amU
Aix Marseille Université

NNT : 2025AIXM0001

THÈSE DE DOCTORAT

Soutenue à AMU — Aix-Marseille Université
le 9 avril 2025 par

Meriam EZZIATI

Searching for high-z quasars in the Euclid Wide Survey

Discipline
Physique et Sciences de la Matière

Spécialité
Astrophysique et Cosmologie

École doctorale
ED 352

Laboratoire/Partenaires de recherche
Laboratoire d'Astrophysique de Marseille

Composition du jury

Giulia RODIGHIERO University of Padova	Rapporteure
Hervé DOLE Université Paris-Saclay -IAS	Rapporteur
Marta VOLONTERI IAP	Examinatrice
Julyan ARBEL Inria Grenoble	Examinateur
Hakim ATEK IAP	Examinateur
Véronique BUAT LAM	Présidente du jury
Roser PELLO LAM	Directrice de thèse
Pierre PUDLO I2M	Co-directeur de thèse
Jean-Gabriel CUBY CFHT/LAM	Membre invité
François-Xavier DUPÉ LIS	Membre invité

LAM LABORATOIRE D'ASTROPHYSIQUE DE MARSEILLE

cnrs

euclid consortium

Affidavit

Je soussignée, Meriam Ezziati, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Roser Pello, dans le respect des principes d'honnêteté, d'intégrité et de responsabilité inhérents à la mission de recherche. Les travaux de recherche et la rédaction de ce manuscrit ont été réalisés dans le respect à la fois de la charte nationale de déontologie des métiers de la recherche et de la charte d'Aix-Marseille Université relative à la lutte contre le plagiat.

Ce travail n'a pas été précédemment soumis en France ou à l'étranger dans une version identique ou similaire à un organisme examinateur.

Fait à Marseille le 11/02/2025

Signature,



List of publications and participation in conferences

List of publications :

1. **M. Ezziati**, R. Pello, J-G. Cuby , P. Pudlo , F-X. Dupé, et al. *Euclid : `Owl-z` : Bayesian tool for selecting $z \gtrsim 7$ quasars*, *Astronomy & Astrophysics*, Vol. 701, A282, 2025. ,
2. **Euclid Collaboration**, K. Koyama, S. Pamuk, et al. (including **M. Ezzaiti**), *Euclid preparation. Simulations and nonlinearities beyond ΛCDM. 4. Constraints on $f(R)$ models from the photometric primary probes*, *arXiv e-prints*, arXiv:2409.03524, 2024.
   Citation : `@ARTICLE{2024arXiv240903524E}`
3. **Euclid Collaboration**, L. Ingoglia, M. Sereno, et al. (including **M. Ezzaiti**), *Euclid preparation : Determining the weak lensing mass accuracy and precision for galaxy clusters*, *arXiv e-prints*, arXiv:2409.02783, 2024.
   Citation : `@ARTICLE{2024arXiv240902783E}`
4. **Cagliari, M.S.**, B.R. Granett, L. Guzzo, et al. (including **M. Ezzaiti**), *Euclid : Testing photometric selection of emission-line galaxy targets*, *Astronomy & Astrophysics*, Vol. 689, A166, 2024.
   Citation : `@ARTICLE{2024A&A...689A.166C}`
5. **Weaver, J.R.**, S. Taamoli, C.J.R. McPartland, et al. (including **M. Ezzaiti**), *Euclid : ERO – NISP-only sources and the search for luminous $z = 6 - 8$ galaxies*, *arXiv e-prints*, arXiv:2405.13505, 2024.
   Citation : `@ARTICLE{2024arXiv240513505W}`
6. **Atek, H.**, R. Gavazzi, J.R. Weaver, et al. (including **M. Ezzaiti**), *Euclid : Early Release Observations – A preview of the Euclid era through a galaxy cluster magnifying lens*, *arXiv e-prints*, arXiv:2405.13504, 2024.
   Citation : `@ARTICLE{2024arXiv240513504A}`
7. **Cuillandre, J.C.**, E. Bertin, M. Bolzonella, et al. (including **M. Ezzaiti**), *Euclid : Early Release Observations – Programme overview and pipeline for compact- and diffuse-emission photometry*, *arXiv e-prints*, arXiv:2405.13496, 2024.
   Citation : `@ARTICLE{2024arXiv240513496C}`
8. **Euclid Collaboration**, F.J. Castander, P. Fosalba, et al. (including **M. Ezzaiti**), *Euclid. V. The Flagship galaxy mock catalogue : a comprehensive simulation for the Euclid mission*, *arXiv e-prints*, arXiv:2405.13495, 2024.
   Citation : `@ARTICLE{2024arXiv240513495E}`

9. **Euclid Collaboration**, Y. Mellier, J.A. Acevedo Barroso, et al. (including **M. Ezzaiti**), *Euclid. I. Overview of the Euclid mission*, *arXiv e-prints*, arXiv:2405.13491, 2024.
Citation : `@ARTICLE{2024arXiv240513491E}`

Participation in conferences and summer schools :

1. *Naples, Italy, September 2023*
AGN and galaxies selection with the first Euclid data
Euclid workshop : Galaxies and AGN with the first Euclid data and beyond
2. *Paris, France, December 2022*
Selection of high-redshift quasars
Euclid France symposium Paris
3. *Online, November 2022*
Galaxies and AGN with the first Euclid data and beyond
Euclid Early careers, 2022 session
4. *Ringberg, Germany, April 2023*
Chair of the session on the Euclid Scientific Challenge 8, and presentation of this data
Euclid QSO workpackage meeting
5. *Copenhagen, Denmark, April 2023*
Plenary session talk (invited) : The selection of high-z quasars in Euclid
Euclid Consortium meeting
6. *Online, December 2022*
Future Euclid high-z quasars selection
Euclid Early careers, 2023 session
7. *Lyon, France, January 2024*
Selection of high-redshift quasars
Euclid France symposium Lyon
8. *Anglet, France, August 2021*
Euclid summer school of cosmology Cycle 1
9. *Banyuls-sur-Mer, France, August 2022*
Euclid summer school of cosmology Cycle 2

Résumé

La formation et l'évolution des trous noirs supermassifs (SMBH) dans l'Univers primordial représentent un défi majeur en astrophysique. Les quasars à très haut redshift ($z > 7$) constituent des laboratoires uniques pour explorer cette époque, apportant des contraintes essentielles sur la croissance des trous noirs, l'évolution précoce des galaxies et la réionisation cosmique. L'identification de ces objets rares dans les grands relevés photométriques, tels que le relevé *Euclid* Wide Survey (EWS), nécessite des techniques de sélection efficaces pour distinguer les quasars des contaminants, notamment les naines brunes et les galaxies elliptiques à redshift intermédiaire.

Cette thèse développe une méthodologie robuste pour la sélection des quasars à haut redshift, en s'appuyant sur une classification probabiliste des sources à partir de données photométriques. Le processus de sélection est optimisé pour équilibrer complétude et pureté, garantissant une efficacité maximale pour la confirmation spectroscopique. Je présente `Owl-z`, un outil bayésien de classification attribuant des probabilités aux quasars et à leurs principaux contaminants. Bien qu'optimisé pour *Euclid*, `Owl-z` est conçu pour être adaptable à d'autres relevés, tels que LSST, Roman et JWST, et peut être étendu à la sélection de galaxies à haut redshift.

Les performances de `Owl-z` ont été évaluées sur des simulations de données *Euclid* couvrant les redshifts $7 \leq z \leq 12$ et les magnitudes $18 \leq H_E \leq 24.5$. Les résultats montrent que le code atteint des valeurs élevées de complétude et de pureté pour les sources brillantes ($H_E \lesssim 22$), avec une dépendance minimale au redshift. Le seuil de probabilité de sélection peut être ajusté après traitement afin d'optimiser l'équilibre entre complétude et pureté et ainsi affiner la sélection des candidats pour les suivis spectroscopiques.

`Owl-z` a été validé sur des données réelles issues du programme *Euclid* Early Release Observations (ERO). Cette analyse a permis d'identifier un sous-ensemble de candidats quasars à haut redshift, tout en mettant en évidence les défis liés à la contamination par des artefacts observationnels, tels que la persistance et les détections erronées. De plus, `Owl-z` a correctement identifié une naine brune ($T7$), confirmée spectroscopiquement, démontrant ainsi sa capacité à distinguer quasars et contaminants stellaires.

Des améliorations futures incluent l'affinement des modèles de contaminants, l'intégration de critères morphologiques et l'ajout de techniques d'apprentissage automatique pour améliorer la classification des sources plus faibles. La découverte d'un échantillon important de quasars à $z > 7$ fournira des contraintes inédites sur la formation des SMBH, l'évolution précoce des galaxies et la réionisation cosmique. En sélectionnant les candidats les plus prometteurs, `Owl-z` jouera un rôle clé dans l'optimisation des suivis photométriques et spectroscopiques dans *Euclid* et d'autres grands relevés.

Mots-clés : Méthodes : statistiques ; Relevés ; Cosmologie : observations ; réionisation ; quasars : général.

Abstract

Understanding the formation and evolution of supermassive black holes (SMBHs) in the early Universe is a fundamental challenge in modern astrophysics. High-redshift ($z > 7$) quasars provide a unique window into this epoch, offering crucial insights into black hole growth, early galaxy evolution, and cosmic reionization. Detecting these rare objects in wide-field photometric surveys, such as the Euclid Wide Survey (EWS), requires efficient selection techniques capable of distinguishing quasars from contaminants, including brown dwarfs and early-type galaxies at intermediate redshifts.

This thesis develops a robust methodology for high-redshift quasar selection, focusing on the probabilistic classification of sources based on photometric data. The selection process optimizes the trade-off between completeness and purity, ensuring the best possible balance for subsequent spectroscopic confirmation. I present `Owl-z`, a Bayesian classification tool designed to assign probabilistic classifications to quasars and their main contaminants. While `Owl-z` has been optimized for Euclid, its flexible framework allows for adaptation to other surveys, such as LSST, Roman, and JWST, and can be extended to the selection of high-redshift galaxies.

The performance of `Owl-z` was extensively tested using Euclid simulations, covering redshifts $7 \leq z \leq 12$ and magnitudes $18 \leq H_E \leq 24.5$. The results confirm that `Owl-z` achieves high completeness and purity for bright sources ($H_E \lesssim 22$), with minimal dependence on redshift. The selection probability threshold can be adjusted after processing to fine-tune the balance between completeness and purity, optimizing the selection of candidates for spectroscopic follow-up.

`Owl-z` was further validated using real data from the Euclid Early Release Observations (ERO) program. This analysis confirmed a subset of high-redshift quasar candidates while highlighting the challenge of contamination by observational artefacts, such as persistence and spurious detections. Additionally, `Owl-z` successfully identified a brown dwarf ($T7$), later confirmed spectroscopically, demonstrating the tool's ability to distinguish between quasars and stellar contaminants.

Future improvements to `Owl-z` include refining contaminant models, incorporating morphological selection criteria, and integrating machine learning techniques to enhance classification accuracy for fainter sources. The discovery of a large sample of quasars at $z > 7$ will provide unprecedented constraints on SMBH formation, early galaxy evolution, and cosmic reionization. By selecting the most promising candidates, `Owl-z` will play a critical role in optimizing photometric and spectroscopic follow-ups in Euclid and other large-scale surveys.

Keywords: Methods: statistical; Surveys; Cosmology: observations; reionization; quasars: general.

Long résumé

L'étude de la formation des premières structures après le Big Bang repose sur des processus physiques complexes. Initialement, l'univers était un mélange chaud de particules fondamentales, empêchant toute formation atomique stable (Guth &Steinhardt 1984; Linde 1982). 380 000 ans après le Big Bang, les électrons et noyaux se sont recombinés, libérant le fond diffus cosmologique (Peebles 1993). Quelques centaines de millions d'années plus tard, la gravité a amplifié les fluctuations de densité ($\sim 10^{-5}$), entraînant l'effondrement des halos de matière noire et la formation des premières étoiles et galaxies, qui marque la fin des "âges sombres" et le début de la réionisation (Barkana &Loeb 2001a; Fan et al. 2006a).

La naissance des premières étoiles et galaxies coïncide également avec celle des premiers trous noirs, et en particulier les trous noirs dits supermassifs (SMBH, pour *SuperMassive Black Hole*), dont la masse dans l'univers local varie typiquement de millions à milliards de masses solaires. Les SMBHs se trouvent au cœur des galaxies et accrètent de la matière, un processus très énergétique et complexe, pouvant dominer l'émission de la galaxie hôte. Nous sommes alors en présence d'une "galaxie à noyau actif" ou AGN (pour *Active Galactic Nuclei*). Un AGN présente une structure qui englobe plusieurs composants clés (voir Figure 1.2) : un disque d'accrétion où le gaz et la poussière s'enroulent en spirale, s'échauffent et émettent un rayonnement intense (Sanders et al. 1989); une région produisant des raies d'émission larges (BLR, *Broad Line Region*) et une région produisant des raies d'émission étroites (NLR *Narrow Line Region*) (Kellermann et al. 1989); un tore de poussière entourant le noyau, et qui obscurcit l'AGN sous certains angles (Netzer 2015). On trouve parfois des jets de particules relativistes qui s'étendent loin au-delà de la galaxie hôte (Blandford &Rees 1978) (voir la figure 1).

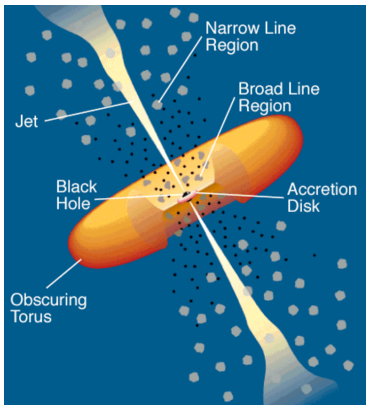


Figure 1 – Représentation schématique d'un AGN. Figure tirée de Urry &Padovani 1995.

La manière d'observer ces régions ainsi que leurs signatures spectrales diverses ont

conduit à une classification complexe des AGN, souvent appelée "zoo des AGN" (Padovani et al. 2017; Antonucci 1993; Urry &Padovani 1995). Dans cette thèse on se concentrera sur un type de AGN dit quasar.

Les quasars sont des AGN extrêmement lumineux, qui témoignent de l'accrétion de matière sur un SMBH. Leur luminosité exceptionnelle permet de les observer à des grandes distances cosmologiques, ce qui fait d'eux des outils clés pour explorer l'Univers jeune (Fan et al. 2001a; Ferrarese &Ford 2005a). Leur spectre couvre une large gamme de longueurs d'onde, du radio aux rayons X, avec une émission UV et optique dominée par le disque d'accrétion (Collinson et al. 2016). Leur abondance, quantifiée par leur fonction de luminosité, évolue avec le redhisft (z, décalage spectral vers le rouge), atteignant un pic à z$\sim$2 et déclinant à plus haut redshift (Boyle et al. 2000; Hasinger et al. 2005). Cette distribution en redshift suit globalement celle de l'histoire de formation stellaire dans les galaxies, et témoigne d'une co-évolution AGN/galaxies, dont les contours complexes sont explorés de nos jours à l'aide d'observations et de simulations cosmologiques.

De nombreux efforts ont été consentis par la communauté scientifique pour explorer l'époque de la réionisation, son étendue, et les processus physiques impliqués. L'observation du profil d'absorption Gunn-Peterson (Gunn &Peterson 1965a) dans les spectres de quasars à z$\sim$5.5-6.6 a permis de retracer l'évolution de la fraction d'hydrogène neutre, et de contraindre pour la première fois l'époque de la réionisation (Fan et al. 2001c, 2006b; Fanidakis et al. 2012). Pendant longtemps les quasars ont été considérés comme jouant un rôle clé dans la réionisation cosmique, à cause de leur intense rayonnement ionisant (McGreer et al. 2015). Cependant, leur faible abondance suggère que les galaxies de faible masse avec un fort taux de formation d'étoiles ont été les principales sources de photons ionisants (Robertson et al. 2015). L'origine et la croissance rapide des SMBH observés à z>7 par des observations récentes du James Webb Space Telescope posent un défi majeur aux modèles de formation des trous noirs (Volonteri et al. 2021). Deux scénarios sont envisagés : l'accrétion rapide sur des trous noirs stellaires issus des premières générations d'étoiles, ou la formation directe de "germes" massifs par effondrement gravitationnel (Madau et al. 2024).

Malgré leur forte luminosité, les quasars à très haut redshift restent rares et difficiles à identifier en raison de leur contamination par des objets rouges tels que les naines brunes et les galaxies poussiéreuses ou précoces (Wang et al. 2019). Avec l'arrivée des nouvelles générations de relevés comme *Euclid*, LSST (*Legacy Survey of Space and Time*) et le télescope spatial Roman, la détection de quasars à z>7 pourrait considérablement améliorer notre compréhension de l'évolution des SMBHs et du processus de réionisation (Euclid Collaboration et al. 2024a; Akeson et al. 2019). La détection et la sélection des quasars à haut redshift (z > 7) est un défi majeur en astrophysique observationnelle, notamment dans le cadre des grands relevés photométriques et spectroscopiques. L'approche Bayésienne permet une modélisation rigoureuse des incertitudes et une intégration efficace des connaissances a priori, facilitant ainsi l'identification des sources rares comme les quasars dans l'univers jeune (Jaynes &Bretthorst 2003; Sivia &Skilling 2006) .

Les méthodes classiques de sélection des quasars à z>7 dans les domaines optique et proche-IR s'appuient sur l'analyse des couleurs et des spectres pour discriminer les quasars des contaminants, tels que les naines brunes et les galaxies rouges à faible redshift

(Bañados et al. 2018; Fan et al. 2001b). Plus récemment, des techniques statistiques avancées ont été introduites, notamment l'utilisation de Random Forests (méthode de machine learning supervisé) et de modèles Bayésiens probabilistes tels que XDHZQSO, qui repose sur l'estimation de la densité des objets en espace de flux (Nanni et al. 2022).

Dans ce travail de thèse, j'ai adapté le formalisme Bayésien (Mortlock et al. 2012; Pipien et al. 2018b) à l'analyse des données du relevé *Euclid* Wide Survey (EWS), en affinant les modèles de contamination et en intégrant les dernières contraintes observationnelles sur les quasars et leurs contaminants (Mortlock et al. 2011; Wang et al. 2019). J'ai développé un outil appelé `Owl-z`, permettant d'assigner des probabilités aux candidats quasars à z> 7, et validé ses performances à l'aide de simulations et d'observations réelles. `Owl-z` a été appliqué aux premières données publiques d'*Euclid*, ouvrant la voie à une identification plus fiable et systématique des quasars les plus lointains.

En ce qui concerne le principe de OWL-z, la sélection Bayésienne permet d'estimer la probabilité qu'une source soit un quasar à haut redshift ou un contaminant, en combinant les probabilités à priori et la vraisemblance des observations. Le formalisme utilisé dans cette thèse est basé sur le travail de Mortlock et al. 2012. La probabilité a posteriori d'une source de type t ayant des données d'observation $\mathbf{D}$ est définie par le théorème de Bayes (nous utilisons la forme pondérée) :

$$\mathbb{P}(t|\mathbf{D}) = \frac{W(t|\mathbf{D})}{\sum_{t' \in \mathcal{T}} W(t'|\mathbf{D})} \tag{0.1}$$

avec $W(t|\mathbf{D}) = \mathbb{P}(t)\mathbb{P}(\mathbf{D}|t)$ représentant le poids de preuve. Ce poids est calculé pour chaque population (quasars à grand-z, étoile ou galaxie contaminante) afin d'évaluer la probabilité qu'une source soit un quasar ou pas. L'estimation des paramètres optimaux se fait par maximisation de la probabilité a posteriori :

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} P(D|\theta)P(\theta). \tag{0.2}$$

Les paramètres optimisés incluent le redshift, la magnitude et le type spectral.

Le prior des quasars est basé sur une fonction de luminosité en double loi de puissance en suivant Matsuoka et al. 2023 :

$$\Phi(M_{1450}, z) = \frac{10^{k(z-7)}\Phi^*}{10^{0.4(\alpha+1)(M_{1450}-M^*_{1450})} + 10^{0.4(\beta+1)(M_{1450}-M^*_{1450})}}. \tag{0.3}$$

La densité surfacique est exprimée par :

$$\rho_{\mathrm{QSO}}(H_E, z) = \frac{1}{4\pi} \times \frac{dV_c}{dz} \times \Phi(H_E - \mu - K_{\mathrm{corr}}(z), z). \tag{0.4}$$

Les principaux contaminants sont les naines brunes et les galaxies elliptiques à $1 < z < 2$. Les naines brunes de types spectraux MLT sont modélisées avec un disque galactique fin Caballero et al. (2008). En ce qui concerne les galaxies, une longue étude nous a permis de conclure que la nature des galaxies qui contaminent le plus sont les galaxies précoces à redshift intérmédiaire ($z \approx 1.5$). Nous avons donc modélisé leur densité numérique à l'aide de la fonction de luminosité, ajustée aux données COSMOS, en utilisant une fonction de

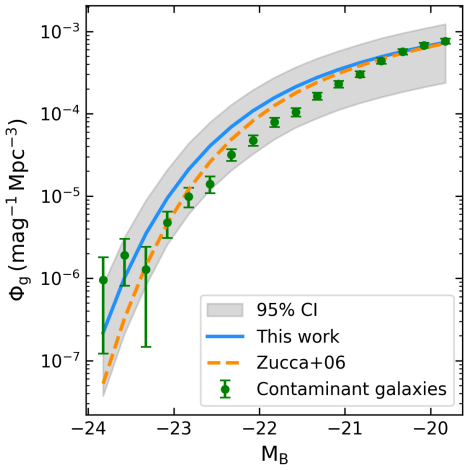


Figure 2 – Ajustement de la fonction de luminosité dans la bande B, avec son intervalle de confiance à 95 % *(gris)*, pour la population de galaxies de type précoce à $1 < z < 2$ dans les données du catalogue E-COSMOS *(vert)*, comparé à l'ajustement de la même fonction par Zucca et al. 2006 *(pointillés orange)*.

Schechter :

$$n(M) = (0.4 \ln 10)\phi^*[10^{0.4(M^*-M)}]^{\alpha+1} \exp\left[-10^{0.4(M^*-M)}\right]. \tag{0.5}$$

Cette évaluation indépendante rejoint les résultats obtenus auparavant par Zucca et al. 2006 pour ce type de galaxie, que nous avons poussé plus loin en magnitude (voir Figure. 2).

L'approche Bayésienne permet de gérer les incertitudes des observations et des modèles tout en optimisant la sélection des quasars. J'ai d'abord analysé puis validé les résultats de OWL-z sur des données simulées d'*Euclid*, puis sur des vrais quasars à z>7, et finalement je l'ai appliqué aux vraies premières données d'*Euclid*, qui sont encore préliminaires.

`Owl-z` est un code de classification Bayésienne flexible, conçu pour identifier les quasars à $z > 7$ dans des relevés photométriques. `Owl-z` compare les populations modélisées aux observations photométriques et attribue une probabilité à chaque objet d'appartenir à une classe donnée. D'abord, `Owl-z` a été appliqué à la sélection des quasars dans l'EWS, en utilisant les modèles de la section précédente pour les quasars et les contaminants. Dans le processus de modélisation, `Owl-z` génère des flux synthétiques pour diverses populations (quasars, étoiles de types spectraux MLT, galaxies) et leur applique un processus Bayésien afin de calculer des probabilités d'appartenance aux différentes populations d'objets.

L'évaluation des performances de l'outil repose sur trois métriques :

— **Complétude** : proportion des quasars correctement identifiés.

