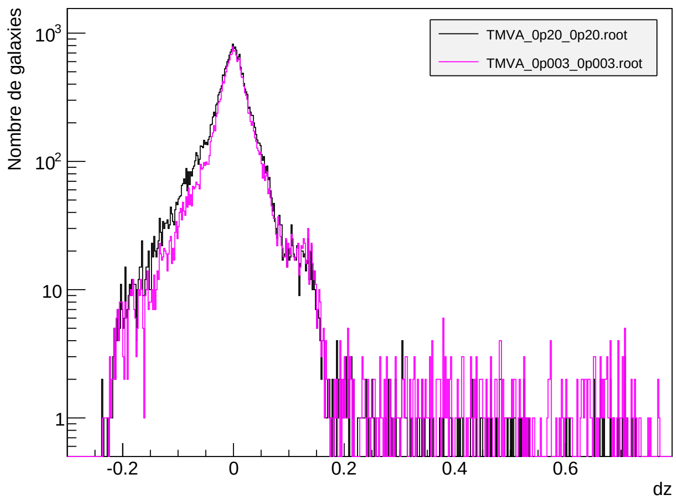


FIGURE 9 – Distribution de galaxies en fonction de l'erreur Δz pour un redshift compris entre $1{,}8 < z < 2{,}1$

La figure 9 montre que la distribution avec la distribution rose (limite à $\Delta z = 0{,}003$) est plus étroite que la distribution noire (limite à $\Delta z = 0{,}15$) au niveau du pic autour d'une erreur nulle. Cependant, nous observons plus de mauvaises reconstructions pour la distribution rose, ainsi qu'une bosse autour de $\Delta z = 0{,}15$ plus importante. Ces valeurs ont tendance à tirer la moyenne légèrement vers la droite et donc ramener le biais vers 0. Cela explique également pourquoi le RMS est plus faible tout en ayant plus de outliers pour la distribution rose.

Il n'est donc pas si évident de conclure quand à quel entrainement est le meilleur aux vues de ces résultats. En effet, au niveau du biais les distributions ayant eu un entrainement avec une limite bonnes/mauvaises galaxies très faible est meilleur. Cependant nous venons de voir que ce bon résultat sur le biais est en quelque sorte lui même biaisé par de plus nombreuses reconstructions aberrantes. Néanmoins nous restons dans les critères LSST au niveau du nombre de valeurs aberrantes qui doit être inférieur à 10%. Donc je conclurai que définir une limite bonnes/mauvaises galaxies très faibles est meilleur au vue des seuls critères LSST, cependant il faudrait pouvoir effectuer des tests sur le calcul des paramètres cosmologiques pour voir la réelle influence de ces différents entrainements. Ces codes n'étant pas encore à disposition il est difficile de conclure au jour d'aujourd'hui.

5 Essais d'autres méthodes : Fisher et MLP

L'outil TMVA, dispose également d'autre méthodes que la méthode BDT. Nous pouvons voir sur la figure 10, sur un autre test que le notre, l'efficacité en fonction de la réjection d'autres méthodes.

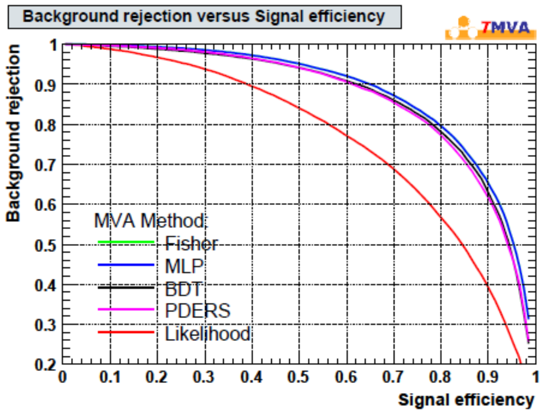


FIGURE 10 – Exemples de courbes efficacité/réjection en fonction de la méthode utilisée. Source : TMVA User Guide [3], page 12.

On remarque que la méthode MLP semble meilleure que la méthode BDT. Nous avons donc testé 2 autres méthodes : Fisher (qui ne se voit pas sur le graphe) et MLP.