— **Pureté** : fraction des objets classés comme quasars qui le sont réellement.

— **F-mesure** : moyenne harmonique de la complétude et de la pureté.

`Owl-z` a été initialement appliqué à des données simulées du EWS. La complétude obtenue avec `Owl-z` sur dcs données est illustrée dans la Fig. 3. Le seuil de sélection des quasars a un impact direct sur la performance, comme illustré dans la Fig. 4 et discuté plus loin.

L'optimisation du seuil de sélection ζ, avec lequel on décide si oui ou non un objet est

un quasar ($P_q > \zeta$), permet d'améliorer les performances d'`Owl-z`. Une valeur plus faible maximise la complétude mais réduit la pureté, tandis qu'une valeur élevée améliore la pureté au détriment de la complétude. La Fig. 4 montre la variation de la F-mesure en fonction de ζ. L'adoption d'un seuil variable optimisé en fonction de la magnitude et du redshift permet une sélection plus efficace des quasars à $z > 7$.

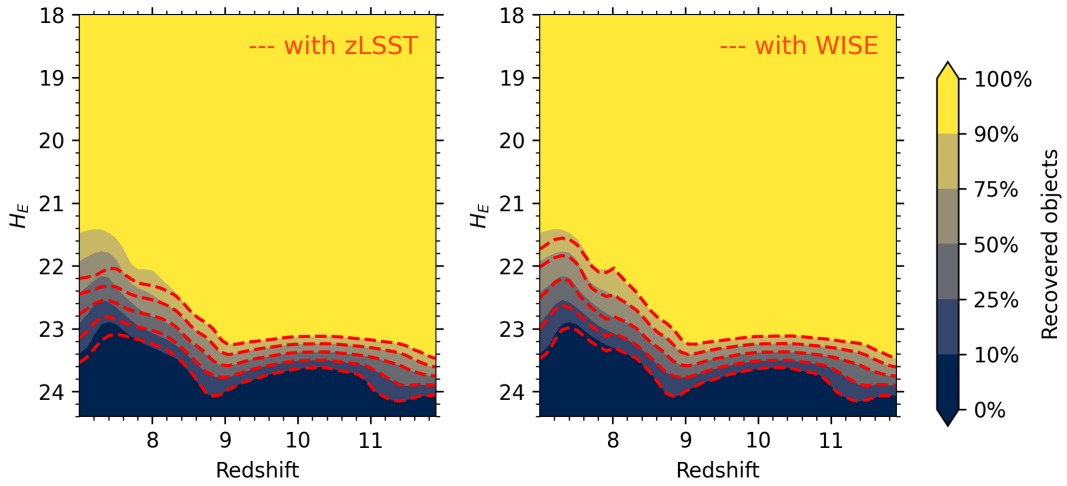


Figure 3 – Les fonctions de sélection *(en rouge)* des quasars à haut redshift déterminées à l'aide de `Owl-z` pour les données *Euclid* dans en utilisant les bandes $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$à gauche et z$Y_{\rm E}J_{\rm E}H_{\rm E}$ou z est la bande z de LSST. La fonction de sélection définit le niveau de complétude par bin de redshift ; la complétude d'un bin est calculée en comptant le pourcentage de quasars dans ce bin ayant une probabilité de $\mathbb{P} > 0.1$.

Ensuite, `Owl-z` a été appliqué à des données photométriques réelles, pour chacun des quasars à z>7 sélectionnés dans des relevés similaires et spectroscopiquement confirmé. Le Tableau 1 résume ces données photométriques, ainsi que le redshift de chaque quasar et celui calculé par mon code. `Owl-z` a correctement attribué une probabilité proche de 1 à l'ensemble de ces quasars, nonobstant le fait que les données photométriques utilisées étaient différentes de celles d'*Euclid*. La Figure 5 montre les résultats de l'ajustement

Les conclusions de cette première phase sur la performance de l'outil peuvent se résumer ainsi :

— `Owl-z` est un outil efficace pour identifier les quasars à $z > 7$ avec une approche Bayésienne flexible.

— La sélection des quasars peut être optimisée en adaptant le seuil ζ selon la magnitude et le redshift.

— L'outil est performant dans les relevés photométriques profonds et peut être appliqué à divers jeux de données.

En 2024, les premières données réduites d'*Euclid* ont été mises à disposition du consortium *Euclid* (et ensuite au public) dans le cadre du "Early Observation Release" (ERO). J'ai pu appliquer `Owl-z` à ces données dans le programme ERO, obtenant des résultats préliminaires sur la sélection des quasars à haut redshift.

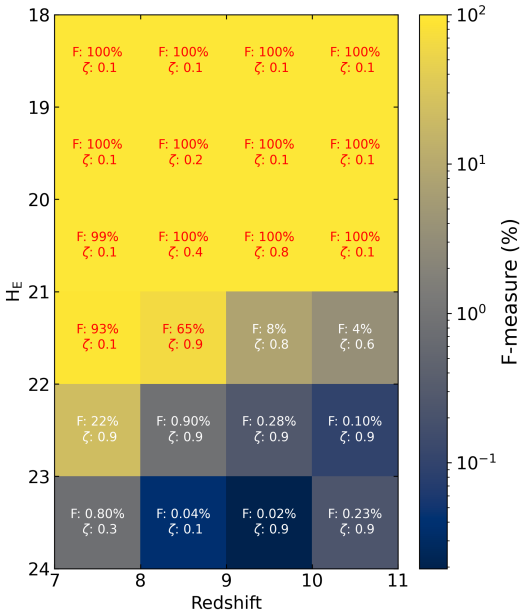


Figure 4 – Évolution de la F-mesure en fonction du seuil ζ. Un seuil plus bas augmente la complétude, et un seuil plus élevé améliore la pureté.

Table 1 – Quasars connus à z>7, sélectionnés dans des relevés similaires, et spectroscopiquement confirmés. On montre les données photométriques utilisées pour la découverte et les redshifts spectroscopiques ($z_{\rm spec}$). Les colonnes $z_{\rm out}$ et P_q montrent respectivement le redshift et la probabilité d'être un quasar calculés par `Owl-z`.

QSO name	Photometry	$z_{\rm spec}$	$z_{\rm out}$	Depth in ref band	Reference
J1342+0928	$z_{\rm DECam}$, Y,J,H,K, $W1$,$W2$	7.54	7.8	J 20.52 5σ	Bañados et al. (2018)
J1120+0641	$z_{\rm DECam}$, Y,J,H,K,$W1$,$W2$	7.08	7.05	J 20.52 5σ	Mortlock et al. (2011)
J0313-1806	$z_{\rm PS1}$,$y_{\rm PS1}$,J,K	7.64	7.7	J 20.52 5σ	Wang et al. (2021)
J0038–1527	$z_{\rm DECam}$,$y_{\rm PS1}$, Y,J,H,K, $W1$,$W2$	7.02	6.95	J 20.52 5σ	Wang et al. (2018)
J1007+2115	$z_{\rm DECam}$,$y_{\rm PS1}$, Y,J,H,K, $W1$,$W2$	7.51	7.65	J 20.52 5σ	Yang et al. (2020)
J0252–0503	$z_{\rm DECam}$, Y, J, $W1$, $W2$	7.02	6.9	J 20.52 5σ	Yang et al. (2019)
J1243+0100	$g_{\rm HSC}$, $r_{\rm HSC}$, $i_{\rm HSC}$, $z_{\rm HSC}$, $y_{\rm HSC}$	7.07	9.9	$y_{\rm HSC}$ 24.4 5σ	Matsuoka et al. (2019b)
J2356+0017	$i_{\rm HSC}$, $z_{\rm HSC}$, $y_{\rm HSC}$	7.01	6.75	$y_{\rm HSC}$ 24.4 5σ	Matsuoka et al. (2019a)

Le programme ERO illustre les capacités d'*Euclid* avant son relevé principal (Cuillandre et al. 2024). Parmi les projets, le programme *Magnifying Lens* cible les amas Abell 2390 et Abell 2764 (Atek et al. 2024). Ces amas amplifient les galaxies d'arrière-plan, facilitant ainsi la détection de galaxies à $z > 6$ (Weaver et al. 2024). Les observations d'Abell 2390 et Abell 2764 incluent trois séquences d'observation de 70 minutes, couvrant les filtres I_E, Y_E, J_E, et H_E. Les images obtenues ont une résolution native de 0.1″ pixel^{-1} pour VIS (imageur dans une partie du domaine optique) et 0.3″ pixel^{-1} pour NISP spectrophotomètre couvrant une partie du proche infrarouge) (Cuillandre et al. 2024). La détection des sources a été effectuée avec `SEP` (Bertin &Arnouts 1996), identifiant 501 994 objets sur deux champs de 0.75 deg^2 .

Les candidats quasars à $z > 6$ ont été pré-sélectionnés en combinant des critères de

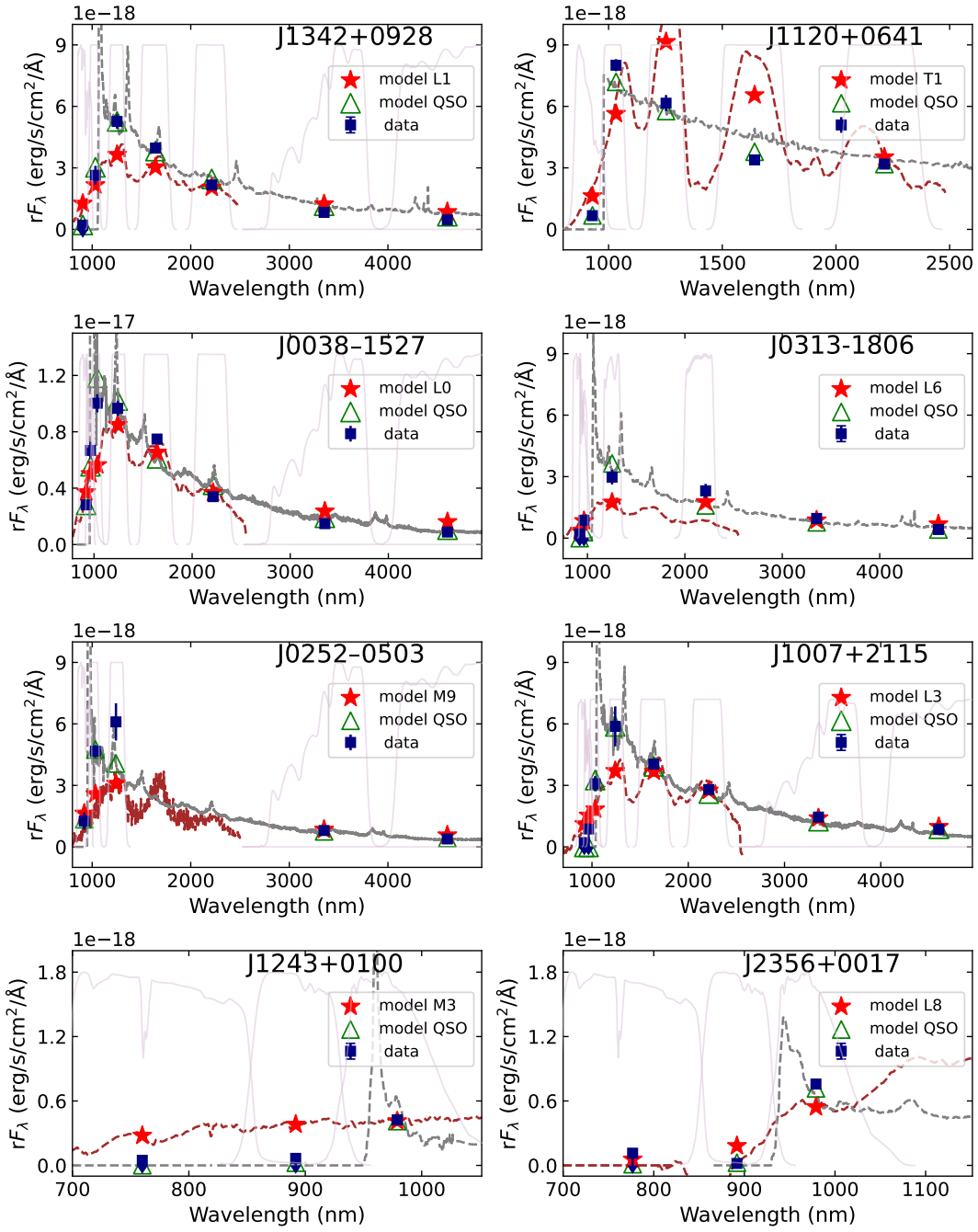


Figure 5 – Résultat de l'ajustement par `Owl-z` des quasars à z>7 montrés dans le Tableau 1. On montre le meilleur ajustement par des modèles de quasars (lignes discontinues grises) et d'étoiles MLT (lignes marron). Les points photométriques d'origine sont en bleu, et le meilleur ajustement pour les modèles de quasar et d'étoile sont en vert et rouge respectivement.

couleur et de rapport signal-sur-bruit (SNR). Seuls les objets détectés dans H_E avec $\mathrm{SNR}_{H_E} > 2$, et non détectés en I_E ($\mathrm{SNR}_{\mathrm{VIS}} < 2$) ont été conservés. Après application de

ces critères, 11 963 sources ont été retenues pour analyse avec `Owl-z`. L'exécution d'`Owl-z` sur les catalogues combinés d'Abell 2390 et Abell 2764 a permis d'identifier 292 candidats avec $P_q > 0.1$, et 187 avec $P_q > 0.9$. 78% des sources présentaient une probabilité de galaxie $P_g > 0.1$ et seulement 3% une probabilité de naine brune $P_s > 0.1$.

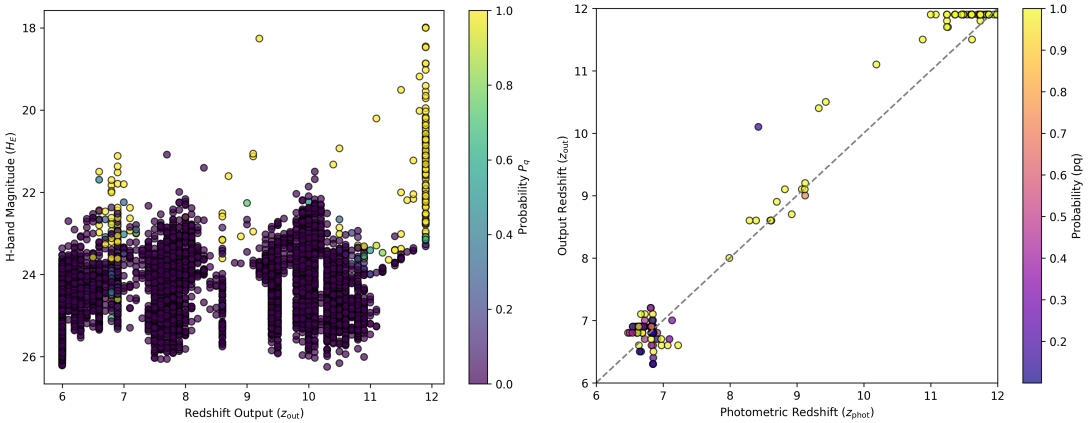


Figure 6 – À gauche : redshift z_{out} vs. magnitude H_E des candidats quasars sélectionnés. À droite : comparaison des redshifts `Owl-z` avec ceux issus de Weaver et al. (2024). Après inspection visuelle, tous les candidats à z∼12 ont été éliminés comme étant des artefacts.

Les catalogues ERO étant entachés de fausses détections et des résidus de persistance dans les images proche-IR, une inspection visuelle des candidats sélectionnés par le code a été nécessaire. Elle a révélé que l'échantillon était dominé par ces artefacts. Dans Abell 2390, seuls 17 quasars potentiels ont été retenus après inspection (Fig. 7), le reste étant des d'artefacts instrumentaux (voir quelques exemples sur Fig. 8).

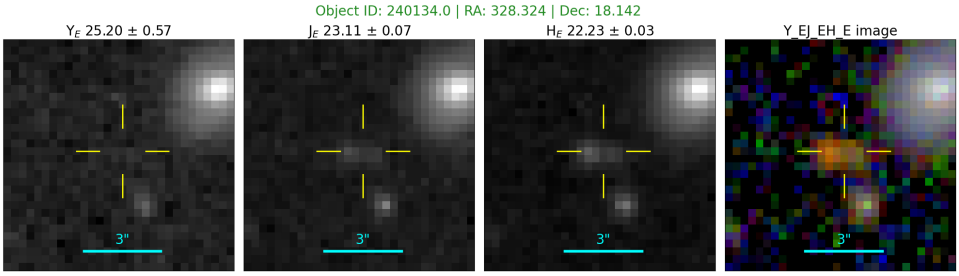


Figure 7 – Exemple de sources sélectionnées par `Owl-z` et apparaissant comme des objets astrophysiques réels.

L'analyse des données ERO a permis de tester `Owl-z` sur des données réelles. Les résultats montrent une capacité à identifier les quasars, bien que l'inspection visuelle reste nécessaire pour éliminer les artefacts. L'amélioration de la qualité des catalogues d'entrée et la réduction des contaminations instrumentales seront essentielles pour le relevé principal d'*Euclid*.

Cette thèse, menée dans le cadre du consortium *Euclid*, a porté sur le développement et l'optimisation d'`Owl-z`, un outil Bayésien de sélection de quasars à $z > 7$ dans les

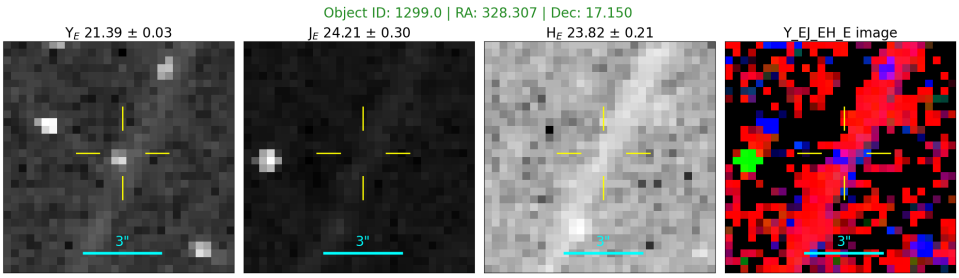


Figure 8 – Exemple d'artefacts détectés comme faux positifs par `Owl-z`.

relevés photométriques. Son efficacité a été démontrée par la réidentification de tous les quasars spectroscopiquement confirmés à $z > 7$, et par des performances optimales sur les simulations de données *Euclid*.

L'application d'`Owl-z` aux premières données *Euclid* (ERO) a validé son utilisation sur des observations réelles, tout en mettant en évidence ses limites face aux artefacts instrumentaux. L'identification d'une naine brune via `Owl-z` et sa confirmation spectroscopique montrent cependant la robustesse du code dans la distinction entre quasars et contaminants stellaires.

Le travail réalisé a mis en évidence quelques perspectives d'ordre technique. L'application d'`Owl-z` à d'autres relevés (LSST, Roman, JWST) permettrait d'élargir son champ d'application, avec l'amélioration des modèles de naines brunes pour affiner la sélection. L'intégration de l'apprentissage automatique pourrait améliorer la pureté des échantillons de candidats.

L'ajout d'une contrainte morphologique offrirait aussi une meilleure distinction entre quasars et galaxies. Il serait également essentiel d'automatiser la détection des artefacts pour réduire la contamination, tout en optimisant les critères photométriques de sélection et l'estimation photométrique du redshift, en particulier pour $z > 9$.

Ce travail ouvre aussi de nombreuses perspectives scientifiques. La découverte de plusieurs centaines de quasars à $z > 7$ transformerait notre compréhension de l'Univers jeune. Elle permettrait de mieux contraindre les scénarios de formation des trous noirs supermassifs et d'étudier l'évolution des galaxies hôtes, le rôle des quasars dans la réionisation et l'enrichissement chimique du milieu intergalactique. Ces quasars fourniraient aussi des cibles idéales pour des observations spectroscopiques avec JWST et l'ELT. Leur distribution spatiale permettrait de cartographier la structure à grande échelle de la réionisation, et même d'explorer des écarts éventuels au modèle ΛCDM. Enfin, leur spectroscopie haute résolution pourrait permettre de révéler des variations des constantes fondamentales, ouvrant la porte à de nouvelles contraintes sur la physique fondamentale.

Mots clés: quasars - statistique Bayésienne - univers à haut redshift - cosmologie observationelle

Acknowledgements

I would first like to warmly thank the reviewers in the jury of my PhD defense, Giulia Rodighiero and Hervé Dole. Their careful reading, insightful comments, and constructive suggestions were essential in improving this work.

I would also like to thank the examiners, Marta Volonteri, Julyan Arbel, and Hakim Atek. Marta's publications opened my eyes to new perspectives and deepened my love for active galactic nuclei. Julyan took the time to read my manuscript and genuinely got excited about my project. Hakim provided guidance throughout my thesis follow-up committees, offering invaluable advice and encouraging words along the way.

I am very grateful to my thesis supervisors, Roser Pelló and Jean-Gabriel Cuby, Pierre Pudlo and François-Xavier Dupé, the supervision was sometimes complex, but it ultimately led to the completion of a fruitful thesis. Their guidance, patience, and scientific rigor were essential to this work.

I want to express my sincere gratitude to my colleagues from the GECO team at LAM. Olivier, Sylvain, Patrice, Raphael, Stéphane, your encouragement, stimulating discussions, and your guidance has been invaluable throughout my PhD. Thank you especially for helping me understand, often from first principles, important astronomical concepts and methods. Thank you all for your kindness, patience, and support not only helped me progress scientifically but also made my experience at LAM truly joyful.

Finally, I want to thank all the members of the Laboratoire d'Astrophysique de Marseille for providing such a stimulating environment. Their encouragement, good humour, and daily scientific discussions made these years of my PhD memorable.

Je tiens à adresser un remerciement tout particulier à Françoise Maxant, qui est sans doute l'une des personnes les plus formidables et attentionnées que j'aie jamais rencontrées. Sa gentillesse, sa bienveillance et son soutien ont illuminé mes journées pendant ma thèse (et après). Même un simple " bonjour " échangé avec elle suffisait à me faire sentir bien. Ses mots encourageants, son sourire et sa disponibilité ont été un énorme soutien moral tout au long de ce parcours. Merci du fond du cœur Françoise.

Je souhaite également remercier Jean-Charles Lambert, à qui je dois non seulement le bon fonctionnement de mon code (fruit de mon travail de thèse) mais aussi, en grande partie, mon équilibre mental tout au long de ce parcours. Ses longues discussions, ses conseils professionnels et son accompagnement attentif ont été précieux.

Je tiens également à remercier Laurent Jorda, avec qui j'ai pu construire une belle amitié professionnelle fondée sur notre amour commun pour la diffusion scientifique. Nous avons travaillé ensemble lors de la Fête de la Science et à d'autres occasions, transmettant notre passion au grand public. Ces expériences ont été très enrichissantes et, grâce à lui, j'ai pu saisir de nombreuses opportunités pour partager mes connaissances scientifiques. Sa présence et son soutien m'ont été extrêmement précieux pour prendre du recul par rapport à ma thèse et m'ouvrir à d'autres perspectives.

Je souhaite également remercier toute l'administration, et en particulier Solange, Miguel et Christine. Leur disponibilité, leur gentillesse et leurs conseils ont été d'une grande aide tout au long de ma thèse.

To my fellow phd students and others:

First and foremost, I want to thank Diana and Gayathri. My PhD would have never been possible without them. They listened to me, supported me, and tolerated me in my lowest moments; always with patience, kindness, and humour. They saw me when I was overcaffeinated, scattered, anxious, or exhausted, and somehow still managed to love me through all of it. Even when distance was imposed on us after they left for their postdocs, our friendship remained strong; we kept in touch, sharing everything from frustrations to laughter, and they never stopped being a grounding presence in my life. The bond we built goes far beyond work: they are truly my chosen family.

Another memeber of my acquired family: Camille, she helped me make sense of everything when nothing seemed to make sense at all. She completes me in a way that's hard to describe. I hope she knows, through these words, how much her friendship means to me and how deeply she shaped this journey.

Martin deserves a very special mention: my office mate during the hardest part of my PhD and a fellow Euclid member. We travelled together for work, shared long discussions, and supported each other through the ups and downs of research life. He was the sound, stable presence I always needed ,calm, reliable, and endlessly kind. Beyond being a wonderful officemate, he was also my running partner, helping me stay motivated when things felt overwhelming. His companionship brought balance to my days, and I am deeply grateful for that.

To Ilias, Sia, Paolo, Alexis, Mathilde and all the other PhD students and postdocs: thank you for all the wonderful social moments we shared. From dinners and barbecues to Eurovision evenings, poker nights, metal concerts (James ;)) and game nights, you made life outside of research vibrant and full of laughter. The hikes we did together, especially the long and hard ones! were unforgettable adventures that bonded us in the best way possible. And truly, to all the PhD students: every outing, every trip, every shared moment made this experience richer and more human.

To all of you: thank you for being there, for making me laugh, for helping me hold on, and for reminding me that even in the most demanding times, friendship and connection are what truly carry us forward. I end this with a famous Gandalf quote: "All we have to decide is what to do with the time that is given us", so "Allons-y, Alonso!" (10th).

Contents

Affidavit **2**

List of publications and participation in conferences **3**

Résumé **5**

Abstract **6**

Long résumé **7**

Acknowledgements **16**

Contents **18**

List of Figures **21**

List of Tables **24**

1 Introduction **25**
1.1 The Active Galactic Nuclei "zoo" . . . 27
1.1.1 Short historical introduction . . . 28
1.1.2 AGN spectral energy distribution . . . 28
1.1.3 AGN Classification . . . 30
1.1.4 Basic phenomenological description . . . 32
1.1.5 Selecting AGN . . . 34
1.1.6 Evolution of the global properties of AGN across cosmic times . . . 35
1.1.7 Quasars and SMBH at the epoch of the reionisation . . . 41
1.2 The selection of astrophysical sources using a Bayesian approach . . . 45
1.2.1 Historical introduction . . . 45
1.2.2 Classical selection methods for high redshift quasars . . . 45
1.2.3 Recent developments . . . 46
1.2.4 Bayesian model comparison applied to the *Euclid* Wide Survey . . . 49

2 The Euclid Mission **51**
2.1 Overview . . . 51
2.2 Observing modes . . . 52
2.3 Survey characteristics . . . 53
2.4 Euclid: a beacon to uncover the mysteries of Dark Matter and Dark Energy 54
2.4.1 Mission objectives . . . 54

2.4.2 Dark matter . . . 55
2.4.3 Dark energy . . . 55
2.5 Legacy Science of Euclid: Insights into the Primeval Universe . . . 56

3 Bayesian Comparison model **57**
3.1 Introduction . . . 58
3.2 Principles . . . 58
3.2.1 Formalism . . . 58
3.2.2 Gaussian truncation . . . 60
3.2.3 The likelihood form . . . 62
3.2.4 Extraction of the Maximum A Posteriori parameters . . . 63
3.3 Bayesian approach to the search for high redshift quasars . . . 63
3.3.1 High-redshift quasars population model . . . 64
3.3.2 MLT stars population model . . . 65
3.3.3 Contaminant galaxy model . . . 72
3.3.4 Unbalanced prior problem . . . 80

4 `Owl-z`: A flexible tool to select high redshift quasars **82**
4.1 `Owl-z`: Ingredients and outputs . . . 83
4.2 `Owl-z`: how it works . . . 85
4.2.1 Configuration . . . 85
4.2.2 Model building . . . 86
4.3 Performance estimators . . . 86
4.3.1 Completeness . . . 86
4.3.2 Purity . . . 87
4.3.3 F-measure . . . 87
4.4 Re-identifying known $z > 7$ quasars . . . 88
4.5 Performance of `Owl-z` based on simulated catalogues . . . 89
4.5.1 Data . . . 89
4.5.2 Completeness . . . 92
4.5.3 Redshift estimation . . . 95
4.5.4 Purity and F-measure . . . 96
4.6 Discussion . . . 98
4.6.1 Influence of the selection threshold . . . 98
4.6.2 Optimising the identification of $z > 7$ quasars . . . 100
4.6.3 Influence of the position in the EWS footprint . . . 101
4.6.4 Influence of thick disc MLT . . . 103
4.6.5 Comparison with previous works . . . 103
4.7 Conclusions . . . 104

5 Application of `Owl-z` to Euclid data **108**
5.1 The Euclid Early Release Observations Program . . . 108
5.1.1 Observations . . . 109
5.1.2 Photometric catalogue . . . 110
5.1.3 Preselection of high-redshift quasar candidates . . . 111

5.2 Results 112
5.2.1 Redshift estimation 114
5.2.2 Visual inspection 115
5.2.3 Spectroscopic confirmation of a Brown Dwarf 120
5.3 Discussion 123

6 Conclusions and Perspectives **124**
6.1 Conclusions 124
6.2 Technical perspectives 125
6.3 Scientific perspectives 127

ANNEXES **129**

Acknowledgements **169**

List of Figures

1 Représentation schématique d'un AGN 7
2 Évolution de la fonction de luminosité des galaxies de type précoce à $1 < z < 2$ dans E-COSMOS 10
3 Complétude de `Owl-z` en fonction du redshift et de la magnitude H_E . . 11
4 Évolution de la F-mesure en fonction du seuil ζ 12
5 Résultat de l'ajustement par `Owl-z` des quasars à z>7 montrés dans le Tableau 1. On montre le meilleur ajustement par des modèles de quasars (lignes discontinues grises) et d'étoiles MLT (lignes marron). Les points photométriques d'origine sont en bleu, et le meilleur ajustement pour les modèles de quasar et d'étoile sont en vert et rouge respectivement. 13
6 redshift $z_{\rm out}$ vs. magnitude H_E des candidats quasars et la comparaison des redshifts `Owl-z` avec les z_{phot} 14
7 Exemple de sources sélectionnées par `Owl-z` et apparaissant comme sources réelles. 14
8 Exemple d'artefacts détectés comme faux positifs par `Owl-z`. 15

1.1 Cosmic timeline of the universe after recombination along with the stages of reionisation. 26
1.2 Schematic representation of an AGN. 27
1.3 A simplified schematic diagram of an AGN SED. 29
1.4 Evolution with redshift of the LF for quasars in the 2dF QSO Redshift Survey. 37
1.5 The space density of AGNs as a function of redshift in different luminosity classes and the sum over all luminosities. 38
1.6 Graphic visual of the Gunn-Peterson effect. 40
1.7 Left panel: Fiducial reionisation histories for the AGN-dominated models discussed in (Madau et al. 2024). Right panel: Constraints on the ionisation state of the IGM from a variety of probes and techniques. 43
1.8 SED Models and Simulated Spectra of High-z Quasars (Left) Input SED models with damping wing profiles for different x(HI) values. (Right) Simulated Blue Grism spectra for z = 7.5 quasars assuming deep survey depth. 49

2.1 Euclid Wide Survey coverage and colour-coded yearly progress in an all-sky Mollweide projection. 52
2.2 Spectral transmission of *Euclid*'s imaging (VIS: $I_{\rm E}$; NISP: $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$) and spectroscopic channels (NISP: Blue grism $\rm BG_E$ and Red grism $\rm RG_E$) at the beginning of the mission. 53

3.1 The expected number of quasars per redshift bin in the 15,000 deg^2 EWS 65
3.2 Sketches illustrating the Milky Way and the galactic coordinate system. . 66
3.3 Brown dwarf number counts as a function of H-band magnitude for the thin and thick discs. 68
3.4 Ratio of the surface density of brown dwarfs of spectral type M6 to T8 to the surface density of $z > 7$ quasars over the *Euclid* footprint shown in Galactic coordinates and galactic projection. 71
3.5 Templates as a function of redshift colour-coded by the density of galaxies that are selected as quasars. On the left, the Starburst99 template results, and on the right, the BC03 template results. 74
3.6 The *Euclid* $Y - J$, $J - H$ colour-colour scatter diagram illustrates the various populations in this study. 75
3.7 The redshift and extinction distributions for the Starburst99 first template, with two populations distinguished by their colours. 76
3.8 The extinction and redshift distributions for the BC03 first template (SSP), with two populations distinguished by their colours. 77
3.9 The luminosity function fit in the B-band for the early-type population. . 79
3.10 Surface densities comparison between galaxies and quasars. 80
3.11 Ratio of MLT to $z > 7$ quasar density as a function of apparent magnitude in two different locations of the *Euclid* footprint. 81

4.1 `Owl-z` logo 83
4.2 Flowchart describing the `Owl-z` Bayesian model comparison code. 85
4.3 WISE W1 − W2 colours of MLT dwarfs from Best et al. (2017). 89
4.4 Photometric data for the eight spectroscopically confirmed quasars at $z > 7$ and the `Owl-z` best-fit solutions for quasars and MLT dwarfs. 90
4.5 The location of the focus regions used to compute the purity. 93
4.6 The selection functions of high redshift quasars determined using `Owl-z`. 93
4.7 Tracks of $J_{\rm E}$-$H_{\rm E}$ colours as a function of redshift of Early-Type galaxies (left panel) and high redshift quasars (right panel). 96
4.8 Top panel shows the comparison between the true redshift and the redshift estimated by `Owl-z`. The bottom panel shows the residual $\Delta z = z_{out} - z_{true}$ as a function of redshift. 97
4.9 Classification performance of `Owl-z` determined using the threshold $\zeta = 0.1$. 98
4.10 Same is in Fig 4.9, for a threshold $\zeta = 0.2$ 98
4.11 Same is in Fig 4.9, for a threshold $\zeta = 0.3$ 99
4.12 Same is in Fig 4.9, for a threshold $\zeta = 0.4$ 99
4.13 Same is in Fig 4.9, for a threshold $\zeta = 0.5$ 100
4.14 Same is in Fig 4.9, for a threshold $\zeta = 0.6$ 100
4.15 Same is in Fig 4.9, for a threshold $\zeta = 0.7$ 101
4.16 Same is in Fig 4.9, for a threshold $\zeta = 0.8$ 101
4.17 Same is in Fig 4.9, for a threshold $\zeta = 0.9$ 102
4.19 F-measure evolution as a function of the $H_{\rm E}$ magnitude in each of the regions shown in Table 4.5, computed in the $7 < z < 8$ interval. 102

4.18 Classification performance of `Owl-z` on the magnitude $H_{\rm E}$ and redshift z_{out} plane. The threshold value of ζ that maximises F-measure is given for each bin. 106
4.20 F-measure evolution as a function of the $H_{\rm E}$ magnitude in each of the regions shown in Table 4.5, computed in the $7 < z < 8$ interval. 107

5.1 The output redshift ($z_{\rm out}$) versus H-band magnitude (H_E) for preselected ERO high-redshift quasar candidates. 113
5.2 Comparison of photometric redshift ($z_{\rm phot}$) computed by Weaver et al. (2024) with the output redshift ($z_{\rm out}$) from OWL-z. 114
5.3 Cutout images (30 × 30 pixels) of sources selected as z>6 quasars by `Owl-z`, and visually resembling real red sources. 115
5.4 Same as in Fig. 5.3, for dubious sources displaying a weird profile or suffering from contamination. 118
5.5 Cutout images (30 × 30 pixels) of artefacts identified in the Euclid NISP images. And misclassified by Owl-z. 119
5.6 The euc328+18_id222452 source, initially identified in Weaver et al. (2024) and suggested to be an artefact due to persistence, and correctly classified by `Owl-z`. 121
5.7 Observed smoothed spectrum of euc328+18_id222452 and best-fit template of T1 brown dwarf. 122

List of Tables

1 Quasars connus à z>7, sélectionnés dans des relevés similaires, et spectroscopiquement confirmés. On montre les données photométriques utilisées pour la découverte et les redshifts spectroscopiques ($z_{\rm spec}$). Les colonnes $z_{\rm out}$ et P_q montrent respectivement le redshift et la probabilité d'être un quasar calculés par `Owl-z`. 12

1.1 The AGN zoo: list of the major classes of AGN. 33

3.1 Quasar Luminosity function parameters 65
3.2 MLT density at the Galactic plane, and model colour data from Euclid Collaboration: Barnett et al. (2019). 70
3.3 Comparison of the Schechter function parameters obtained in this work with the results of Zucca et al. (2006). 78

4.1 Known spectroscopically confirmed quasars at $z > 7$ and their photometric data and spectroscopic redshift. 88
4.2 Scenarios explored in the simulations for the completeness estimation. . . 91
4.3 Photometric bands and depth. The *Euclid* bands characteristics. 94
4.4 Scenarios explored in the simulations for the completeness estimation. . . 94
4.5 Properties of areas and catalogues simulated across the EWS footprint. . 104

1 Introduction

Imagine the universe in its very first moments, right after the Big Bang—when everything we see today was condensed into an extremely hot, dense point. This primaeval universe had no atoms, light, or structure. It was a chaotic soup of fundamental particles—quarks, gluons, electrons—so tightly packed that they could not form anything stable. However, as the universe expanded, it began to cool down, and gradually, these particles started to combine. Quarks formed protons and neutrons, which eventually joined to create the nuclei of the first atoms. Still, it was too hot for electrons to settle down into orbits around nuclei, so the universe remained a seething plasma for some time, opaque to light (Guth &Steinhardt 1984; Linde 1982).

About 380,000 years after the Big Bang, the universe was cold enough for electrons to combine with protons and form neutral hydrogen atoms in an event known as "recombination". This was when light could finally travel freely across space, marking the beginning of the Cosmic Microwave Background (CMB), a faint glow that we can still see, giving us a snapshot of this ancient time (Peebles 1993). Nevertheless, it was still dark for hundreds of millions of years, filled with neutral hydrogen devoid of stars to illuminate it. This era is often called the 'Dark Ages'. Slowly, gravity began to work its magic. Although the universe at recombination was highly uniform, small spatial fluctuations in the dark matter density field ($\sim 10^{-5}$), generated in the early universe, started to grow over time due to gravitational instability. These small density fluctuations grow over time until they lead to the dynamical collapse of dark-matter halos, beginning the formation process of bounded objects, stars, galaxies and large-scale structures. Gas infalls and cools in these dark-matter halos until the first stars ignition and the first galaxies' assembly. This marked the beginning of the epoch of reionisation, roughly 400 million years after the Big Bang, when these first stars and galaxies produced intense ultraviolet light that started to ionise the surrounding neutral hydrogen gas, making the universe transparent once again (Barkana &Loeb 2001a; Fan et al. 2006a). By the end of this era, bright stars and galaxies filled the universe, and the cosmos as we know it today began to take shape (see Figure 1.1).

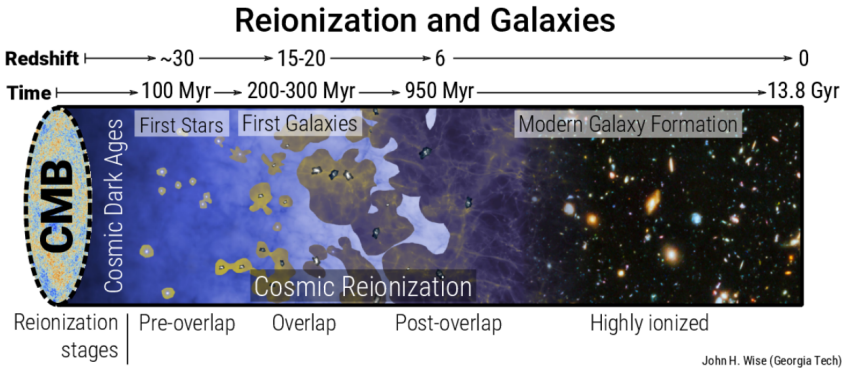


Figure 1.1 – Cosmic timeline of the universe after recombination along with the stages of reionisation. Source: Wise 2019.

The birth of the first stars also marks the birth of the first black holes, which could have given rise to "mini-quasars" with complex spectra growing by the accretion of surrounding material. The impact of these objects on the reionisation process is expected to be high. It has been discussed by different authors during the last 20 years (see, e.g. Haiman et al. 2000; Barkana &Loeb 2001a, and the references therein). However, there is currently very little observational evidence for the formation of supermassive black holes at very high redshift (z>7). Indeed, the number of spectroscopically confirmed sources unambiguously identified as quasars at z>7 is dramatically small. Here, "quasars" stands for objects whose emission is dominated by the accretion of matter into a massive black hole.

In the following, Active Galactic Nuclei (AGN) define an astrophysical source powered by the accretion onto a supermassive black hole (SMBH). The electromagnetic spectrum of galaxies hosting such an accreting black hole in their cores is dominated by this emission, covering the entire electromagnetic spectrum over more than twenty orders of magnitude in frequency.

Quasars are bright objects, but their identification at the highest redshifts is particularly challenging. This thesis is motivated by the arrival of massive cosmological surveys, such as *Euclid* (Laureijs et al. 2011; Euclid Collaboration et al. 2024a), the Nancy Grace Roman Space Telescope (Akeson et al. 2019) and the Large Synoptic Survey Telescope (LSST) (Ivezić et al. 2019a), which will provide sufficiently extensive data to unearth these rare objects. Whatever the outcome of this search for z>7 quasars, it will provide essential information for understanding the formation of the first stars and black holes and their impact on the reionisation process.

This thesis focuses on identifying the most distant AGNs formed in the universe, in particular during the epoch of the reionisation. We have developed a Bayesian comparison method to select such sources in wide photometrical surveys for this need. In this chapter, we provide an introduction to both the astrophysical sources targeted in this study (Sect. 1.1), and the specific method developed for this need (Sect. 1.2).

1.1 The Active Galactic Nuclei "zoo"

Active galactic nuclei (AGN) are galaxies whose emission is dominated by an extraordinarily energetic region found in their core. This central region hosts a supermassive black hole (SMBH) that powers the AGN. The emitting region of an AGN is complex, showing several different components, as shown in Figure 1.2:

— An accretion disk, where gas and dust spiral inwards, heating up and emitting intense radiation.
— The broad-line (BLR) and narrow-line regions (NLR), where fast-moving (for BLR) and slower-moving (for NLR) gas clouds produce emission lines.
— A dust torus surrounding the nucleus obscures the AGN from certain angles.
— The relativistic particles' jets extend far into space beyond the host galaxy.

The observability of these regions and the signatures induced in the electromagnetic spectra motivate the classification of AGN in different "families" presented in Sect. 1.1.3. According to the unified model of AGN, the different AGN types are mainly associated with the different orientations of the accretion disc concerning the observer (see, e.g. Antonucci 1993; Urry &Padovani 1995).

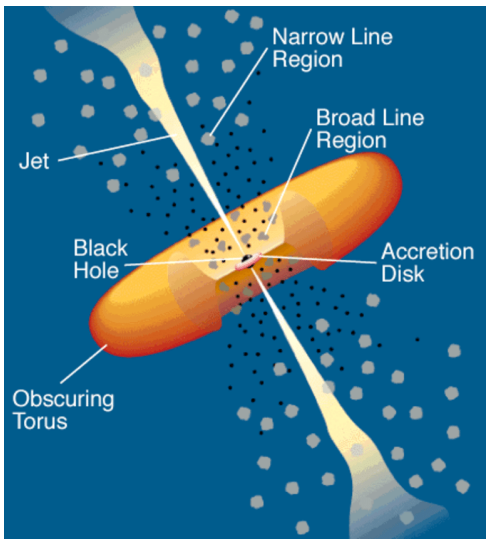


Figure 1.2 – Schematic representation of an AGN. Figure from Urry &Padovani 1995

SMBHs in AGNs have masses typically ranging from millions to billions of solar masses, drawing in surrounding gas and dust, releasing vast amounts of energy across the electromagnetic spectrum, from radio waves to gamma rays, making AGNs some of the most luminous objects in the universe. More details about the spectral energy distribution of AGNs are provided in Sect. 1.2.2. The radiation emitted by the AGN is so intense that it can outshine the host galaxy, leading to various observational signatures. The variety of these observational signatures has resulted in a complex classification system often called the "AGN zoo" (Padovani et al. 2017). Some authors have proposed different ways to simplify this complex observational view (see, e.g. Brotherton 2014).

1.1.1 Short historical introduction

The study of AGN dates back to the 20th century. The discovery of the first quasar, 3C 273, in 1963 by Maarten Schmidt (Schmidt 1963) marked the beginning of a new era, as it was the first time an object outside our galaxy was recognised as the source of such exceptional luminosity. The path leading from the observation of the first quasars to the idea of an accreting SMBH in the core of galaxies exhibiting AGN is well described in Ferrarese &Ford (2005b).

The term "quasar" (for quasi-stellar radio source) was introduced in the 1960s when astronomers discovered extraordinarily luminous and distant objects emitting vast amounts of energy. Shortly after this discovery, other similar sources were found, with the same characteristics but no radio emission. QSO (for a quasi-stellar object) was then used to design all these luminous objects appearing starlike, irrespective of their radio emission. Nowadays, throughout this thesis, we use the term "quasar," which is used to designate both radio-loud and radio-quiet sources, which are subsets of AGN (see Sec. 1.1.3).

As research progressed, the classification of AGN expanded to include various types, such as Seyfert galaxies (Seyfert 1943), which were identified by their unique spectral features, and blazars, characterised by their rapid variability and strong jets (Ulrich et al. 1997; Padovani et al. 2016).

Our mastery of multi-wavelength observational techniques significantly evolved in the late 20th and early 21st centuries. This advancement facilitated the study of these objects across the electromagnetic spectrum, leading to a better comprehension of the physical processes involved in AGNs.

This advancement in the study of AGN has led to the conclusion that many of the seemingly different classes can result from a limited number of underlying physical processes, prompting a shift towards a more unified framework for understanding AGN (Padovani et al. 2017). Today, AGN research continues to be a dynamic field, addressing fundamental questions about black hole physics, galaxy evolution, and the role of AGN in the cosmic landscape. However, many questions remain unanswered pertaining the galaxy-AGN co-evolution and the precise origin of SMBH seeds (Volonteri et al. 2021). The identification and follow-up studies of quasars at z>7 could provide invaluable clues in this area.

1.1.2 AGN spectral energy distribution

The spectral energy distribution (hereafter SED) represents the flux density emitted by an astronomical source across a range of wavelengths or frequencies. For AGN, the SED is essential for studying its physical processes in all regions, from the central SMBH to the large-scale jets. The SED helps us understand the energy generation mechanisms, the emitting regions, and how components interact, providing an overall knowledge of AGN physics. A detailed description of AGN emission processes is beyond the scope of this work. We provide below a summary, and we invite the reader to find more detailed presentations in specialised reviews (see, e.g. Sanders et al. 1989; Ferrarese &Ford 2005b; Collinson et al. 2016).