5.1 Fisher

Nous avons utilisé la méthode Fisher principalement pour tester notre algorithme, car très rapide. Elle prend une dizaine de secondes quand la méthode BDT prend une dizaine de minutes. Cependant les résultats de cette méthode ne sont pas très bons, comme on peut le constater sur la figure 11 où l'on voit les mauvaises galaxies encore très étalées, comme on avait pu le voir sur la figure 6 où on avait effectué un entrainement très large avec la méthode BDT. L'efficacité en fonction de la réjection est sans étonnement moins bon que la méthode BDT.

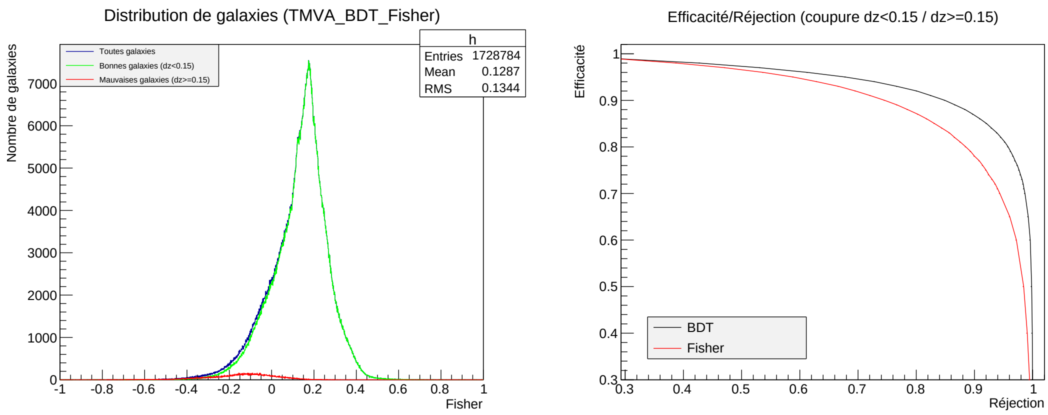


FIGURE 11 – À gauche : distribution des galaxies en fonction du Fisher. À droite : efficacité/réjection pour la méthode Fisher (en rouge) et la méthode BDT (en noir). Nous avons utilisé l'entrainement de référence pour les 2 méthodes avec une limite bonnes/mauvaises galaxies défini pour $\Delta z = 0{,}15$.