Several spectral features are seen in the SED of AGNs. On one hand, there is a smooth

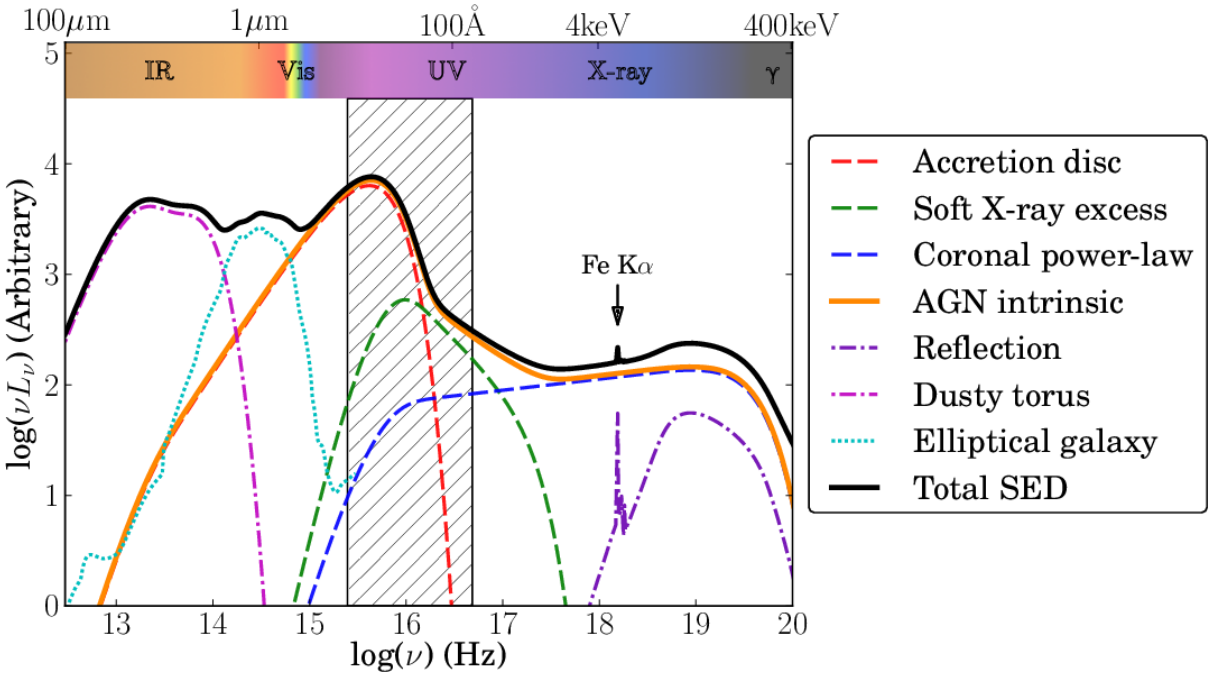


Figure 1.3 – A simplified schematic diagram of an AGN SED, showing the approximate shape and extent of the various components discussed in the text. Dashed lines denote AGN intrinsic emission, dash-dot lines show emission reprocessed by the surrounding material, and the dotted line shows starlight from an elliptical host galaxy. The hatched region highlights the spectral range heavily obscured by absorption in the IGM. We only show an elliptical galaxy here; galaxies with active star formation have stronger UV/IR contributions, Figure and caption from Collinson et al. 2016

continuum resulting from different physical processes like synchrotron radiation, dust re-radiation, and accretion disc emission. On the other hand, emission lines arise from the gas near the AGN, providing information about its composition, motion, and ionisation state. We briefly review these two components below.

1.1.2.1 Continuum

The SED of AGN spans from radio to gamma rays (see Figure 1.3). Synchrotron radiation dominates at the **radio wavelengths**, resulting from the relativistic electrons spiralling in magnetic fields (Sanders et al. 1989). This radio emission is often associated with jets and lobes extending far from the central engine. The **submillimeter and far-infrared** (FIR) regions exhibit thermal radiation from the dust grains heated by the AGN or the surrounding star-forming regions (Sanders et al. 1989). The dust then re-radiates the absorbed energy in the **mid-infrared** (Collin-Souffrin et al. 1986). Substantial contributions from the accretion disc are present in **the optical and ultraviolet (UV)** domains, producing a feature called "big blue bump" (Collinson et al. 2016; Collin-Souffrin 1979; Malkan &Sargent 1982; Czerny &Elvis 1987). In addition to the broad and narrow

emission lines coming from the photoionisation of the gas in the broad-line region (BLR) and narrow-line region (NLR), absorption features from the surrounding gas are also seen in the continuum (Sanders et al. 1989). The inner regions of the AGN contribute to the **X-ray** continuum through Compton scattering of UV photons by a corona of hot electrons (Sanders et al. 1989). This part of the SED also provides information about the geometry and conditions of the central regions and the presence of a reflection component from the inner accretion disk. Finally, in the **gamma-ray** regime, AGN emission often comes from the relativistic jets, including inverse Compton scattering and hadronic interactions (Sanders et al. 1989).

1.1.2.2 Emission lines

Emission lines in AGN provide key insights into the processes in the different regions. The main emission lines in AGN are from the BLR and the NLR, and are primarily present in the UV, optical, and infrared domains.

— **UV**: The main lines come from highly ionised gas in the BLR. Emission lines like Lyα (hydrogen Lyman-alpha), C IV, N V, and Mg II help us understand the ionisation and gas composition around the accretion disk.

— **Optical**: Emission lines come from the ionised gas in BLR and NLR.

Broad Emission Lines, like hydrogen Balmer series (e.g., Hα, Hβ), helium lines, and highly ionised species like C IV. These lines are Doppler-broadened because of the high velocity of the gas close to the SMBH (Sanders et al. 1989).

Narrow Emission Lines. The Forbidden transitions like [O III] and [N II] are the most prominent. These narrow lines come from slower-moving gas farther away from the SMBH. (Sanders et al. 1989).

— **Infrared**

Mid-Infrared. The lines here are associated with polycyclic aromatic hydrocarbons (PAHs) as well as forbidden lines from low-ionisation species like [Ne II] and [Ne III]. These are tracers of the dusty torus and star-forming regions (Sanders et al. 1988, 1989).

Far-Infrared (FIR). Emission lines like [C II] and [O I] are tracers for the gas in the galaxy hosting the AGN. They are an indicator of starburst activity around the AGN (Sanders et al. 1988, 1989).

1.1.3 AGN Classification

AGN classification has evolved since the discovery of the first quasars. Here, we follow the classification proposed by Padovani et al. 2017, who classify the AGNs into classical and less classical classes based on their observational properties, namely the spectral characteristics and jets' presence (or absence). A summary is provided in Table 1.1. Here are the main classes of AGN:

— **Radio-quiet AGN** have a relatively weak emission in the radio wavelengths compared to their emission at other wavelengths, such as optical or X-ray (Padovani et al. 2016; Gaur et al. 2019). The radio-loudness parameter is given by (Gaur et al.

2019):

$$R = \frac{F_{\text{radio}}}{F_{\text{optical}}}, \tag{1.1}$$

where F_{radio} is the radio flux density, measured at a specific frequency, typically 5 GHz, and F_{optical} is the optical flux density, typically measured at 4400 Å. A radio-quiet source has a small R (e.g. R<10).

— **Radio-loud AGN** have a relatively significant emission in the radio wavelengths compared to other wavelengths, such as optical or even X-ray (Padovani et al. 2016).

— **Quasars** are among the most luminous AGNs, and therefore, they can be observed up the highest redshifts. They often outshine their host galaxies as they emit enormous amounts of energy across the spectrum. Quasars are characterised by their strong emission lines. Nowadays, the term QSO (Quasi-Stellar Object) and quasar are used interchangeably. However, in some studies, the distinction can still be made based on radio properties. In that case, a quasar would be radio-loud, whereas a QSO includes radio-loud and radio-quiet AGN.

— **Seyfert galaxies** have an active nucleus and belong to the spiral galaxy type. They are radio-quiet sources. There are two main types of Seyfret galaxies:

1. **Seyfert 1(Sey 1)**: exhibit broad and narrow emission lines in the spectra. These broad emission lines indicate the presence of fast-moving gas close to the SMBH. Seyfert galaxies are typically more luminous than Seyfert 2 galaxies. In conclusion, their main physical property is the gas that causes a Doppler broadening of emission lines due to the high velocities. The broadening happens because the gas moves both toward and away from us, creating a broad spectral line.
2. **Seyfert 2 (Sey 2)**: exhibit narrow emission lines in their spectra. The fact that only narrow emission lines are seen could indicate that the central engine is obscured by dust. This region is close to the central SMBH, where the high-speed gas produces broad emission lines. Due to this obscuration, the broad emission lines in Seyfert 2 galaxies are not visible.

— **QSO 2** have comparable properties to quasars in their total energy output but are radio-quiet. They are obscured, much like Sey 2 galaxies, but with higher power.

— **Blazars** are jetted AGN. Their main characteristic is the strongly variable and polarised non-thermal radiation from radio to γ-rays. They exhibit two broad bumps in their spectral energy distributions: a lower one attributed to synchrotron radiation and a higher one to inverse Compton scattering. Their jets are oriented close to the line of sight from Earth, improving the observation of their radiation due to relativistic effects. The intrinsic properties of blazars can vary significantly, influenced by the peak frequency of their synchrotron emission (Blandford &Rees 1978; Falomo et al. 1993; Urry &Padovani 1995; Padovani et al. 2017). Blazars have two subclasses: flat-spectrum radio quasars (FSRQs) with broad emission lines and BL Lac objects with featureless spectra.

1. **FSRQ (Flat-spectrum radio quasars)** are called "flat spectrum" because their radio emission shows a relatively flat spectral index (α_r). The spectrum is considered flat if this index is small (close to 0). This flatness is due to

synchrotron self-absorption in their relativistic jets. Their spectra have strong, broad emission lines (similar to quasars) (Blandford &Königl 1979; Kellermann et al. 1989). They have more pronounced thermal components in their emission than BL lac.

2. **BL Lacs** exhibit extreme variability and are highly polarised. They have weak or absent emission lines. Their emission is dominated by a relativistic jet pointed close to the line of sight of the observer (Giommi et al. 2012).

— **Type 1 AGN** have broad (FWHM $>$ 1000 km/s) and narrow emission lines (FWHM $<$ 1000 km/s) in their optical spectrum. They allow for a direct view of their central engine with no obscuration by gas or dust, and they have strong X-ray emission from the accretion disk. Examples of Type 1 AGN are Sey1 and quasars

— **Type 2 AGN** only have narrow emission lines. Their central engine is obscured behind a dense torus of dust and gas. Their X-ray emission is absorbed (due to the obscuring gas and dust). Examples of Type 2 AGN are BL Lac and FSRQs (Blazars).

— **Fanaroff-Riley radio sources** is a classification of radio galaxies based on the morphology of their radio emission. These sources are divided into two categories (Fanaroff &Riley 1974), Fanaroff-Riley Type I (FR I) and Fanaroff-Riley Type II (FR II), based on the relative brightness of the radio jets and lobes.

1. **FR I** (Fanaroff-Riley class I radio sources): Their radio emission is the brightest near the core and fades outward. They are low-luminosity sources with weak jets that become turbulent or diffuse as they extend (Fanaroff &Riley 1974).

2. **FR I** (Fanaroff-Riley class II radio sources): Their radio emission is the brightest at the lobes' edges (hot spots). They are characterised by high luminosity and powerful collimated jets, extending to large distances (Fanaroff &Riley 1974).

— **Jetted AGNs** exhibit relativistic jets extending beyond their host galaxy. These jets produce strong emissions in the radio wavelengths. Examples of jetted AGN are Blazars, radio galaxies, and some quasars (Padovani et al. 2016).

— **Non-jetted AGNs** do not exhibit any prominent relativistic jets. Their emission is dominated by the accretion disc and not by jets. Examples of non-jetted AGN are Seyfert galaxies and radio-quiet quasars (Padovani et al. 2016).

1.1.4 Basic phenomenological description

This section provides a basic phenomenological description of the processes taking place in AGNs. More details can be found in various reviews (see, e.g. Ferrarese &Ford 2005a; Padovani et al. 2017, and the references therein). Several key aspects can be used to describe AGN, including theoretical models, evidence of accretion around the central SMBH, the concept of Eddington luminosity, and the duty cycles of activity.

The **"unified model" of AGN** was introduced by Antonucci 1993. According to this model, the differences in AGN types are mainly due to the orientation of the accretion disc and the gas and dust surrounding it with respect to the observer. This model explains how the same physical processes can lead to different observational characteristics depending

Table 1.1 – The AGN zoo: list of the major classes of AGN. The last column describes the main properties. Table from Padovani et al. 2017. The FWHM (Full Widtch Half Maximum) is the width of spectral emission lines at half of their maximum intensity

Class/Acronym	Meaning	Main Properties/Reference
Quasar	Quasi-stellar radio source (originally)	Radio detection no longer required
Sey1	Seyfert 1	FWHM $\geq$ 1,000 km s^{-1}
Sey2	Seyfert 2	FWHM $\leq$ 1,000 km s^{-1}
QSO	Quasi-stellar object	Quasar-like, non-radio source
QSO2	Quasi-stellar object 2	High power Sey2
RQ AGN	Radio-quiet AGN	see Padovani et al. 2016
RL AGN	Radio-loud AGN	see Padovani et al. 2016
Jetted	AGN with strong relativistic jets	see Padovani et al. 2016
Non-jetted	AGN without strong relativistic jets	see Padovani et al. 2016
Type 1	Sey1 and quasars	
Type 2	Sey2 and QSO2	
FR I	Fanaroff-Riley class I radio source (Fanaroff &Riley 1974)	radio core-brightened see Fanaroff &Riley 1974
FR II	Fanaroff-Riley class II radio source	radio edge-brightened (Fanaroff &Riley 1974)
BL Lac	BL Lacertae object	see Giommi et al. 2012
FSRQ	Flat-spectrum radio quasar	RL AGN, $\alpha_r \leq 0.5$
Blazar	BL Lac and quasar	BL Lacs and FSRQs

on the line of sight to the central engine (Urry &Padovani 1995). Before the unified model, AGNs were largely classified based on observational differences without a unifying framework. This model was later revisited by Netzer 2015, who proposed to include other parameters in the model besides orientation and luminosity. The revised model acknowledges that the unified model cannot explain significant differences among AGN and that the central obscurer (torus) is more complex and clumpy than previously thought. This revised model introduces the evolution of black holes and their host galaxies as an additional complexity, particularly in merging systems, which may not conform to the original unification scheme.

There was a long path between the observation of the first quasars and the **evidence for an accreting SMBH** in the core of galaxies exhibiting AGN (Ferrarese &Ford 2005b). This fact is well established today and supported by numerous observations, such as the detection of broadened emission lines in the optical spectrum, which indicate high-velocity gas orbiting the SMBH. X-ray emissions and variability observations at different wavelengths (constraining the size of the emitting region) are additional proofs of the accretion processes. The infall of matter into the SMBH generates a tremendous amount of gravitational energy converted into radiation, and this is the only known process which is able to provide such an amount of energy (Ferrarese &Ford 2005a; Kormendy &Ho 2013).

Regarding the emission of an AGN, it is essential to introduce the concept of **Eddington Luminosity**, often used in the literature because it provides a measure of the accretion rate. The Eddington Luminosity (L_E) is the luminosity at which radiation pressure on free electrons balances the force of gravity. This luminosity limit is significant for AGN as it sets the scale for the accretion rate and the resulting luminosity. L_E is independent of

the distance but depends on the mass of the SMBH (M) as follows:

$$L_E = \frac{4\pi G M m_p c}{\sigma_T} \sim 1.3 \times 10^{46} \frac{M}{10^8 M_\odot} erg/s \tag{1.2}$$

where m_p is the proton mass and σ_T is the Thomson cross-section. Above the Eddington luminosity, the source cannot maintain steady spherical accretion. The relationship between the mass of the SMBH and its Eddington luminosity helps to explain the observed range of AGN luminosities and their evolution over cosmic time (Shankar et al. 2013; Jones et al. 2016). Today, direct observational evidence exists for a relationship between accretion rate and the SMBH mass (Rodighiero et al. 2015).

The mass (and the corresponding luminosity) of a SMBH accreting at the Eddington rate, with constant (dM/dt)/M, will increase exponentially, with e-folding time t_S:

$$t_S = \frac{\sigma_T c}{4\pi G M m_p} \sim 4 \times 10^8 \epsilon yr \tag{1.3}$$

where ϵ is the efficiency of conversion of mass into energy. This quantity depends on the spin of the black hole and varies typically between 6% when a black hole is not spinning and 42% if the black hole reaches maximum spinning. t_S is often called the "Salpeter time" and represents the time needed for a black hole radiating at the Eddington luminosity a time to dissipate its entire rest mass.

To estimate the total energy output of an AGN over cosmic time, a determination of its **duty cycle** is also needed. The duty cycle of AGN is the ratio of time intervals during which the AGN is active and shut down. Here, "active" means a phase during which the AGN emits significant radiation due to accretion into the SMBH. The value of the duty cycle for a certain type of AGN can be deduced from the time evolution of its global Luminosity Function and SMBH mass function (see Sect. 1.1.6), assuming a value for ϵ in the equation of t_S (see e.g. Delvecchio et al. 2020). Direct observational measurements for individual sources are challenging (see, e.g. Shulevski et al. 2015). Understanding the history of AGN accretion and its connection with the galaxy and large-scale structure assembly is of great importance. Needless to say, AGN accretion history has implications for galaxy evolution, as the feedback from AGN influences star formation and the growth of the host galaxy (Hasinger 2008).

1.1.5 Selecting AGN

The goal of this thesis is to propose and test a Bayesian method to select high-z quasars in wide-field optical to near-IR cosmological surveys. Before describing this method, it is important to remember the main techniques used to identify AGNs since the discovery of the first quasars. AGNs can be selected using different strategies based on their observational characteristics across the electromagnetic spectrum. A complete description can be found in Padovani et al. (2017). Here are the main approaches used for AGN selection:

— **Radio-selected AGN**. Radio emissions inform us of relativistic jets associated with jetted AGN. Radio selection is particularly effective for identifying radio-loud

AGN, including radio galaxies and blazars. This is important as they might not be detected in the optical due to obscuration by dust (Kellermann et al. 1989).

— **Infrared-selected AGN**. Infrared selection can help identify AGN that is heavily obscured in other wavelengths, providing a complete view of the AGN population as the dust re-emits in the far infrared the energy absorbed in the UV and visible bands (Assef et al. 2013).

— **Optically-selected AGN**. This selection method focuses on the main spectral features available in this domain, namely broad emission lines and breaks in the continuum, such as the Lyman break for high-z quasars selected in the optical bands. Different techniques have been proposed, such as the use of colour cuts, photometric redshifts, narrow-band filters, or blind spectroscopic surveys. The optical selection of AGN is interesting for several reasons, in particular, to search for high-z AGNs with important spectral features in the visible and NIR domains. Large surveys significantly enhance the effectiveness in finding these rare objects (2dF Colless et al. 2001, SDSS Margon 1999, VVDS Le Fèvre et al. 2005, *Euclid* Euclid Collaboration: Scaramella et al. 2022 and LSST Ivezić et al. 2019b).

— **X-ray selected AGN**. X-ray luminosity is a crucial activity indicator in AGN because material accretion produces high-energy X-rays. This method is also used to identify obscured AGN (Brandt &Hasinger 2005).

— **Gamma-ray selected AGN**. Gamma-ray emissions are associated with the most energetic processes in AGN, like blazars. They are an indicator of high-energy particles near the SMBH (Abdo et al. 2010).

— **Variability-selected AGN**. The variability in the flux emitted by an AGN over time can be used to identify the source. Variability is connected to the accretion around the centre of the SMBH. By observing the brightness of potential AGN candidates over time, significant changes in their luminosity can be observed, strongly indicating AGN activity (Ulrich et al. 1997).

Different approaches can be combined to increase the reliability of the selection. For instance, variability or X-ray emission can be helpful when selecting quasars in the optical wavelengths.

1.1.6 Evolution of the global properties of AGN across cosmic times

The evolution in the global properties of AGN captures the evolution of SMBH formation across cosmic times. In the following, we give some indications of the evolution of the AGN Luminosity Function and the mass function of SMBH. Since this thesis seeks to improve our knowledge of the epoch of reionisation, it is also important to recall the role played by quasars in our understanding of this epoch, particularly the observation of the Gunn-Peterson trough. All these aspects are briefly presented in this section.

1.1.6.1 Abundance and Luminosity function of AGNs

Quantifying the abundance of AGN and its evolution through cosmic time is fundamental for studying the growth and evolution of the central SMBH and their host galaxies over

cosmic time. As discussed in the review by Padovani et al. (2017), each selection method described above has weaknesses and strengths related to the specific characteristics of AGNs they are sensitive to. The relative efficiency is given by the ability to identify a large number of AGNs with a minimum cost in terms of exposure times. In the case of a selection based on rest-frame optical and NIR data, as in this thesis, the identification will be sensitive to the emission by the accretion disk.

The Luminosity Function (hereafter LF) gives the number density of AGN (per unit comoving volume and luminosity/magnitude bin) as a function of luminosity (or absolute magnitude). Different studies have addressed the LF of quasars and its evolution in redshift based on rest-frame optical and NIR data coming from wide-field cosmological surveys, like the 2dF QSO Redshift Survey at 0.3≤z≤2.3 (Boyle et al. 2000), or the Sloan QSO survey at 3*leq*z≤5 (Fan et al. 2001d). Figure 1.4 shows the evolution with redshift of the LF for quasars in the 2dF QSO Redshift Survey. Type 1 AGNs have also been extensively selected and studied in the X-rays, based on data from ROSAT, XMM-Newton and Chandra surveys (see, e.g. Hasinger et al. 2005). These results show a luminosity-dependent density evolution with redshift, as illustrated in Fig. 1.5, for both optical and X-ray selected samples. Simple models have been used to understand these trends, like the Pure-Luminosity evolution model (PLE), or Luminosity-Dependent Density evolution model (LDDE) (see, e.g. Boyle et al. 1988; Hasinger et al. 2005).

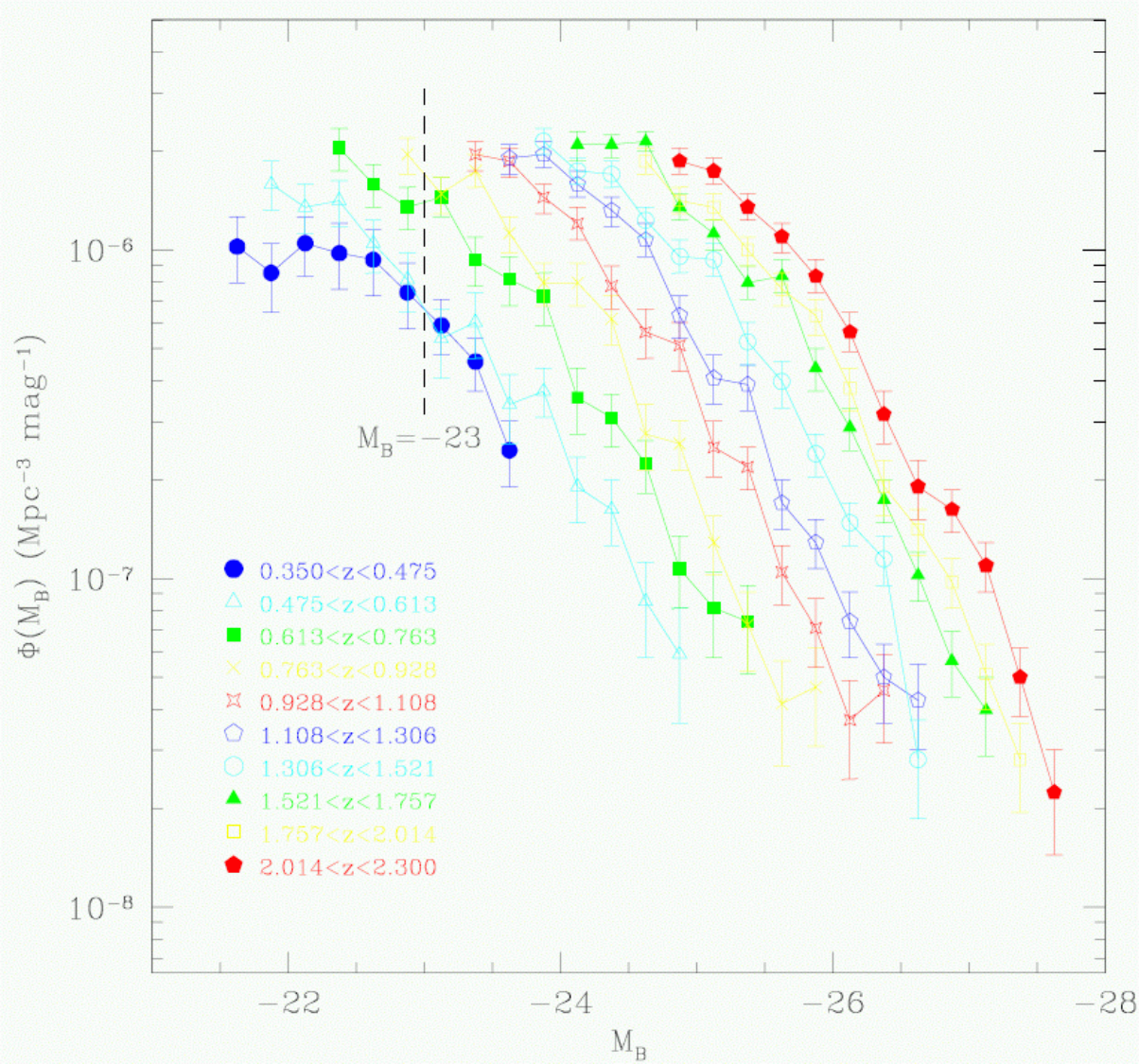


Figure 1.4 – Evolution with redshift of the LF for quasars in the 2dF QSO Redshift Survey. Credits: Boyle et al. (2000).

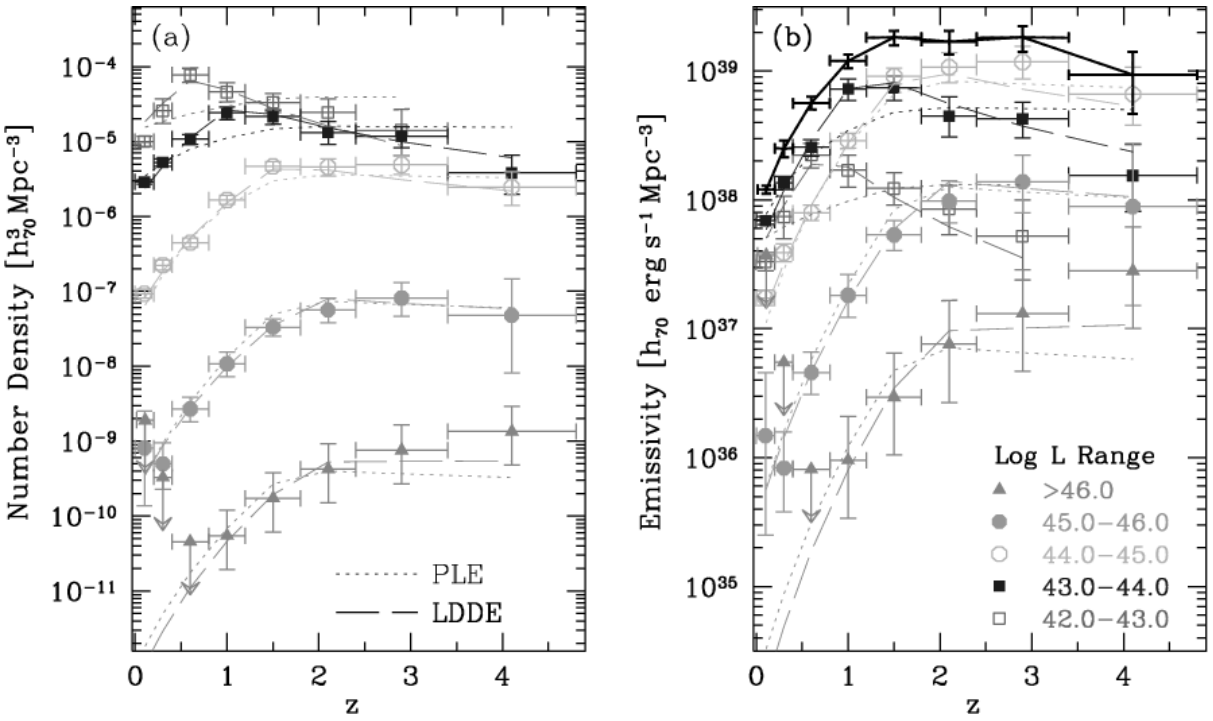


Figure 1.5 – Left: The space density of AGNs as a function of redshift in different luminosity classes and the sum over all luminosities with log $L_x \geq 42$. Densities expected from two evolutionary models (Pure-Luminosity evolution model-PLE-, and luminosity-dependent density evolution model-LDDE) are overplotted with solid lines. Right: The same as Left, except that the soft X-ray emissivities are plotted instead of number densities. The uppermost curve (black) shows the sum of emissivities in all luminosity classes plotted. Credits: Hasinger et al. (2005).

Results from different studies, both in the optical and X-rays, show that AGN activity has gone through significant evolution from $z = 0$ to $z = 2$. It peaks at redshift ∼2 (Schmidt &Green 1983; Boyle et al. 1988, 2000; Hasinger et al. 2005; Fanidakis et al. 2012), corresponding to a period of the highest accretion and star formation in the universe. At lower redshifts (z<2), AGN numbers decline (Warren et al. 1994; Fan et al. 2001a; Wolf et al. 2003). This implies the transition of galaxies from active to more quiescent phases. Observations also demonstrate the diversity of AGN as a function of luminosity, from the most luminous quasars to lowest-luminosity Seyfert galaxies (Boyle et al. 2000). These studies are extremely important for understanding the co-evolution of black holes and galaxies, the physics of accretion processes, and AGN's role in shaping large-scale cosmic structures.

1.1.6.2 Mass function of SMBH

Since the first review on the mass of SMBHs published by Kormendy &Richstone (1995), the progress has been rapid in the detection of SMBHs using a variety of methods in the local universe (see Ferrarese &Ford 2005a, for a review). Measuring SMBH masses in

high-z quasars is more challenging. It requires some assumptions, such as the fact that the (bolometric) luminosity comes only from gas accretion onto the central SMBH, with a fixed conversion efficiency (see Sect. 1.1.4). Additionally, it is assumed that SMBHs are originally in place with a minimum mass and cannot be formed or destroyed in subsequent phases. Under these assumptions, the total mass density or mass function can be derived from quasars' luminosity function (LF). As pointed out by Ferrarese &Ford (2005a), this derivation is sensitive to the above assumptions and, to a lesser extent, the LF used in the analysis.

For instance, the magnitude limits of the 2dF survey ($0.3 \leq z \leq 2.3$; Boyle et al. 2000) and the Sloan QSO survey ($3 \leq z \leq 5$; Fan et al. 2001d) correspond to Eddington limits on the SMBH masses of $4.5 \times 10^7 M_\odot$ and $7.3 \times 10^8 M_\odot$, respectively. This implies that, at high redshifts, quasar surveys probe only the most massive supermassive black holes, in contrast with the samples built in the local universe.

A strong connection has emerged between SMBHs and their host galaxies, in the sense that the mass of the SMBH is directly correlated to the mass of the dark matter (DM) halo in which it resides (see, e.g., Ferrarese &Ford 2005a, for a review).

The mass function of SMBHs shows a trend of "downsizing," where the abundance of less massive black holes declines rapidly with increasing redshift (Li et al. 2012), indicating that more massive black holes were more predominant in the early universe Magorrian et al. 1998. The space density of SMBHs remains relatively stable since redshift z~2, suggesting that while the most massive black holes continue to grow, the mechanisms driving their growth evolve, with significant mass buildup occurring primarily through hot-halo accretion and mergers at lower redshifts McLure &Dunlop 2002. Additionally, the mass density of black holes increases significantly from z=6.2 to z=2 due to intense accretion during the starburst phase (Fanidakis et al. 2012) and the strong correlation between SMBH mass and the host galaxy mass (Ferrarese &Merritt 2000). However, the increase becomes modest from z=2 to z=0, where it only doubles as the growth of the most massive black holes becomes less efficient (Marconi &Hunt 2003). The mass function reflects the dominance of different accretion modes at various redshifts, with the starburst mode being more significant at high redshifts, leading to varying slopes in the mass function for high-mass versus low-mass black holes (Häring &Rix 2004). Furthermore, the models of the black hole mass function (BHMF) indicating downsizing do not provide a unique interpretation of the observational data, highlighting the complexity of factors influencing the observed properties of SMBHs, as shown by (Ferrarese &Merritt 2000; Gebhardt et al. 2000). These observational facts feed the cosmological simulations, such as HORIZON-AGN and NEWHORIZON, which can trace the SMBH/galaxies co-evolution by following the relevant scales in mass for the MBH population (see, e.g. Volonteri et al. 2020).

1.1.6.3 AGNs as tracers of the IGM and reionisation

The Lyman forest refers to a series of absorption lines seen in the spectra of distant galaxies and AGN caused by the intervening neutral hydrogen gas H I along the line of sight (Rauch 1998). These lines are imprints of the Ly_α and Ly_β transitions of hydrogen at the wavelength of 1215,17 Å and 1026 Å, respectively, in the rest-frame of the absorbers,

which is highly sensitive to the presence of H I in the intergalactic medium (IGM) (Wise 2019). Ly_α and Ly_β photons have cross sections large enough for exciting a neutral hydrogen atom at its ground state n = 1 to the n = 2 and n = 3 states, respectively. When the atom falls back to its fundamental state, the photon is randomly scattered and not re-emitted in the same direction, which causes an absorption at the corresponding wavelength. Due to redshift, light emitted at the wavelengths of the Lymann series from distant sources passes through multiple clouds of neutral hydrogen at different redshifts (Meiksin 2009), leading to the characteristic "forest" of absorption lines at wavelengths shorter than the source's redshifted Ly_α and Ly_β lines (Gunn &Peterson 1965a; Scheuer 1965; Shklovskii 1965; Weymann et al. 1981). Figure 1.6 illustrates this effect.

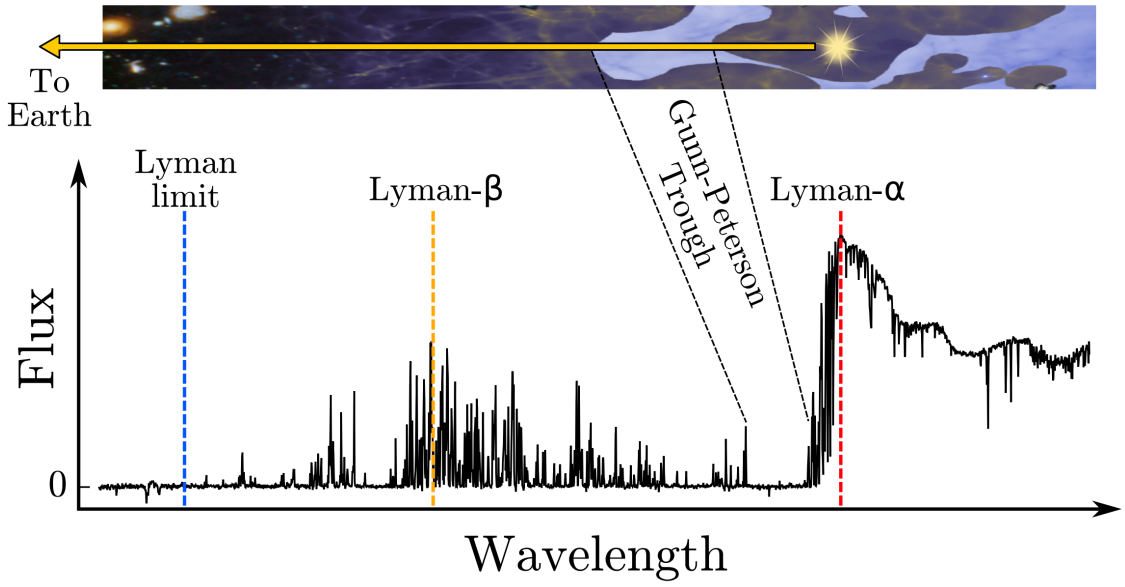


Figure 1.6 – light from distant quasars, powered by supermassive black holes, can probe the ionisation and thermal state of the intergalactic medium. Overdense clumps of intergalactic gas absorb some fraction of light from the intrinsic spectrum (bottom) when the photons ionise any neutral hydrogen. Only lines in the Lyman series, down to the Lyman limit (912 Å), with Ly_α (1215 Å) and Ly_β (1026 Å) being the strongest. Absorption lines from clouds at various redshifts create a Ly_α forest. When these lines become numerous enough, a Gunn-Peterson trough is created that indicates an elevated neutral hydrogen fraction and the end of the Epoch of Reionisation. Image Credits: Wise (2019).

The number and intensity of the Lyman_α absorption lines show the column density and distribution of neutral hydrogen in the IGM. Low column-density systems correspond to diffuse gas in the cosmic web, while higher column-density systems (Lyman limit systems or damped Lyman-alpha systems) trace denser structures (Schaye et al. 2000; Fechner &Reimers 2007). The thermal broadening of absorption lines in the Lyman_α forest provides information on the temperature of the IGM. The relationship between the

IGM's temperature and density helps constrain its equation of state, often parametrised as $T \propto \rho^{\gamma-1}$, where γ describes the thermal state of the IGM (Hui &Gnedin 1997; Schaye et al. 2000).

The distribution of absorption features in the Lyman$_\alpha$ forest helps estimate the fraction of baryonic matter in the IGM, which constitutes the majority of the universe's ordinary (baryonic) matter (Fukugita et al. 1998; Schaye et al. 2010). Last but not least, the associated absorption features from metals (e.g., C IV, Si IV) in the Lyman$_\alpha$ forest offer crucial information on the enrichment of the IGM with heavy elements, tracing feedback processes from galaxies and AGN (Simpson et al. 2016; Kakiichi et al. 2018; Frank et al. 2010).

The intricate connection between the Lymann$_\alpha$ forest and the ionising background radiation—governing the ionisation state of the IGM—sets the stage for exploring the role of AGN in general and SMBHs in particular, and their contributions to reionisation (McDonald et al. 2000; Schaye et al. 2000; Dayal et al. 2013). Spectroscopic observations of luminous quasars at z∼6 have set the first constraints on the epoch of reionisation (see, e.g. Fan et al. 2006b, and the references therein). Indeed, the study of the Gunn-Peterson absorption indicates a rapid increase in the neutral fraction of the IGM, from x(HI) $<$ 10^{-4} at z $\leq$ 5.5, to x(HI) $>$ 10^{-3} at z∼6. The Gunn-Peterson absorption saturates above this threshold. This observation is considered the first evidence for a neutral IGM above z∼6.

1.1.7 Quasars and SMBH at the epoch of the reionisation

1.1.7.1 The epoch of the reionisation

Cosmic reionisation presents a pivotal era in the evolution of the early universe. It is the transition that the universe underwent from the "cosmic dark ages" to a transparent universe filled with ionised gas (Gnedin &Madau 2022). This process started approximately 150 million years after the Big Bang and ended approximately one billion years later (z∼6). Following the formation of the first stars and galaxies, UV light emitted from these objects triggered the process of ionising the surrounding hydrogen gas. For more extensive reviews presenting the early structure formation and reionisation process, see, e.g. Barkana &Loeb (2001b); Fan et al. (2006b).

The UV light from the first luminous sources formed in the early universe (stars, galaxies, and AGN) is likely to be the origin of the radiation responsible for reionisation. The whole process was complex and took hundreds of millions of years to reach completion (Furlanetto et al. 2006). Several phases can be distinguished. A first phase, relatively slow, occurred when each UV source was isolated within its own Strömgren sphere, followed by a second phase, which was relatively fast, as these spheres eventually grew and merged. The overlap phase is a relatively local phenomenon, with significant differences along lines of sight.

Observations performed before the arrival of the James Webb Space Telescope (JWST) set strong constraints on the epoch of reionisation. They led to the idea that ionising photons responsible arose from high-redshift, low-mass, star-forming galaxies (see, e.g., Robertson et al. 2015, and references therein). Among the key observations used to derive

these constraints are spectroscopic studies of $z \sim 6$–8 QSOs (see Sect. 1.1.6), the UV luminosity function of star-forming galaxies and its evolution with redshift (together with the star-formation rate density), and the integrated optical depth of Thomson scattering to the cosmic microwave background (Robertson et al. 2010; Planck Collaboration et al. 2020).

Simulations such as the SPHINX (Rosdahl et al. 2018) or THESAN project (Kannan et al. 2022) provide new insights into this era. These numerical simulations are among the most detailed, capturing how light from early galaxies interacted with gas, driving reionisation. Cosmological simulations help us understand which types of galaxies/sources were responsible for reionisation (Davé et al. 2019; Hummels et al. 2013; Pallottini et al. 2019), and the precise interpretation of current observational results in this framework.

1.1.7.2 Quasars, SMBH and reionisation

The main contributors to cosmic reionisation are supposed to be high-redshift, low-mass star-forming galaxies, as discussed in Sect. 1.1.7.1 The contribution of AGNs has been considered negligible based on the evolving far-UV luminosity density produced by AGN at high redshift, derived from the quasars UV LF (see Sect. 1.1.6) and the fact that it declines rapidly from z∼4 to z∼6. This trend has been already shown in the previous section. It is further supported by the recent work of Matsuoka et al. (2023), who derived the LF for z∼7 quasars and confirmed that the overall LF shape remains similar from z= 4 to 7. At the same time, the amplitude decreases dramatically at z≥ 6. This LF is later used in Chapter 3 to model the expected abundance of high-z quasars. This is also the most up-to-date LF for quasars at z≥7, based on a relatively small number of spectroscopically confirmed sources at 6.55≤z≤7.15, covering the -28<M_{UV} <-23 domain in luminosity. As discussed later in Chapter 4, the number of spectroscopically-confirmed quasars at z≥7 is dramatically small, and it is one of the motivations of this thesis work.

Based on the estimated ionising flux from the LFs mentioned above, these results indicate that quasars are not the dominant contributors to cosmic reionisation at z∼6-7. Their contribution to the ionising flux is less than 1% of the critical rate to keep the IGM ionised at z=7. The reader should remember that only the most massive SMBHs have been identified in the surveys used to derivate the previous LFs (typically above a few $\times 10^8$ M$\odot$). The ongoing observations with JWST can probe down to masses around $M_\odot \sim 10^6 - 10^7\, M_\odot$ at redshifts typically larger than 4, up to z=6, hinting at the existence of even fainter quasars (Larson et al. 2023; Maiolino et al. 2024d,a; Matthee et al. 2024). In recent JWST observations, AGN signatures have been found up to (at least) redshifts z∼11 (Maiolino et al. 2024d; Oesch et al. 2016), suggesting the considerable presence of accreting SMBHs in the early universe. However, cosmological simulations struggle to reproduce the observed SMBH densities, highlighting the necessity for a more refined approach to black hole seeding mechanisms (Ni et al. 2022). By synthesising observations and theoretical models, the history of SMBH formation is gradually interpreting the interplay between black holes and their hosting environments over time. The continued evolution of these models is important for understanding the origins and growth of the universe's most massive objects (Li et al. 2024).

These recent findings reveal an emerging population of AGNs at 4≤z≤13 have led

Figure 4. from
Piero Madau et al 2024 Astrophys. J. 971 doi:10.3847/1538-4357/ad5ce8
https://dx.doi.org/10.3847/1538-4357/ad5ce8
© 2024. The Author(s). Published by the American Astronomical Society.

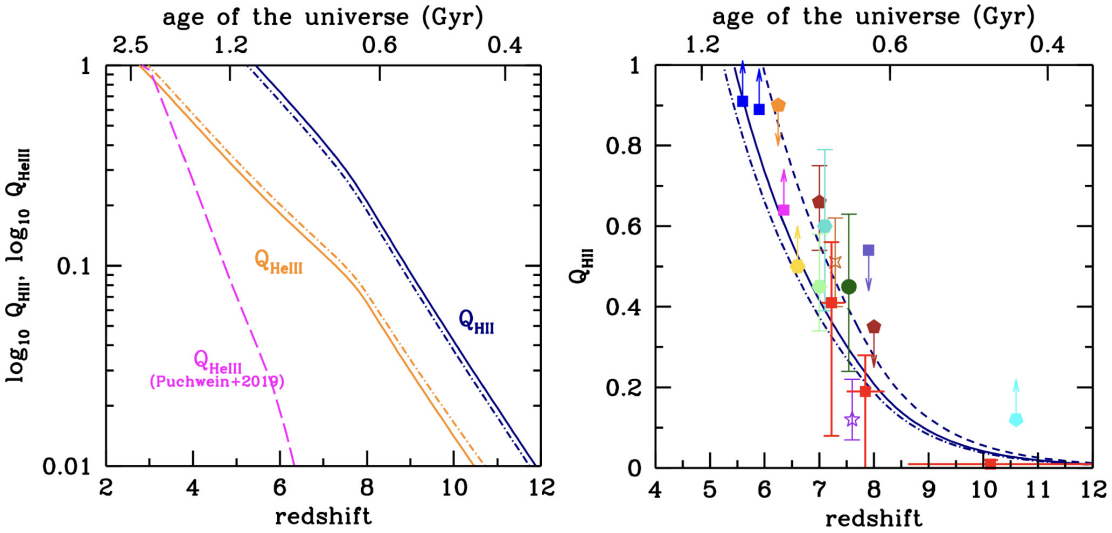


Figure 1.7 – Left panel: Fiducial reionisation histories for the AGN-dominated Models I (solid lines) and II (dotted–dashed lines discussed in (Madau et al. 2024). The evolving H II and He III ionised volume fractions are depicted in blue and orange colours, respectively. Hydrogen in the IGM is fully reionised when Q HII = 1, while helium is double reionised when Q HeIII = 1. The dashed curve shows a model of He II reionisation history driven by rare, luminous quasars (the "fiducial model" of Puchwein et al. 2019). Right panel: constraints on the ionisation state of the IGM from various probes and techniques. The solid, dotted–dashed, and dashed curves denote the ionised volume filling factor Q HII (linear scale) for our Model I, Model II, and "modified Model I," respectively (see the main text for details). The data points are from McGreer et al. 2015; (blue squares), Schenker et al. 2014 (firebrick pentagons),Greig et al. 2017 (turquoise pentagon),Schroeder et al. 2013 (orange pentagon), Ouchi et al. 2010 (gold hexagon), Gallerani et al. (2008; magenta square), Bañados et al. 2018 (green dot),Hoag et al. 2019 (purple star), Greig et al. (2022; brown star), Mason et al. (2019; blue square),Whitler et al. 2020 (light-green hexagon), Nakane et al. 2024 (red squares), and Bruton et al. 2023 (cyan pentagon). Image and caption credits: Madau et al. (2024).