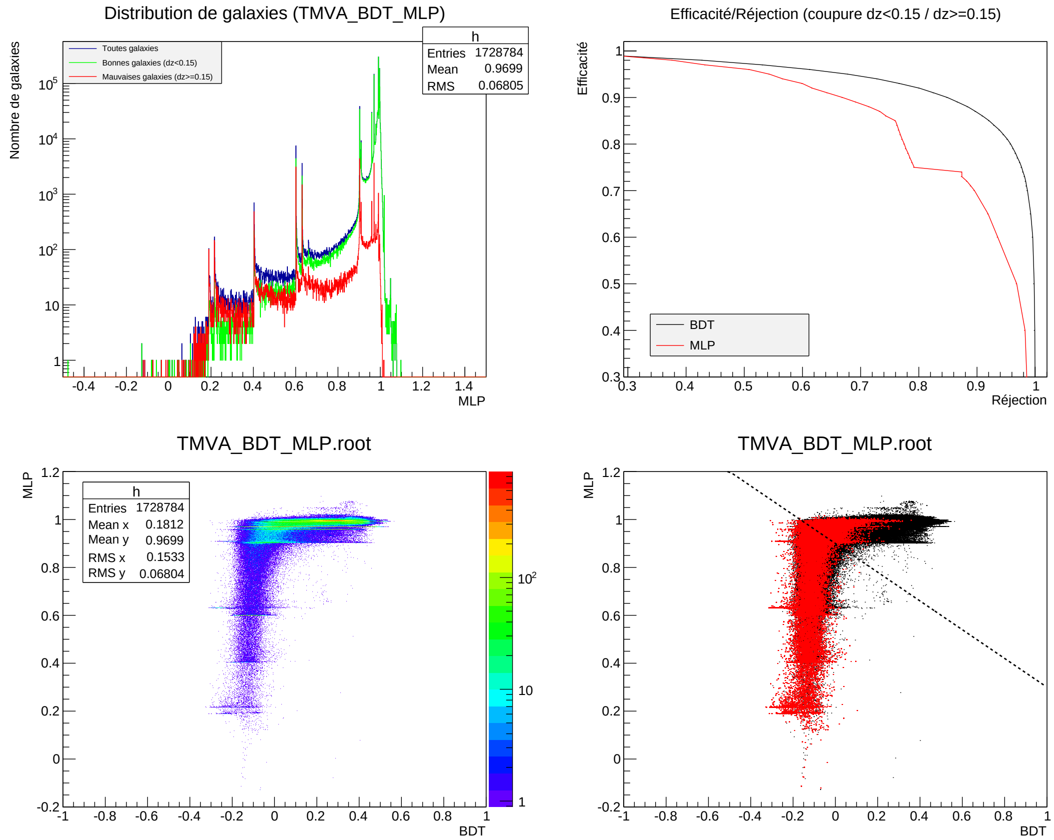


FIGURE 12 – En haut à gauche : distribution des galaxies en fonction du MLP (échelle log). En haut à droite : efficacité/réjection pour la méthode MLP (en rouge) et la méthode BDT (en noir). Nous avons utilisé l'entrainement de référence pour les 2 méthodes avec une limite bonnes/mauvaises galaxies défini pour $\Delta z = 0{,}15$. En bas à gauche : histogramme 2D représentant la corrélation entre le MLP et BDT (l'échelle de couleur représente la concentration en galaxies). En bas à gauche : recouvrement des galaxies totales (en noir) et des mauvaises galaxies (en rouge). Droite en pointillé : $MLP = -0{,}6.BDT + 0{,}9$.

5.2 MLP

La méthode MLP semble meilleure que la méthode BDT d'après le TMVA User Guide, donc nous l'avons testée dans notre cas. Voici sur la figure 12 les résultats.

La distribution des galaxies est très différente des autres et présente des pics assez marqué pour certaines valeurs de MLP. Contrairement à ce que l'on attendait, cette méthode n'est pas meilleure que la méthode BDT comme on peut le voir sur la courbe efficacité/réjection. La corrélation entre les deux méthodes est assez intéressante et présente deux lignes perpendiculaires caractéristiques. Nous avons alors essayé d'améliorer l'efficacité en fonction de la réjection selon la droite représenté sur la figure en bas à gauche. Cependant nous n'arrivons pas à obtenir des résultats meilleurs qu'avec la coupure de base sur le BDT.

De plus la méthode MLP présente un gros désavantage : la vitesse. En effet, la durée d'exécution est de plus d'une heure et demie tandis que la méthode BDT ne prend pas plus que 10 minutes. Aux vues des résultats sur ces deux méthodes supplémentaire, il n'y a pas de doute que la méthode BDT dans notre cas est bien meilleure.

Conclusions

On rappelle le cadre général de cette étude qui est basé sur la reconstruction du redshift photométrique à partir de la mesure des magnitudes dans les 6 bandes LSST. La mesure des magnitudes permet d'en déduire des densités de probabilités pour le redshift, le type et le taux de poussières (qui sont les 3 paramètres caractérisant une galaxie). Cette méthode n'est pas parfaite, dû à la mauvaise résolution photométrique. On définit alors plusieurs variables caractérisant ces distributions afin d'éliminer au maximum de mauvaises reconstruction sur le redshift (qui est le paramètre qui nous intéresse).

Pour cela, nous utilisons un algorithme puissant d'analyse multivariable TMVA, qui permet de combiner toutes ces variables en une seule via la méthode BDT. Cet algorithme nécessite un échantillon d'entrainement dont il connait les vraies valeurs (dans notre cas nous travaillons sur une simulation donc ce n'est pas un problème, en pratique il y aura une étude en parallèle de LSST pour obtenir le redshift spectroscopique d'un petit échantillon de galaxies).

Nous effectuons ensuite une coupure sur la variable BDT afin d'éliminer le plus de mauvaises reconstructions tout en conservant le maximum de bonnes. Nous vérifions ensuite les critères LSST. Notre travail est basé sur une simulation, et le but principal est d'améliorer au maximum la méthode de reconstruction photométrique pour que dans la pratique cette méthode soit la plus performante et fiable possible.