Madau et al. 2024 to wonder whether we should not return to the idea that AGNs could have a significant contribution to reionisation once the full population (including the smallest SMBHs) is taken into consideration. In the past, (Madau &Haardt 2015) already discussed this idea of quasars being primarily responsible for cosmic reionisation, given their strong radiation field resulting from accretion onto SMBHs (Miralda-Escudé et al. 2000; Madau &Haardt 2015). Madau et al. 2024 reassess a scenario where AGNs are the main/sole contributors of the reionisation process, based on the fraction of type 1 AGN among high-z galaxies and a high mean escape fraction of Lyman continuum photons. In this reionisation scenario Madau &Haardt 2015 obtain results fully consistent with current observations, as illustrated in Fig. 1.7, in particular a hydrogen reionisation of ∼99% at z=5.5. Their results further indicate that delays in helium reionisation—brought about by factors such as internal absorption and the steepness of ionising spectra—are consistent with the observed characteristics of the He II Lyman-alpha forest (Madau &Haardt 2015; Worseck et al. 2016). This analysis, together with new observational evidence (e.g. Maiolino et al. 2024b), help us reshape our understanding of the IGM and set the stage for future studies aimed at untangling the complexities of cosmic evolution and the role of AGNs in shaping the early universe.

The emergence of a significant population of active SMBHs at high redshift (Maiolino et al. 2024e,f,c; Scholtz et al. 2024) may play a crucial role in how we view cosmic reionisation in the future. Recent studies by Grazian et al. 2023, 2024; Greene et al. 2024 suggest that these AGNs exhibit unique characteristics, such as broad Balmer lines and a compact morphology, indicating their unobscured nature typical of type 1 AGNs. Notably, the "Little Red Dots" have been identified as probable AGNs at high redshift, further showing their prevalence in the early Universe (Goulding et al. 2023). Additionally, there is an ongoing debate regarding their ionising emissivity and escape fraction of ionising photons, which are essential for understanding their contribution to the reionisation process (Grazian et al. 2018; Matthee et al. 2024). The properties of this "Little Red Dots" population discovered by JWST have started to be addressed by cosmological simulations (Volonteri et al. 2024).

As a final remark, the quasars I intend to identify in this thesis using Bayesian model comparison applied to wide photometric surveys are bright and rare sources corresponding to the most massive SMBHs. Despite their high intrinsic luminosity, high-redshift quasars appear faint at $z > 7$, and their photometric selection on wide cosmological surveys is particularly challenging. Indeed, their red colours, often used to identify them with the appropriate colour cuts or photometric redshifts, are subject to severe contamination by brown dwarfs and intermediate-redshift galaxies. With the arrival of massive cosmological surveys, such as *Euclid* (Laureijs et al. 2011; Euclid Collaboration et al. 2024a), the Nancy Grace Roman Space Telescope (Akeson et al. 2019) and the Large Synoptic Survey Telescope (LSST) (Ivezić et al. 2019a), it becomes possible to identify significant samples of quasars at high redshifts, which will significantly improve our understanding of galaxy and SMBH formation mechanisms. However, for this, building efficient and optimised methods and tools is essential. This is precisely the purpose of my thesis.

1.2 The selection of astrophysical sources using a Bayesian approach

1.2.1 Historical introduction

Astronomy and statistics have been deeply connected from early times, with astronomers aspiring to quantify and describe celestial phenomena and infer underlying physical principles from observations (Feigelson et al. 2009). Statistical methodologies upgraded from data collection to mathematical analysis. In the 19th century, probabilistic approaches gained prominence through the formalism of the least squares method for refining orbital calculations (Feigelson et al. 2009). This method gained popularity in celestial mechanics, and Gauss further advanced statistical theory by devising techniques for handling observational errors and introducing the Gaussian, or "normal," distribution, which was widely referred to as the "astronomical error function" (Gauss &Stewart 1995; Hughes 2023). Astronomers continued to employ the method of least squares as a statistical tool. One example was Hubble's Law, which describes the universe's expansion by establishing a linear equation between the recessional velocity of external galaxies and their proper distance (Hubble 1929). That was until the Bayes' Theorem by Thomas Bayes (Bayes &Price 1763) and the Maximum Likelihood Estimation by Ronald Fisher (Fisher 1922) were employed in astronomy throughout the early 20th century. By the end of the century, maximum likelihood estimation had already had significance in various fields of astronomy. An example closely related to the topic of this thesis: it helped calculate the luminosity function of extragalactic sources (Efstathiou et al. 1988).

1.2.2 Classical selection methods for high redshift quasars

The different selection methods based on the SED of quasars have been described in Sect. 1.1.5. Here, we focus on selection in the visible-NIR domains, following the needs of this thesis work. Quasar selection typically involves several key steps, including photometric and spectroscopic techniques:

1. *Photometric Selection:* This initial step involves identifying potential quasar candidates based primarily on their brightness. Surveys measure the flux of objects across different wavelengths (optical and infrared) to find sources that exhibit excess brightness or particular colours characteristic of quasars. Criteria like colour indices (using photometric filters) help distinguish quasars from other types of celestial objects. This is a very effective method that allowed for the discovery of numerous high-redshift quasars (see details in Belladitta et al. 2020, Bañados et al. 2018 and Bañados et al. 2016).
2. *Spectroscopic Confirmation:* Once candidates are selected photometrically, follow-up observations are conducted to obtain their spectra. This confirmation involved detecting broad emission lines, such as Lyα, Mg II and/or C IV. If the spectra have a good enough Signal-to-Noise (SNR) and resolution, and the lines are correctly identified, the sources are successfully confirmed. Spectroscopic confirmation allows us to access the redshift and rest-frame flux emitted (or absorbed) by the different

components of the quasar, hence having a more detailed idea of the source in question.

3. *Multi-wavelength Data:* Multi-wavelength observations are employed to improve the accuracy of quasar selection and guarantee the proper estimation of their redshifts. This can include data from X-ray, radio, or infrared wavelengths, providing a more comprehensive view of the object's characteristics and helping reduce contamination from other red sources (in the case of high-z quasars).

1.2.3 Recent developments

In the last decade, as telescopes have improved in terms of their capability to collect photons (both in the area covered and in depth) and in spatial and spectral resolution, the ambition of astronomers to find faint and distant sources has increased. Many methods have been set to select high redshift quasars. These methods are based on photometry, spectroscopy, or combining both techniques.

1.2.3.1 Methods based on statistics

Statistical source selection is fundamental to identifying and classifying objects or phenomena based on observational and photometric data in astronomical sources. Traditional methods of source selection often rely on deterministic criteria, such as cut-based thresholds on measurable attributes (e.g., brightness or colour), which can lead to significant biases and misclassification, particularly in the presence of noise or incomplete data (Hewett et al. 1995; Goto 2006). Consequently, more robust statistical approaches that account for uncertainties and incorporate prior knowledge into the classification process were needed (Sivia &Skilling 2006). In particular, the Bayesian approach provides a robust statistical framework for selecting and classifying astronomical sources by integrating prior knowledge with probabilistic modelling of uncertainties. Unlike deterministic classification methods, Bayesian inference estimates the probability that an astronomical source belongs to a particular category (or class of sources) (Jaynes &Bretthorst 2003; Sivia &Skilling 2006). This statistical perspective is advantageous for analysing our inherently noisy, incomplete, and high-dimensional data in astronomical surveys (Kaiser et al. 2002; Emerson et al. 2004). Previous studies such as Fan et al. 2001b and Mortlock et al. 2011 have established the effectiveness of the Bayesian approach in selecting high-redshift quasars amidst various contaminating sources (Hughes 2023).

1.2.3.2 The XDHZQSO method

In recent years, numerous statistical methods have been developed for selecting high-redshift quasars, employing machine learning and probabilistic modelling techniques. One notable approach is the XDHZQSO method, which utilises a probabilistic Bayesian framework based on density estimation in flux space to compute the probability of an object being a high-redshift quasar. The XDHZQSO method performs the following steps (Nanni et al. 2022):

1. **Density Estimation**: The method employs extreme deconvolution (XD) techniques (Bovy et al. 2011b,a) to model the distribution of quasars in a multidimensional space, accounting for heteroscedastic noise. The quasar's probability density function (PDF) is a Gaussian mixture.
2. **Bayesian Framework**: Using a Bayesian formalism, the probability $P(QSO|O)$ of a given object O being a quasar is computed by integrating the quasar density over the observed fluxes. This requires knowledge of the prior distribution of quasars and contaminating sources, such as M-type dwarfs and early-type galaxies (Mortlock et al. 2012; Reed et al. 2017).
3. **Classification**: Each object is assigned a probability of being classified as a quasar. The classification step determines whether O belongs to the class' high-z quasar' or 'contaminant'. The assignment is based on evaluating the likelihood ratios derived from the density estimations.

When **applied to surveys**, the XDHZQSO technique has been shown to achieve completeness rates of $\geq 56\%$ and efficiency rates of $\geq 5\%$ for selecting quasars in the redshift range $6 < z < 8$ (see Nanni et al. 2022). Given the low expected surface density of $z \geq 7$ quasars compared to more abundant contaminants (Wang et al. 2019), this sophisticated approach is paramount for large-scale astronomical surveys.

1.2.3.3 Random Forests

Random Forests is a supervised machine-learning method that operates by constructing multiple decision trees during the training phase and outputting the mode of their classifications for regression tasks. In the context of selecting high-redshift quasars, Random Forests handle the complex and high-dimensional photometric data from telescopes like Pan-STARRS and WISE (and soon *Euclid*). This technique allows for identifying quasars even in overlapping colour spaces with other celestial objects, such as brown dwarfs and low redshift galaxies, which complicate their selection. By creating a representative training set of known quasars and contaminants, the algorithm can learn to distinguish between them, ultimately leading to high completeness and efficiency in quasar selection, even in challenging regions of the parameter space (Wenzl et al. 2021).

1.2.3.4 Selection from spectra and SED

The method for selecting quasars using the Blue Grism of the *Euclid* NISP spectrophotometer involves using photometric and spectroscopic data. The Blue Grism covers from 0.92 to 1.25 μm (see Chapter 2); this spectral coverage is useful to capture key spectral features associated with quasars, such as the Lyα emission line. The redshift range in which we can observe the Lyα line in the blue grism falls between $z \approx 6.57$ and $z \approx 9.28$. For *Euclid* members, the detailed process is explained in the Blue grism Science Case presented in `https://euclid.roe.ac.uk/dmsf/files/20929/view` Here is an overview of the two-step process to select high redshift quasars:

1. Using the *Euclid* images, a photometric pre-selection is applied. This pre-selection starts with a photometric survey where objects in the sky are measured for fluxes

and magnitudes at various wavelengths. Quasars exhibit distinct colours in their photometric data due to their strong emission lines and Lyman Break (see Sect.1.1.6) compared to objects like stars or brown dwarfs. This pre-selection step is a colour-based selection to provide a preliminary list of potential quasar candidates.

2. Spectroscopic follow-up with Blue Grism: Once candidates are identified photometrically, the Blue Grism is employed to obtain spectroscopy for these objects. By dispersing light into a spectrum, the Grism allows for observing individual spectral features in low resolution. Key features such as the presence of the Lyα emission line and its associated absorption features are analysed to confirm the quasar's nature. The spectral resolution provided by the Blue Grism is usually enough to enable a detailed examination of these features. The main advantage of using the Blue Grism for quasar selection is its ability to detect faint high-redshift quasars that might otherwise be buried among more numerous contaminants, such as brown dwarfs or intermediate-redshift galaxies. The unique spectral features of quasars help to reduce false positives by distinguishing quasars from these other sources.

To demonstrate the strength of the *Euclid* Blue Grism, we refer to Figure 1.8, which presents the simulated spectra of high-redshift quasars, highlighting their characteristics when observed with the *Euclid* Blue Grism. The figure is composed of two main parts:

1. Original Spectrum: The left side of the figure displays the original spectrum of the quasar ULAS J1342+0928 (z=7.54), discovered by Bañados et al. 2018. This serves as a reference to assess how different models, incorporating varying hydrogen neutral fractions, appear when processed through the Blue Grism.
2. Model Variants: The figure also presents three simulated spectra, each corresponding to different values of the hydrogen neutral fraction (x(HI)=0.1, 0.5, 1.0). These models illustrate how the strength of the quasar's damped Lyman-α absorption feature varies with the neutral hydrogen content along the line of sight. In astrophysics, this variation is crucial, as the presence of neutral hydrogen directly influences light absorption, thereby affecting the observed properties of the quasar. The flux is modelled as zero for wavelengths below the Lyman break (which occurs at 121.6 nm), reflecting the expected physical conditions at high redshift, where intergalactic neutral hydrogen absorbs light bluewards with respect to Lyα. By comparing the spectral features across these models, researchers can infer key properties of the intergalactic medium and the environmental conditions at early cosmic epochs.

While we are still waiting for actual observations from the Blue Grism, the Red Grism has already demonstrated its value by identifying a new quasar at z>5 and discovering several previously unknown brown dwarfs (Banados et al. 2025 (submitted)).

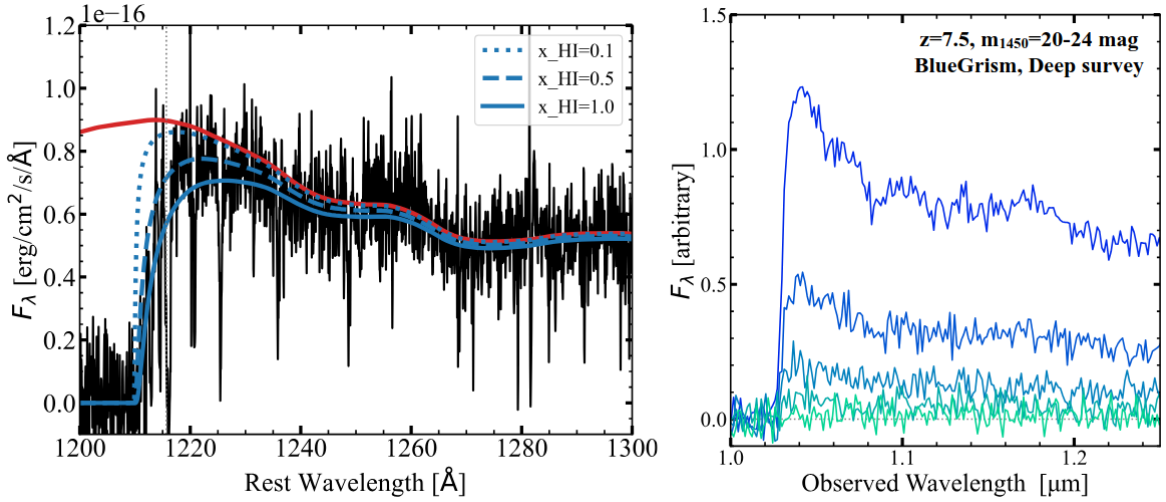


Figure 1.8 – *(left)* The input SED models for high-redshift quasars. The black line is the original spectrum presented in Bañados et al. (2018), in the rest frames and the red line, is the PCA model that they reconstructed from the observed continuum redward of Lymanα. The blue lines show our input quasar SED models with the damping wing profile (hydrogen neutral fraction x(HI) = 0.1, 0.5, 1.0. (Right:) Simulated Blue Grism spectra for z = 7.5 quasars (x(HI) = 0.5) using fastSpec, where the deep survey depth is assumed. The input spectra have m_{1450} = 20–24 magnitudes, with a step larger than one mag at (1 + z) × 1216Å

1.2.4 Bayesian model comparison applied to the *Euclid* Wide Survey

Bayesian formalism has been widely used in classical and modern machine-learning approaches. Traditional applications include works such as Mortlock et al. 2012, Pipien et al. 2018b, and Euclid Collaboration: Barnett et al. 2019, while more innovative methods, like the XDHZQSO technique, have also emerged. In this thesis, I present my work using Bayesian model comparison (Chapter 3) as introduced by Mortlock et al. 2012 to identify high-redshift z>7 quasars. I first adapt the formalism to large-scale surveys, refining the contamination models using the most up-to-date ingredients for quasars and contaminant sources (i.e. MLT brown-dwarfs and red intermediate-z galaxies). The method and associated code are tuned and tested to be used on the *Euclid* Wide Survey data, meaning that the modelling considers the detailed observational characteristics of this survey.

Building on this framework, I have developed an efficient, flexible, and fast tool called `Owl-z` to assign probabilities to high-redshift quasar candidates. I validate and characterise the performance of this tool on simulated *Euclid* data, extracting statistical performance estimators such as completeness, purity and F-measure, a performance metric combining completeness and purity. I also extend the validation by re-identifying all the spectroscopically-confirmed z>7 quasars known today. This work is performed by adapting `Owl-z` to the analysis of the best photometric data available for these objects, different

from the original *Euclid* data and framework used to validate the tool (Chapter 4). In the last step, I apply `Owl-z` to Euclid's first public data release—the Early Observation Data Release—demonstrating its practical use in the real environment of the survey (Chapter 5). Finally, in Chapter 6, I offer general conclusions and perspectives for this work.

2 The Euclid Mission

Contents

2.1 Overview 51
2.2 Observing modes 52
2.3 Survey characteristics 53
2.4 Euclid: a beacon to uncover the mysteries of Dark Matter and Dark Energy 54
2.4.1 Mission objectives 54
2.4.2 Dark matter 55
2.4.3 Dark energy 55
2.5 Legacy Science of Euclid: Insights into the Primeval Universe 56

2.1 Overview

Today the universe is in a rapid and intricate expanse, filled with galaxies, stars, and planetary systems. Yet, despite centuries of astronomical progress, much of its fundamental composition remains inaccessible. Around 95 % of the universe consists of dark matter and dark energy—mysterious components that cannot be directly detected but profoundly shape cosmic structures and their evolution. The nature of dark matter and dark energy is still one the biggest mysteries in modern cosmology, and Euclid's mission is to unravel it. The European Space Agency (ESA) launched the Euclid mission on July 1, 2023, to uncover their nature. This ambitious space-based telescope is designed to study the universe's geometry, the influence of dark energy, and the distribution of dark matter. By providing observations and maps of large-scale structures with exceptional precision, the Euclid mission has as a main objective to shed light on how galaxies form, how invisible forces shape galaxy clusters, and how cosmic expansion has evolved over time (Euclid Collaboration et al. 2024a).

Euclid is now conducting two surveys: the Euclid Wide Survey (EWS hereafter) and the Deep Survey, each designed to maximize its ability to study and access the universe's structure. The Wide Survey will cover by the last data release 14,679 deg ^{2} (Euclid Collaboration: Scaramella et al. 2022), covering over half of the observable extragalactic sky (see Figure 2.1), excluding the portion of the sky that is obscured by the Milky Way—primarily due to interstellar dust and the high density of stars along the Galactic Plane. This vast area observed by Euclid enables the detection of billions of galaxies, providing a large-scale map of cosmic structures influenced by dark matter and dark energy. On the other hand, the Deep Survey focuses on a much smaller region —about 50 square degrees— but observes it with significantly greater sensitivity. This allows Euclid to detect fainter sources, providing a deeper look into the early universe.

In this chapter we summarize the main characteristics of the Euclid Survey that are necessary to understand the work done in this thesis. The reader can find more details in the reference article by Euclid Collaboration et al. (2024a).

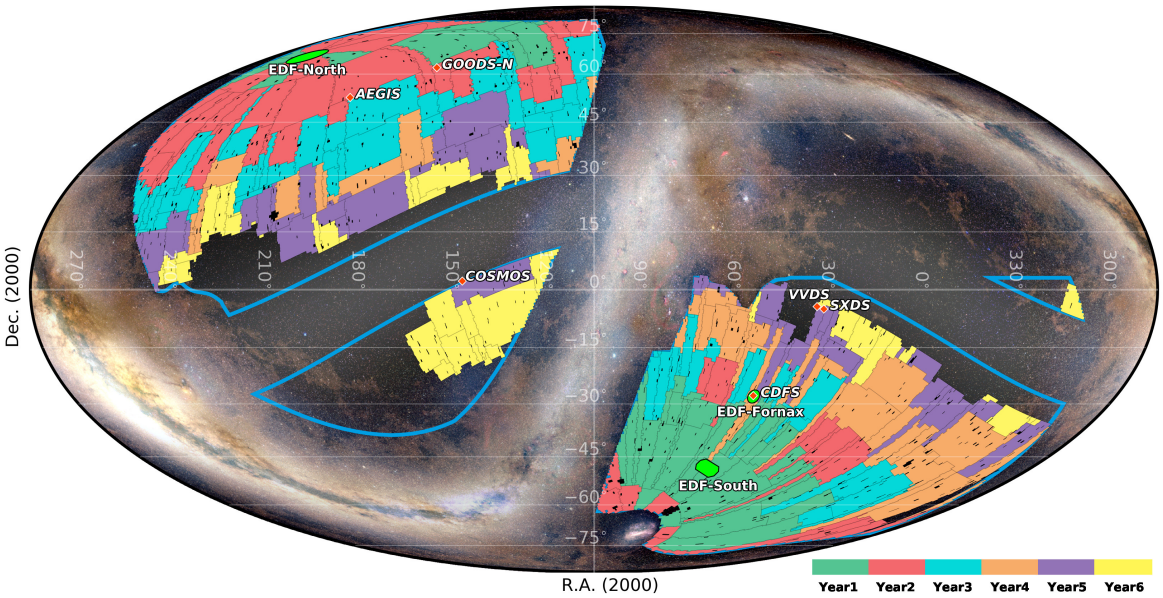


Figure 2.1 – EWS coverage and colour-coded yearly progress in an all-sky Mollweide projection. The blue borders enclose the 16,000 deg^2 ROI that contains the 13,416 deg^2 observed sky of the EWS. Small dark regions within the EWS are masks for stars brighter than 4 mag. Image credits: Euclid Collaboration: Scaramella et al. (2022)

2.2 Observing modes

The Euclid telescope hosts two main instruments:

1. The Visible Imager (VIS): A high-resolution imager that captures light in the visible spectrum. VIS is essential for the precise measurements of galaxy shapes, which are crucial for gravitational lensing studies.
2. The Near Infrared Spectrometer and Photometer (NISP): This instrument captures images and spectra in the near-infrared range. Using the NISP instrument, we can measure the redshifts of galaxies, allowing us to determine their distances and create a three-dimensional cosmic map. This instrument was partly made at the Laboratoire d'Astrophysique de Marseille (LAM)

The VIS and NISP instruments allow Euclid to survey the universe across a broad range of wavelengths, providing detailed data on galaxy shapes, distances, and large-scale structures.

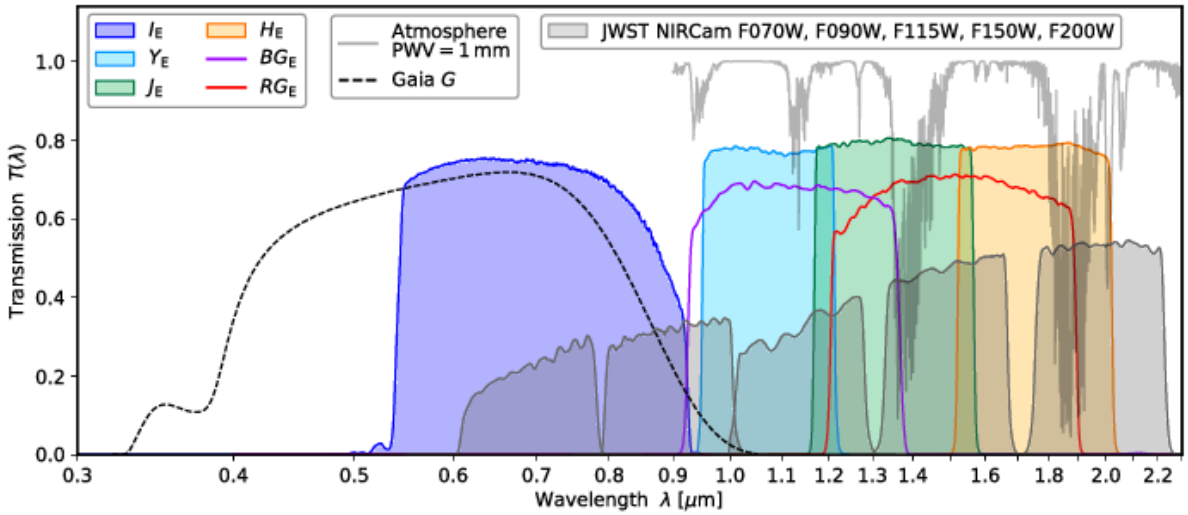


Figure 2.2 – Spectral transmission of *Euclid*'s imaging (VIS: I_E; NISP: Y_E, J_E, H_E) and spectroscopic channels (NISP: Blue grism BG_E and Red grism RG_E) at the beginning of the mission. The expected transmission loss at the end of the mission due to space weathering and non-volatile contamination is at most 0.05. For reference, the Gaia G passband is shown from their third data release (Gaia Collaboration et al. 2023b), and some of the *James Webb* Space Telescope passbands of their Near Infrared Camera Rieke et al. NIRCam; 2005. Figure and caption from (Euclid Collaboration et al. 2024a).

To achieve its scientific objectives, the Euclid mission employs a suite of filters across its imaging and spectroscopic instruments, specifically in the visible (VIS) and near-infrared (NIR) ranges. The I_E channel utilizes a single filter, while the NIR channel incorporates three filters: Y_E, J_E, and H_E(see Figue 2.2) (Euclid Collaboration et al. 2022), which are designed to enhance photometric redshift precision and improve the detection of low-surface-brightness features. In comparison, the Gaia mission primarily focuses on optical wavelengths with its G band. At the same time, JWST offers a broader range of NIR filters, including F070W, F090W, F115W, F150W, and F200W, which are optimized for deep-space observations. Euclid's filters' spectral responses are designed to complement the capabilities of both Gaia and JWST, allowing for a comprehensive analysis of cosmic structures and the universe's evolution. Notably, while Gaia excels in astrometric measurements and large-scale surveys, and JWST provides high-resolution imaging and spectroscopy, Euclid's unique combination of VIS and NIR filters aims to bridge these capabilities, focusing on weak lensing and the large-scale structure of the universe (Euclid Collaboration et al. 2024a).

2.3 Survey characteristics

We provide below the main characteristics of the EWS and ancillary data used in this thesis. The reader can find additional details in Euclid Collaboration: Scaramella et al. (2022). The areas covered by the EWS, as shown in Fig. 2.1, are adjusted in extent and

size to maximise the over- lap with the ground-based surveys providing the complementary data needed for photometric redshift estimation. The use of these ancillary data is also important for our needs.

Regarding the VIS imaging, it is expected to achieve a SNR $\geq$10 for extended sources within a diameter of 1.2 times the FWHM of the PSF at $I_{\rm E}$= 24.5 AB mag. For NISP imaging, a SNR$\geq$5 will be obtained for point sources with m_{AB} = 24.0 in all three NISP bands. Complementary ground-based data is (or will be) available as a result of coordinated campaigns with different telescopes. It can be summarized as follows:

— In the Southern hemisphere, DES (completed in 2019) covers about 3750 deg^2 on the EWS. The depht for a 10σ detection of a point source in a 2arcsec diameter aperture is: g = 24.5, r = 24.1, i = 23.6, and z = 23.4. This photometry will be superseded by the Large Synoptic Survey Telescope (LSST) (Ivezić et al. 2019a), in the same filters. The LSST overlaps with 7534 deg^2 of the EWS.

— In the Northern hemisphere, photometric data will be provided by a a new "collaboration of collaborations" set up in 2017: the Ultraviolet Near Infrared Optical Northern Survey (UNIONS), a wide field ugriz survey of the extragalactic sky. It combines the efforts of observations made at the Canada-France-Hawaii Telescope (CFHT), the Panchromatic Survey Telescope and Rapid Response System (Pan-STARRS), the Wide Imaging with Subaru-Hyper Suprime-Cam Euclid Sky (WISHES), and the Waterloo-Hawaii-IfA g-band Survey (WHIGS). The depths achieved by UNIONS are: u = 23.6, g = 24.5, r = 24.1, i = 23.7, z = 23.4 (10σ for a point source in a 2arcsec diameter aperture)

2.4 Euclid: a beacon to uncover the mysteries of Dark Matter and Dark Energy

2.4.1 Mission objectives

The Euclid mission has several scientific objectives:

1. Mapping the universe's large-scale structure, mapping the positions of billions of galaxies, creating a three-dimensional map of the universe. This map will allow us to study how galaxies are distributed and cluster together under the influence of dark matter and energy.
2. Measuring cosmic acceleration: Euclid will determine how fast the universe has expanded at different times by observing distant galaxies and galaxy clusters. This will help us understand the nature of dark energy and whether it behaves as a constant or varies over time.
3. Gravitational lensing measurements: Euclid will measure weak gravitational lensing—the subtle distortion of light from distant galaxies caused by massive objects (such as dark matter) in the foreground.
4. Exploring the nature of dark energy: The mission will study the "cosmic distance ladder," using galaxy clustering and lensing to probe the geometry and expansion of the universe.

2.4.2 Dark matter

Dark matter is an invisible component that does not interact with electromagnetic radiation, making it impossible to observe directly using our electromagnetic radiation emitters/receivers. However, we can map its distribution and study its effects indirectly using the kinematic properties of galaxies and large-scale structures, as well as gravitational lensing, which is sensitive to the total amount of matter (dark and baryonic matter).

One of Euclid's main scientific objectives is to answer questions about dark matter by providing high-resolution optical and infrared observations over a vast extragalactic sky, enabling precision measurements of cosmic structures. Using its weak gravitational lensing probe, Euclid will measure the distortion of light from distant galaxies (lensed sources) as it passes near massive objects (lenses), exposing the mass distribution of both baryonic (visible) and dark matter. Euclid will also observe galaxy clustering to trace the history of cosmic structure formation, providing insights into the dark matter distribution; in addition, it will perform cosmic shear measurements to derive the matter density of the universe (Euclid Collaboration et al. 2024a). This makes it a pivotal stage given the evidence for a dark universe presented in previous studies such as those by Riess et al. 1998 and Perlmutter et al. 1999, which highlighted the acceleration of the universe's expansion and proved by implication the gravitational effects of dark matter. Euclid's data will complement findings and observations from the WMAP (Hinshaw et al. 2013) and Planck satellite missions (Planck Collaboration et al. 2020), which have established foundational cosmological models incorporating dark matter. In addition, theoretical frameworks will be tested directly on Euclid's observations, allowing for a better understanding of the interplay between dark matter and dark energy (Albrecht et al. 2006; Amendola et al. 2018). Euclid will refine the characterization of dark matter's distribution and dynamics through its measurements of cosmic structures, unravelling some of the remaining ambiguities pointed out in recent analyses (Battaglieri et al. 2017; Eisenstein et al. 2005; Betoule et al. 2014) and seeking to clarify the properties of dark matter.

2.4.3 Dark energy

In addition to dark matter, the Euclid mission will also indirectly probe dark energy, as it is believed to be the main driver for the universe's accelerated expansion. To do so, Euclid will use diverse cosmological probes, such as weak lensing, galaxy clustering, and baryon acoustic oscillations (BAO). These probes will allow us to gather data that can inform models of dark energy (Euclid Collaboration et al. 2024a). By measuring how these aspects evolve, researchers hope to learn about the properties of dark energy. Weak gravitational lensing allows for the mapping of dark matter and will also give information about the universe's expansion rate influenced by dark energy (Wright et al. 2010b). The galaxy clustering probe will help us to understand how dark energy affects cosmic evolution (Schaye et al. 2015). The BAO, which are acoustic waves from the early universe imprinted in the distribution of galaxies, help measure the universe's expansion history (Schmidt 1963). Lastly, the large survey area will provide an excellent opportunity to do statistical analyses of dark energy's effects across vast cosmic distances (Kauffmann et al.

2003).

2.5 Legacy Science of Euclid: Insights into the Primeval Universe

Beyond its main mission objectives—mapping the universe's large-scale structure and probing dark matter and dark energy—the Euclid mission has a big potential for legacy science. Legacy science refers to the long-term scientific contributions a mission can make by providing vast datasets that future generations of researchers can explore and analyse outside the scope of the mission's main objectives, with particular importance for studying the early universe, also known as the primaeval or primordial universe.

During my PhD, I have been an active member of the Euclid Consortium, particularly in the science working group "Primeval Universe", focusing on the first sources formed in the early universe. In this respect, I have participated in many activities, mainly in the subgroups of this science working group. I have contributed to the studies conducted in the LBG (Lyman Break Galaxies) group, and the Quasars group. The goal of these groups is to select, confirm and study high-redshift (z>7) LBG galaxies and quasars. My thesis work precisely focuses on the study of quasars.

Euclid's deep observations will allow us to explore the high redshift universe. This era is crucial as it was when galaxy formation and evolution were still in their early stage. In addition, Euclid will allow us to study in more depth the epoch of reionisation, which aligns well with the objectives of this thesis. Indeed, the large field of view covered by the EWS is ideally suited to identify and study rare objects at very high-redshifts, such as quasars at z>7. As it is discussed in the next Chapter, these objects are particularly difficult to identify, making the study of their properties particularly challenging.

3 Bayesian Comparison model

Contents

3.1 Introduction . . . 58
3.2 Principles . . . 58
3.2.1 Formalism . . . 58
3.2.2 Gaussian truncation . . . 60
3.2.2.1 Some Hypotheses . . . 60
3.2.2.2 Notations . . . 60
3.2.2.3 Gaussian Distributions . . . 60
3.2.2.4 Truncated Gaussian Distributions for Fluxes . . . 61
3.2.2.5 Conversion to Magnitudes . . . 61
3.2.2.6 Case Where $F^\star = 0$ . . . 62
3.2.2.7 Non-detection Case . . . 62
3.2.3 The likelihood form . . . 62
3.2.4 Extraction of the Maximum A Posteriori parameters . . . 63
3.3 Bayesian approach to the search for high redshift quasars . . . 63
3.3.1 High-redshift quasars population model . . . 64
3.3.2 MLT stars population model . . . 65
3.3.2.1 Thin disc model . . . 69
3.3.2.2 Model templates . . . 72
3.3.3 Contaminant galaxy model . . . 72
3.3.3.1 Understanding the nature of contaminant galaxies . . . 73
3.3.3.2 Luminosity Function of contaminant galaxies . . . 76
3.3.4 Unbalanced prior problem . . . 80

3.1 Introduction

The Bayesian comparison model is of great significance in the field of statistical inference, offering a systematic approach for evaluating disparate hypotheses or models in light of empirical evidence. The incorporation of prior information and its updating with evidence enables the Bayesian approach to offer a more nuanced understanding of uncertainties and probabilities than is possible with traditional frequentist methods. This model is particularly significant in fields that deal with complex, noisy, or incomplete data, such as astronomy. This point is illustrated by the work ofMortlock et al. (2012), where the Bayesian model played a pivotal role in identifying the most distant known quasar at that time. The Bayesian approach enabled the researchers to meticulously integrate prior knowledge with new observational data despite the challenges of incomplete datasets and observational uncertainties. This exemplifies the significance of Bayesian methods in astronomy, where the capacity to compare models with varying degrees of uncertainty is crucial for discoveries and precise measurements in the vast, often obscure, Universe.

The principal aim of this study is to employ Bayesian selection methodology to ascertain the posterior probability that a specific observed astronomical source, initially identified as non-detected in optical bands, is classified as a high-redshift quasar or as a contaminant. This is achieved by considering the prior surface density of the quasars and contaminants. Contaminant populations of high-redshift galaxies and quasars in near-IR and Mid-IR bands are supposed to be well known (see e.g. Mortlock et al. 2012; Euclid Collaboration: Barnett et al. 2019) and consist of early-type galaxies at intermediate redshifts ($1 \lessapprox z \lessapprox 2$, see e.g. Euclid Collaboration: van Mierlo et al. 2022) and low-mass dwarf stars of spectral types late-M, L or T (MLT hereafter) (Stern et al. 2007; Caballero et al. 2008; Wilkins et al. 2014; Hainline et al. 2024a).

This chapter presents the probabilistic selection method employed for the identification of high-redshift quasars in the *Euclid* Wide Survey. (hereafter EWS). The method may be adapted for use with other wide-field surveys and for the selection of other high-redshift astronomical sources. This method aims to identify and select a tinct set of potentially high-redshift sources, referred to as candidates, by effectively distinguishing them from potentially low-redshift sources designated as contaminants. Our approach is directly inspired by the Bayesian method developed by Mortlock et al. (2012) and Euclid Collaboration: Barnett et al. (2019) to search for quasars at high redshifts.

In this chapter, I first present the Bayesian formalism in a general context before adapting it to the selection of high-redshift quasars in the EWS.

3.2 Principles

3.2.1 Formalism

Bayesian model selection (Robert 2007) is a relevant method to establish the nature of a source in a survey given photometric data (see Mortlock et al. 2012, and references therein). The posterior probability that a given source is of a type $t \in \mathcal{T}$, given the data $\mathbf{D}$, is driven by two terms: the prior probability $\mathbb{P}(t)$ of the source being of type t, and

the integrated likelihood or the evidence of the data given the type of source, $\mathbb{P}(\mathbf{D}|t)$. The latter is the marginal likelihood of the data, which is the integral of the likelihood of the data given the parameters of the model, $\mathbb{P}(\mathbf{D}|\boldsymbol{\theta}_t, t)$, with respect to the prior distribution of the parameters, $\mathbb{P}(\boldsymbol{\theta}_t|t)$. The marginal likelihood is the key quantity in Bayesian model selection. Unlike methods based on best-fit parameters and type, Bayesian model selection accounts for the diversity of photometric data of a given type. The posterior probability of a source being of type t is then given by Bayes' theorem:

$$\mathbb{P}(t|\mathbf{D}) = \frac{\mathbb{P}(t)\mathbb{P}(\mathbf{D}|t)}{\sum_{t'\in\mathcal{T}} \mathbb{P}(t')\mathbb{P}(\mathbf{D}|t')} \tag{3.1}$$

However, in this work, I use the marginalised evidence form of Bayes's theorem:

$$\mathbb{P}(t|\mathbf{D}) = \frac{W(t|\mathbf{D})}{\sum_{t'\in\mathcal{T}} W(t'|\mathbf{D})} \tag{3.2}$$

where $W(t|\mathbf{D}) = \mathbb{P}(t)\mathbb{P}(\mathbf{D}|t)$ is the weighted evidence of the model t given the data $\mathbf{D}$. The weighted evidence of type t is given by:

$$\begin{aligned} W(t|\mathbf{D}) &= \mathbb{P}(t)\mathbb{P}(\mathbf{D}|t) = \mathbb{P}(t)\int \mathbb{P}(\boldsymbol{\theta}_t|t)\mathbb{P}(\mathbf{D}|\boldsymbol{\theta}_t, t)\mathrm{d}\boldsymbol{\theta}_t \\ &= \int \rho_t(\boldsymbol{\theta}_t)\mathbb{P}(\mathbf{D}|\boldsymbol{\theta}_t, t)\mathrm{d}\boldsymbol{\theta}_t \end{aligned} \tag{3.3}$$

where $\boldsymbol{\theta}_t$ is the vector of parameters of t-type object and $\rho_t(\boldsymbol{\theta}_t)$ is the prior value of surface density of detected objects of type t and parameters $\boldsymbol{\theta}_t$ in the sky. A photometric dataset is a set $\widehat{\mathbf{F}} = (\widehat{F}_1, \ldots, \widehat{F}_N)$ of N measurements of the flux of a source in N photometric bands, and their standard error $\widehat{\boldsymbol{\sigma}} = (\widehat{\sigma}_1, \ldots, \widehat{\sigma}_N)$. The statistical model of each type of source is built on a set of spectral energy densities (SEDs) of n_t templates. Given the i-th SED of type t, we can compute expected fluxes in each band, $\mathbf{F}_{t,i}(\boldsymbol{\theta}_t) = (F_{t,i,1}(\boldsymbol{\theta}_t), \ldots, F_{t,i,N}(\boldsymbol{\theta}_t))$ at value $\boldsymbol{\theta}_t$ of the parameter. The likelihood of the data given $\boldsymbol{\theta}_t$ and t is set as the mixture of multivariate densities

$$\mathbb{P}\Big[\widehat{\mathbf{F}}\Big|\boldsymbol{\theta}_t, t\Big] \sim \frac{1}{n_t}\sum_{i=1}^{n_t} p\Big(\widehat{\mathbf{F}}\Big|\mathbf{F}_{t,i}(\boldsymbol{\theta}_t), \widehat{\boldsymbol{\sigma}}\Big) \tag{3.4}$$

centred at $\mathbf{F}_{t,i}(\boldsymbol{\theta}_t)$ and with dispersion given by $\widehat{\boldsymbol{\sigma}}$.

The above mixture sets a uniform prior distribution on the SED templates modelling each population. If the photometric bands do not overlap, we can neglect the correlation between the fluxes in different bands. Thus, we rely on a product of Gaussian distributions, given by

$$p\Big(\widehat{\mathbf{F}}\Big|\mathbf{F}_{t,i}(\boldsymbol{\theta}_t), \widehat{\boldsymbol{\sigma}}\Big) = \prod_{j=1}^{N} \mathcal{L}_j \tag{3.5}$$

where each Gaussian distribution is given by

$$\mathcal{L}_j = \frac{\exp\left(-\frac{1}{2}\left(\frac{\widehat{F}_j - F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)^2\right)\mathbb{I}_{\widehat{F}_j \geq 0}}{\widehat{\sigma}_j\sqrt{2\pi}} \tag{3.6}$$

3.2.2 Gaussian truncation

3.2.2.1 Some Hypotheses

— Fluxes are positive quantities.
— The magnitude is related to the flux by an equation of the form:

$$F = F_0 \exp(-am), \quad a > 0.$$

3.2.2.2 Notations

If x follows a standard normal distribution, we denote it as $x \sim \mathcal{N}(0,1)$. Its probability density function is given by:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Its cumulative distribution function is given by:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(t)\, dt.$$

Using the error function, we can write:

$$\Phi(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right).$$

3.2.2.3 Gaussian Distributions

If x follows a Gaussian distribution centered around $F^\star$ (model flux) with a standard deviation $\sigma^\star$, we write $x \sim \mathcal{N}(F^\star, \sigma^\star)$.

The density of x is given by:

$$\frac{1}{\sigma^\star}\varphi\left(\frac{x - F^\star}{\sigma^\star}\right).$$

The cumulative distribution function of x is given by:

$$\Phi\left(\frac{x - F^\star}{\sigma^\star}\right).$$

I do not want to use such a distribution for the observed flux F, which is affected by a Gaussian error with standard deviation σ around $F^\star$. If I did so, there would be a

non-zero probability that F could be negative, which is nonsensical. This probability is given by the value of the cumulative distribution function at zero, that is:

$$\Phi\left(-\frac{F^\star}{\sigma^\star}\right).$$

As soon as $F^\star \leq 2\sigma^\star$, this probability becomes significant (greater than 5%). In particular, when $F^\star$ is close to zero, this probability approaches 1/2.

3.2.2.4 Truncated Gaussian Distributions for Fluxes

I, therefore, want to use a distribution for the observed flux F that remains positive. To achieve this, I will truncate the Gaussian distribution $\mathcal{N}(F^\star, \sigma^\star)$ to the interval $[0, \infty)$. Here, truncation means conditioning the Gaussian distribution to remain positive.

We denote this truncated Gaussian distribution as $\mathcal{N}^+(F^\star, \sigma^\star)$. Let $F \sim \mathcal{N}^+(F^\star, \sigma^\star)$. The density of F is given by:

$$\frac{1}{\lambda^\star \sigma^\star}\varphi\left(\frac{F - F^\star}{\sigma^\star}\right)\mathbb{1}_{F\geq 0}, \quad \text{where } \lambda^\star = 1 - \Phi\left(-\frac{F^\star}{\sigma^\star}\right).$$

The cumulative distribution function of F is given by:

$$\frac{1}{\lambda^\star}\left(\Phi\left(\frac{F - F^\star}{\sigma^\star}\right) - \Phi\left(-\frac{F^\star}{\sigma^\star}\right)\right).$$

3.2.2.5 Conversion to Magnitudes

We assume that the magnitude m is related to the flux F by the relation:

$$F = F_0 \exp(-am), \quad a, F_0 > 0.$$

Thus, the magnitude m is given by:

$$m = -\frac{1}{a}\log\left(\frac{F}{F_0}\right) = -\frac{1}{a}\log(F) + m_0, \quad \text{where } m_0 = \frac{1}{a}\log(F_0).$$

Let $F \sim \mathcal{N}^+(F^\star, \sigma^\star)$. The Jacobian of the transformation is given by:

$$\left|\frac{dF}{dm}\right| = aF.$$

The density of m is given by:

$$\frac{1}{\lambda^\star \sigma^\star}\varphi\left(\frac{F - F^\star}{\sigma^\star}\right) aF\mathbb{1}_{F\geq 0}.$$

The cumulative distribution function of m is identical to that of F.

3.2.2.6 Case Where $F^\star = 0$

If $F^\star = 0$, then $\lambda^\star = 1/2$ and the density of F is given by:

$$\frac{2}{\sigma^\star}\varphi\left(\frac{F}{\sigma^\star}\right)\mathbb{1}_{F\geq 0}.$$

And the density of m is given by:

$$\frac{2}{\sigma^\star}\varphi\left(\frac{F}{\sigma^\star}\right)aF.$$

This density is indeed zero if $F = 0$. However, in this case, it is likely below the detection threshold of the instrument. I suggest treating it as a case of non-detection.

3.2.2.7 Non-detection Case

We assume that there is a threshold $F_{\min}$ below which the flux is not detected. We assume that $F_{\min} > 0$. In the case of non-detection, the likelihood should reflect the probability of being below the threshold rather than a probability density.

The probability of being below the threshold $F_{\min}$ is given by the cumulative distribution function at $F_{\min}$:

$$\frac{1}{\lambda^\star}\left(\Phi\left(\frac{F_{\min} - F^\star}{\sigma^\star}\right) - \Phi\left(-\frac{F^\star}{\sigma^\star}\right)\right).$$

When $F^\star = 0$, this probability is given by:

$$\frac{1}{2}\left(\Phi\left(\frac{F_{\min}}{\sigma^\star}\right) - \frac{1}{2}\right) = \frac{1}{4}\operatorname{erf}\left(\frac{F_{\min}}{\sigma^\star\sqrt{2}}\right).$$

3.2.3 The likelihood form

I rewrite the Likelihood in its truncated form as follows:

$$\mathcal{L}_j = \frac{\exp\left(-\frac{1}{2}\left(\frac{\widehat{F}_j - F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)^2\right)\mathbb{I}_{\widehat{F}_j\geq 0}}{\widehat{\sigma}_j\sqrt{2\pi}\left(1 - \mathcal{N}\left(-\frac{F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)\right)} \tag{3.7}$$

where $\mathbb{I}_{\widehat{F}_j\geq 0}$ is the indicator function that the flux is non-negative and $\mathcal{N}$ is the cumulative distribution function of the standard normal distribution given by

$$\mathcal{N}(z) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right).$$

In a general context, faint sources are measured with a low signal-to-noise ratio (SNR) in one or more bands. Forced photometry is a practice that is applied in many surveys,

including *Euclid*, using raw images from which aperture fluxes are measured for all undetected sources. The likelihood is then calculated using Eq. (3.7) even when the forced photometry flux has a negative or null value. However, when this is impossible, i.e. there is no estimation of the flux in a particular band, we re-write the likelihood with an unknown measured flux. This allows us to calculate the probability that a source to be observed with a measured flux below the stated detection limit. This probability can be expressed as follows:

$$\mathcal{L}_j = \frac{\mathcal{N}\left(\frac{\widehat{F}_{lim} - F_{t,i,j}(\boldsymbol{\theta}_t)}{\sqrt{\widehat{\sigma}_j}}\right)}{1 - \mathcal{N}\left(-\frac{F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)} \tag{3.8}$$

3.2.4 Extraction of the Maximum A Posteriori parameters

In general, a χ^2-criteria that is used to produce the best-fit parameters in the frequentist approach appears in the exponent of the likelihood given in Eq. (3.7). However, we use a Maximum A Posteriori (MAP) in this work.