La première partie de mon étude a consisté à étudier l'influence de la définition des bonnes et de mauvaises galaxies lors de l'entrainement pour la méhode BDT, avec l'objectif principal : dissocier au maximum la distribution des bonnes et des mauvaises galaxies tout en respectant les critères LSST. Le premier critère étant basé sur la définition des *outliers* (très mauvaises reconstructions) correspondant à une erreur sur le redshift de $\Delta z \geqslant 0{,}15$, cela impose un entrainement correspondant à la même définition pour obtenir la meilleure efficacité en fonction de la réjection. Néanmoins, nous avons vu dans la section 4.2, que nous avons une certaine marge sur le pourcentage d'outliers (qui est définit à 10% pour LSST), afin d'améliorer le biais sur le redshift. Un entrainement ayant une définition des bonnes et des mauvaises galaxies beaucoup plus faible améliore notablement le biais et l'écart type malgré le fait qu'il conserve plus de très mauvaises reconstructions. Pour le moment, rien ne nous permet de conclure quant à quel entrainement est meilleur, il faudra attendre de pouvoir tester ces résultats dans la détermination des paramètres cosmologiques.

Dans un deuxième temps, nous nous sommes plus focalisé sur l'amélioration même de l'outil TMVA utilisant la méthode BDT afin de dissocier au maximum les bonnes des mauvaises reconstructions du redshift photométrique. Pour cela, nous avons utilisé le test de référence qui est celui correspondant à l'entrainement basé sur le critère LSST des bonnes et des mauvaise galaxies à $\Delta z = 0{,}15$. Première conclusion : plus on augmente les valeurs des options NTrees, MaxDepth et nCuts, et meilleur sera l'efficacité en fonction de la réjection, tout en sachant que MaxDepth à tendance à rallonger légèrement le temps de calcul. Au niveau de l'option BoostType, la valeur AdaBoost est la meilleure. Par contre, il sera intéressant d'étudier plus en détail la paramètre AdaBoostBeta, qui en modifiant sa valeur, augmente les performances. L'option SeparationType GiniIndex donne par défaut les meilleurs résultats. Au niveau du nombre de variables, il serait intéressant d'essayer d'en rajouter quelques unes ayant un bon pouvoir de séparation quitte à enlever les 2 ou 4 dernières qui n'influent que très peut sur les performances. En particulier l'étude d'ajout des magnitudes dans chaque bande semble être une bonne piste.

Pour finir, nous avons essayé de voir si une autre méthode ne serait pas plus performante que la méthode BDT. Cependant, aucune des méthodes testées n'a apporté de meilleurs résultats que la méthode BDT. Il serait tout de même intéressant d'essayer de trouver une combinaison de deux méthodes afin de chercher une meilleure façon d'éliminer les mauvaises reconstructions (ex : droite, cercle, ellipse, etc.).

Au niveau personnel ce stage m'aura beaucoup apporté autant que sur le côté que professionnel. En effet, j'ai pu découvrir la vie de chercheur en m'immergeant pendant plusieurs semaines dans un domaine qui me passionne. J'ai également eu l'opportunité de compléter mes notions de cours et découvrir des nouveaux outils de travail, comme root/c++ par exemple. Je remercie toute l'équipe du LPSC qui m'a accueilli lors de ce stage. Je remercie en particulier mon tuteur Jean-Stéphane ainsi que toutes les personnes qui ont pû me guider et m'apporter de l'aide pendant ces 2 mois.

Pour conclure, ce stage m'aura été vraiment bénéfique et il m'aura permis de conforter mes ambitions pour devenir chercheur en Astrophysique et mon goût pour la simulation qui permet de mixer informatique et science.

Références

[1] A. Gorechi, A. Abate, R. Ansari, A. Barrau, S. Baumont, M. Moniez, and J.S. Ricol. A new method to improve photometric redshift reconstruction. *Astronomy and Astrophysics*, 2013.

[2] A. Gorecki. *Cosmologie observationnelle avec le Large Synoptic Survey Telescope.* PhD thesis, Université de Grenoble, 2011.

[3] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss. *TMVA 4 (Toolkit for Multivariate Data Analysis with ROOT) Users Guide.* CERN, October 2013.

A Tableau récapitulatif des coupures BDT en fonction de l'efficacité

	BDT cut					
Ratio	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
100%	-0.30055	-0.32645	-0.30275	-0.30555	-0.34465	-0.39465
90%	-0.02445	-0.04255	-0.06915	-0.11305	-0.14505	-0.18795
80%	0.04345	0.02805	-0.00665	-0.06365	-0.10635	-0.15275
70%	0.10505	0.09305	0.04995	-0.02045	-0.07285	-0.12125
60%	0.16095	0.15475	0.12335	0.02845	-0.03705	-0.08765
50%	0.20595	0.20525	0.20855	0.09235	0.00575	-0.05195

	Nombre de galaxies						
Ratio	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Théorique
100%	1728784	1728784	1728784	1728784	1728784	1728784	1728784
90%	1555884	1555915	1555949	1555866	1555939	1556090	1555906
80%	1383082	1383007	1382974	1383091	1382896	1383289	1383027
70%	1210017	1210086	1210149	1210016	1210356	1210355	1210149
60%	1037128	1037392	1037362	1037410	1037383	1037236	1037270
50%	864500	864296	864378	864436	864522	864416	864392

	Erreur relative sur le nombre de galaxies					
Ratio	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
100%	0.000%	0.000%	0.000%	0.000%	0.000%	0.000%
90%	0.001%	0.001%	0.003%	0.003%	0.002%	0.012%
80%	0.004%	0.001%	0.004%	0.005%	0.009%	0.019%
70%	0.011%	0.005%	0.000%	0.011%	0.017%	0.017%
60%	0.014%	0.012%	0.009%	0.013%	0.011%	0.003%
50%	0.012%	0.011%	0.002%	0.005%	0.015%	0.003%

FIGURE 13 – Test 1 : TMVA_0p20_0p20 ; Test 2 : TMVA_0p15_0p15 ; Test 3 : TMVA_0p10_0p10 ; Test 4 : TMVA_0p05_0p05 ; Test 5 : TMVA_0p02_0p02 ; Test 6 : TMVA_0p003_0p003 ;

B Récapitulatif des variables utilisées pour la méthode BDT

[0] : densité de probabilité du redshift photométrique (z)
[1] : densité de probabilité du type (Type)
[2] : densité de probabilité du taux de poussière (ebv)

Variables actuellement utilisées (18) :

- **npeak[0,1,2] :** nombre de pics
- **log(Likelihood2[0,1,2]) - log(Likelihood1[0,1,2]) :** différence logarithmique entre la probabilité du second pic avec le grand pic
- **rapportInt0 :** intégrale du plus grand pic
- **rapportInt1 :** intégrale du second pic
- **rapportInt2 :** intégrale du troisième pic
- **Abs(varp[0] - varp_marg[0]) :** différence entre 2 estimateurs de z
- **Abs(varp[0] - varp_marg1[0]) :** différence entre 2 estimateurs de z
- **podds :** intégrale du plus grand pic à 3σ divisée par l'intégrale totale
- **loglmax :** relié au χ^2 via la relation $-0{,}5 \log \chi^2$
- **mag(i-1) - mag(i) :** différence entre la magnitude de la bande photométrique i-1 et i [6]

Variables test :

- **varp_marg1[0,1,2] :** estimateur de z, Type, ebv
- **mag[3] :** magnitude dans la bande photométrique 3

Rank	Variable	Separation
1	podds	5.715e-01
2	npeak[0]	4.678e-01
3	raportInt0	3.603e-01
4	log(Likelihood2[0])-log(Likelihood1[0])	2.116e-01
5	TMath::Abs(varp[0]-varp_marg1[0])	1.642e-01
6	TMath::Abs(varp[0]-varp_marg[0])	1.256e-01
7	mag2-mag3	1.048e-01
8	mag3-mag4	7.166e-02
9	raportInt2	6.527e-02
10	log(Likelihood2[2])-log(Likelihood1[2])	4.767e-02
11	log(Likelihood2[1])-log(Likelihood1[1])	4.598e-02
12	raportInt1	4.220e-02
13	mag0-mag1	3.061e-02
14	mag4-mag5	2.708e-02
15	mag1-mag2	2.493e-02
16	loglmax	2.155e-02
17	npeak[1]	2.002e-02
18	npeak[2]	1.145e-04

FIGURE 14 – Rang des 18 variables par rapport à leur pouvoir de séparation. Test : TMVA_0p15_0p15.root.

6. Pour LSST il y a 6 bandes photométriques, soit $i \in [0; 5]$