Maximum A Posteriori estimation is a fundamental concept in Bayesian statistics used to extract parameters that maximise the posterior distribution given the observed data. It plays a crucial role in various fields, such as machine learning, signal processing, and cosmology, where the goal is to infer the most likely parameters of a model based on available data while incorporating prior knowledge. The MAP estimation is particularly valuable when dealing with noisy or incomplete datasets, where maximum likelihood estimation might not yield meaningful results. In Bayesian inference, we are interested in estimating the probability of model parameters $\boldsymbol{\theta}_t$ given the observed data $\mathbf{D}$.

Since the evidence $\mathbb{P}(\mathbf{D})$ is constant with respect to $\boldsymbol{\theta}_t$, it can be ignored during the optimisation process, leading to the simplified MAP criterion:

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta} P(D|\theta)P(\theta) \tag{3.9}$$

In the following chapters, I use θ_{MAP} as parameters to analyse the results and these parameters can be redshift, magnitude, spectral templates, and stellar spectral types.

3.3 Bayesian approach to the search for high redshift quasars

The formalism introduced in the previous section can be applied to the search of high-redshift quasars using Bayes' law. The first step is to identify all the relevant populations. In this case, these are the brown dwarfs (s), the contaminant galaxies (g) and, last but not least, the quasars (q) Thus giving us a $\mathcal{T} = (q, s, g)$

The probability of an object with a given set of photometric data points, designated as

D, being a quasar can now be expressed as follows:

$$\mathbb{P}(q|\mathbf{D}) = \frac{W(q|\mathbf{D})}{W(q|\mathbf{D}) + W(s|\mathbf{D}) + W(g|\mathbf{D})} \tag{3.10}$$

In the following sections, we define in detail each component of this probability $W(q|\mathbf{D})$, $W(s|\mathbf{D})$ and $W(g|\mathbf{D})$. As previously discussed, the likelihood term is identical for the three weighted evidences. The following sections will focus on the modelling of the prior surface densities for each of the populations.

3.3.1 High-redshift quasars population model

The statistical model of the quasar population is based on a collection $n_{\rm QSO}$ of SEDs (also referred to as spectral templates or templates) and a set of parameters $\boldsymbol{\theta}_{\rm QSO}$ describing the QSO properties. The parameters are the redshift z and the apparent magnitude in a reference band, which we choose to be the *Euclid* H band, denoted $H_{\rm E}$. This allows us to compute the expected fluxes in each band, $\mathbf{F}_{{\rm QSO},i}(\boldsymbol{\theta}_{\rm QSO})$ at the $\boldsymbol{\theta}_{\rm QSO} = (H_{\rm E}, z)$ value of the parameter set given the i-th SED model. The QSO likelihood of the data given $\theta_{\rm QSO}$ is defined as the mixture of multivariate densities in Eq. (3.4). For the purposes of this manuscript, we let the redshift vary in the range $7 < z < 12$, since we aim for a maximum redshift of 12 as the *Euclid* bands will cover up to the H band and the Lyman Alpha emission line, stays in the J band until $z \approx 12$.

To characterise the prior distribution of high-redshift QSOs, we adopt the double power law parametric Luminosity Function (LF) from Willott et al. (2010) and adjust the parameters fit to Matsuoka et al. (2023):

$$\Phi(M_{1450}, z) = \frac{10^{k(z-7)}\Phi^*}{10^{0.4(\alpha+1)(M_{1450}-M^*_{1450})} + 10^{0.4(\beta+1)(M_{1450}-M^*_{1450})}}, \tag{3.11}$$

where M_{1450} is the quasar rest-frame absolute magnitude at $\lambda = 1450$Å and $\Phi(M_{1450}, z)$ is the number of quasars per magnitude bin, and unit volume at redshift z. The parameters are the normalisation Φ^*, the break magnitude M^*_{1450}, the bright end slope β and the faint end slope α, whose values can be found in Table 3.1. Using the apparent magnitude in the $H_{\rm E}$band, we can write the surface density of QSO in the sky in $\mathrm{mag}^{-1}\times\mathrm{deg}^{-2}\times\ \mathrm{d}z^{-1}$ units as:

$$\rho_{\rm QSO}(H_E, z) = \frac{1}{4\pi} \times \frac{dV_c}{dz} \times \Phi\left[H_E - \mu - K_{corr}(z), z\right], \tag{3.12}$$

where

— μ represents the distance modulus, defined as a function of the luminosity distance D_L by

$$\mu = \frac{5\log_{10}(D_L/10\ \mathrm{pc})}{\mathrm{pc}}, \tag{3.13}$$

— $K_{\rm corr}(z)$ denotes the K-correction, which converts the absolute magnitude at rest-frame $\lambda = 1450$Å into the observed H-band magnitude, and
— $\frac{1}{4\pi} \times \frac{dV_c}{dz}$ is the comoving volume element per steradian and per redshift interval dz.

We show in Fig. 3.1 the number of quasars per redshift bin and over the whole 15,000 deg^2 of the EWS (Euclid Collaboration: Scaramella et al. 2022), derived from Eq. 3.12. This shows that *Euclid* will detect hundreds of quasars between redshifts 7 and 8, and possibly a few up to $z \sim 11$.

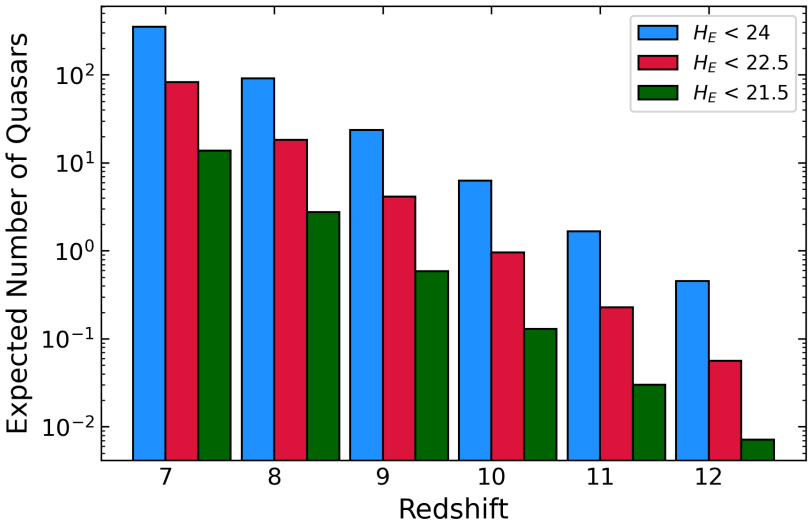


Figure 3.1 – The expected number of quasars per redshift bin in the 15,000 deg^2 EWS given the high-redshift QSO Luminosity Function adopted in this paper. In blue are shown the numbers of quasars expected up to $H_{\rm E}$<24, in crimson with $H_{\rm E}$<22.5 and in green with $H_{\rm E}$<21.5

Table 3.1 – Quasar Luminosity function parameters

α	β	k	M^*_{1450}	ϕ^*10^{-8} (Mpc^{-3}mag^{-1})
−1.2	−3.34	−0.47	−24.38	0.475

3.3.2 MLT stars population model

To define $\rho_{\rm s}$, we use the standard model for the Milky Way proposed by Caballero et al. (2008) and Pipien et al. (2018b). This model has been proposed and implemented for determining the spatial density of brown dwarfs per magnitude interval and spectral type. the parameters for the brown dwarfs model are then $\boldsymbol{\theta}_{\rm s} = (H_{\rm E}, spt)$. where $H_{\rm E}$is

the reference magnitude in the *Euclid* H-band, and *spt* is the spectral types from M6 to T9. Using the notation described in Sect. 3.2 The Milky Way can be effectively modelled using a double exponential function that accounts for both thin and thick discs and a halo power law. Fig. 3.2 shows these components as well as the galactic coordinates (l,b) used throughout this study. The thin disc is characterised by a higher density of stars, a more pronounced structure, and a relatively younger stellar population. In contrast, the thick disc is more rarified and extended, containing older stars and a lower density overall. In the context of studying late-type dwarfs, Caballero et al. (2008) argue that it is more appropriate to focus solely on the thin disc model for determining their spatial density.

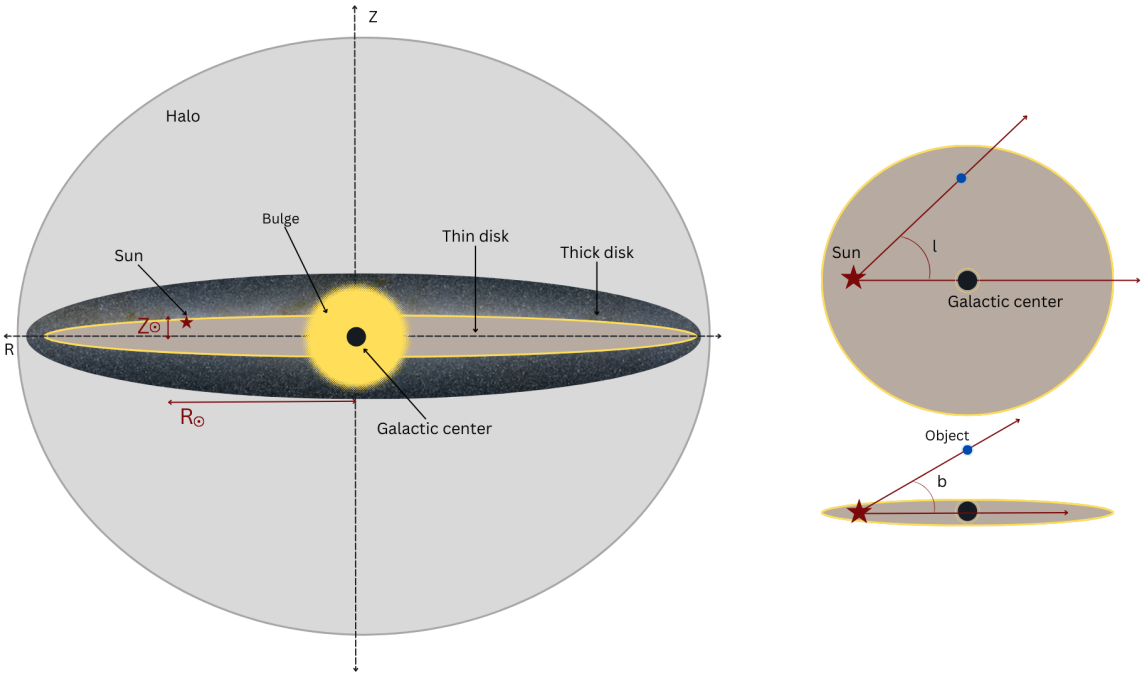


Figure 3.2 – The sketches on the left and right illustrate the Milky Way and the galactic coordinate system, respectively. The right diagram defines the galactic longitude and latitude of an object, which are noted as l and b, respectively.

This decision is based on several key factors. Firstly, the thick disc and halo brown dwarfs are thought to contribute minimally to the overall sample of late-type dwarfs due to their relative faintness. As these objects age, they become increasingly dimmed, making them difficult to detect in deep surveys. For instance, very old thick-disc substellar objects with masses significantly below the hydrogen burning limit may fade to exceedingly faint magnitudes, preventing their identification in observational studies.

Moreover, the normalisation factor for the thick disc in the solar vicinity is relatively low, estimated to be around 13 per cent. This is significantly higher than previous reports, which suggested values as low as 2 percent. The authors highlight that the low normalisation factor, combined with the faintness of thick-disc objects, suggests that the contribution of these stars to the overall population of late-type dwarfs is negligible.

At the magnitudes corresponding to the depth of the EWS (~ 24), the distance to T dwarfs varies from ~ 150 pc for the coldest to ~ 450 pc for the hottest. With a scale

height of the thin disc of the order of 300 pc (Gaia Collaboration et al. 2023a; Vieira et al. 2022), we infer that the sole thin disc approximation will reasonably well represent the population of T dwarfs detectable in the EWS. However, the same will not be true for M dwarfs. An M6 dwarf at an apparent magnitude ≈ 24 is at a distance of ~ 4000 pc, extending well beyond the thin disc into the thick disc or the halo. As a matter of fact, the *James Webb* Space Telescope has revealed in extragalactic fields a significant number of extremely faint and low-temperature brown dwarfs extending well into the thick disc and possibly into the halo, at distances of up to 2 kpc (Hainline et al. 2024b,a; Burgasser et al. 2024). An analysis of the thick disc contamination is therefore warranted.

For this analysis, we follow a similar approach to that used in Ryan &Reid (2016) and compare the distribution of stellar dwarfs in thin and thick discs as a function of magnitude. We adopt a value of 330 pc for the scale height of the thin disc (Caballero et al. 2008), 800 pc for the scale height of the thick disc (Vieira et al. 2023), 2250 pc (Caballero et al. 2008) for the value of the scale length of both discs and a normalisation factor between the thick and thin discs local densities of 10%. Using Eq. 12 of Pipien et al. (2018b), we compare brown dwarf densities for different spectral types in the thin and thick discs for two lines of sight: $b = 90°$ and $(l, b) = (90°, 30°)$ (the lowest Galactic latitude of the EWS is $b = 23°$). For local volume densities in the Galactic plane and absolute *Euclid* magnitudes by spectral type, we refer to Table 2 of Euclid Collaboration: Barnett et al. (2019) and the references therein (Dupuy &Liu 2012; Skrzypek et al. 2016; Bochanski et al. 2010). Late-type M dwarfs can potentially contaminate searches for galaxies or quasars at high redshift in some surveys, whereas, as we will see later, only L and T-type brown dwarfs contaminate the search for $z \geq 7$ with *Euclid*. Consequently, for the purposes of this manuscript, we restrict our analysis to the brown dwarf population within the M dwarf population, that is, to dwarfs of a type later than approximately M6.

The analysis (see Fig. 3.3) indicates that up to *Euclid* magnitude $H_{\rm E} \sim 24$, the thin disc dominates the number of L and T dwarfs at low and high Galactic latitudes. For M6 to M9 dwarfs, the thick disc is the main contributor above H ~ 22.5 magnitudes at high Galactic latitudes, whereas at low Galactic latitudes, the thin disc dominates up to $H_{\rm E} \sim 24.5$ and beyond. For the purposes of our paper, Galactic latitude is, therefore, the dominant parameter affecting the population of contaminating stars. We explore the effect of Galactic latitude on the performance of our code in Sect. 4.5.

This simple analysis assumes that stellar dwarf populations are the same in both thin and thick discs. This oversimplification ignores the differences in age and metallicity between the constituents of the thin and thick discs and, therefore, the significant differences that may exist between the stellar dwarf populations within them. Early M-type dwarfs (M0 to M5) have masses above the hydrogen-burning minimum mass, and their luminosity evolves little on time scales up to 10 billion years, whereas brown dwarfs below this limit cool on much faster time scales (Reid 2013). As a result, late M-type dwarfs evolve into later spectral types, and their luminosities drop rapidly over timescales of a Gyr or less. The thick disc, which is older than the thin disc, should, therefore, be depleted of late-type brown dwarfs (M6 to M9 and L), and T dwarfs should be significantly less luminous in the thick disc than in the thin disc (see also the discussion in Caballero et al. 2008). This suggests that the stellar densities reported in Fig. 3.3 are likely to be seriously

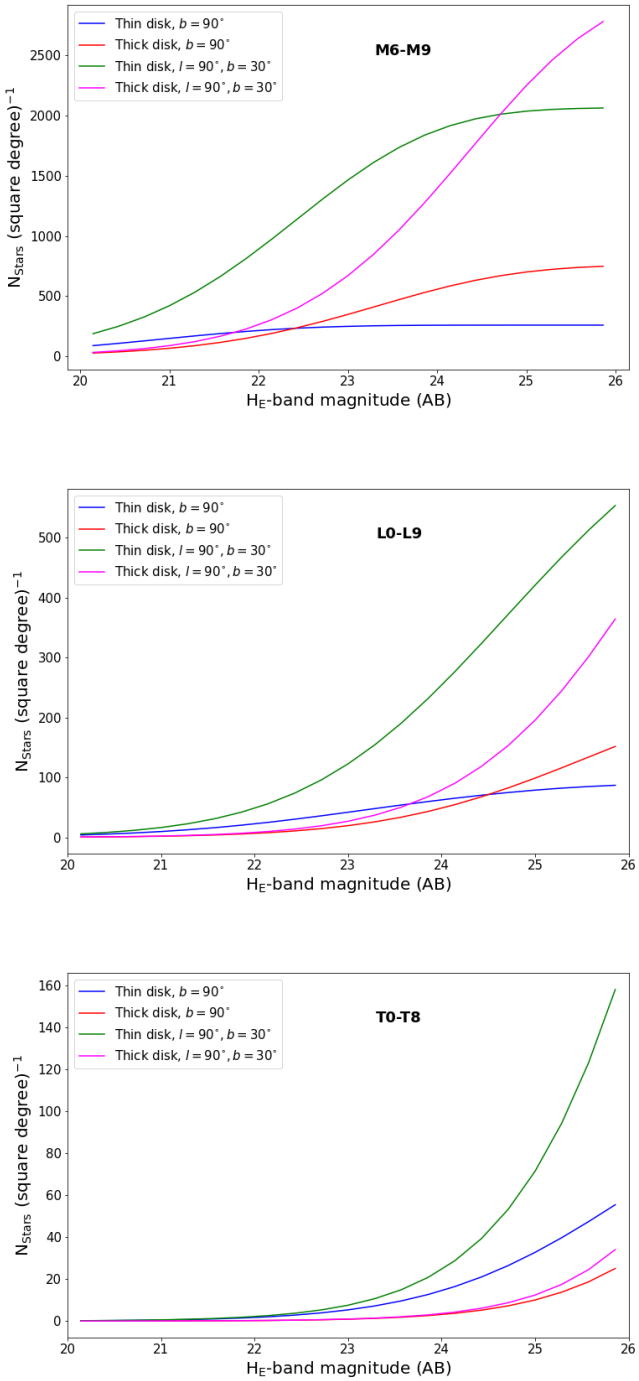


Figure 3.3 – Brown dwarf number counts as a function of H-band magnitude for the thin and thick discs in two different field locations (b=90 and (l,b)=(90,30)). Top: M6 to M9 spectral types, middle: L0 to L9 and bottom: T0 to T8.

overestimated in the thick disc at a given magnitude for spectral types later than M5.

Additionally, Caballero et al. (2008) note that the error associated with assuming only the thin disc model is less than or comparable to the uncertainties in determining other critical parameters, such as the solar galactocentric distance ($R_\odot$), the height above the galactic plane ($Z_\odot$), and the scale height (h_R). This further reinforces the rationale for focusing on the thin disc model, as it simplifies the analysis while still providing a reliable framework for understanding the distribution and density of late-type dwarfs in the Galaxy. By concentrating on the thin disc, researchers can more accurately assess the characteristics and demographics of these objects without the complicating factors introduced by the thick disc.

As a conclusion to this analysis, we restrict the modelling of the Galaxy's brown dwarf population to the thin disc and analyse the impact of stellar density through its dependence on Galactic latitude. In the coming years, *Euclid* and *James Webb* Space Telescope data will enable a more detailed understanding of the relative populations of brown dwarfs in the thin and thick discs.

3.3.2.1 Thin disc model

We now return to the formulation of our Bayesian model for the description of the stellar dwarf population; we write the weighted evidence of this population as follows:

$$W_s = \int_{M3}^{T9} \int_{-\infty}^{+\infty} \rho_s(H_{mod}, spt)\mathbb{P}(D|H_{mod}, spt)\mathrm{d}H_{mod}\mathrm{d}spt.$$

where $\rho_s(H_{mod}, spt)$ is the surface density of the contaminating brown dwarf population of *Euclid* magnitude H_{mod} and spectral type spt, at a given Galactic longitude l and latitude b. We model the spatial distribution across the Galaxy in $\mathrm{mag}^{-1} \times \mathrm{deg}^{-2} \times \mathrm{spt}^{-1}$ units as a function of the heliocentric distance to the star d, assumed to be far smaller than the solar galactocentric distance $R_\odot$ (Caballero et al. 2008):

$$\rho_s(d, l, b, spt) \approx \rho_{0,spt} \exp\left(\frac{\mp Z_\odot}{h_Z}\right) \mathcal{R}(d, l, b), \tag{3.14}$$

where,

$$\mathcal{R}(d, l, b) = \exp\left[-d\left(-\frac{\cos(b)\cos(l)}{h_R} \pm \frac{\sin(b)}{h_Z}\right)\right] \tag{3.15}$$

where $\rho_{0,spt} \exp\left(\frac{\mp Z_\odot}{h_Z}\right)$ is the local volume density of brown dwarfs of spectral type spt, $Z_\odot$ the height of the Sun above the Galactic plane (assumed to be 27 pc), h_Z and h_R the scale height and length of the thin disc as mentioned above, and the sign convention indicates whether the source is above or below the Galactic plane (see e.g. Caballero et al. 2008). The values adopted in this manuscript for $\rho_{0,spt}$ can be found in Table 3.2. Finally, H_{mod} and d are related through:

$$H_{mod} - M_{H_{mod}} = 5\log(d) - 4 + 3.09\ E(B-V) \tag{3.16}$$

where H_{mod} and $M_{H_{mod}}$ are the apparent and absolute magnitudes in the reference band $H_{\rm E}$ and $E(B-V)$ the interstellar reddening, respectively.

Table 3.2 – MLT density at the Galactic plane, and model colour data from Euclid Collaboration: Barnett et al. (2019).

SpT	ρ_0 (pc^{-3})	M_J	$z-Y$	$O-Y$	$Y-J$	$J-H$
M0	2.4×10^{-3}	6.49	0.23	0.55	0.27	0.15
M1	2.7×10^{-3}	7.07	0.29	0.78	0.26	0.08
M2	4.4×10^{-3}	7.71	0.35	1.00	0.23	0.05
M3	7.8×10^{-3}	8.28	0.41	1.23	0.19	0.04
M4	1.0×10^{-2}	8.90	0.46	1.46	0.17	0.04
M5	1.1×10^{-2}	9.53	0.52	1.68	0.16	0.05
M6	7.8×10^{-3}	10.85	0.58	1.91	0.18	0.07
M7	2.2×10^{-3}	11.66	0.82	2.26	0.21	0.10
M8	1.7×10^{-3}	12.08	1.06	2.61	0.26	0.13
M9	1.1×10^{-3}	12.33	1.30	2.96	0.33	0.18
L0	6.7×10^{-4}	12.54	1.54	3.32	0.40	0.23
L1	4.3×10^{-4}	12.79	1.56	3.32	0.47	0.29
L2	3.8×10^{-4}	13.11	1.64	3.42	0.54	0.35
L3	3.6×10^{-4}	13.50	1.72	3.51	0.59	0.41
L4	3.0×10^{-4}	13.93	1.75	3.56	0.63	0.47
L5	4.1×10^{-4}	14.38	1.75	3.55	0.65	0.52
L6	2.2×10^{-4}	14.80	1.72	3.53	0.65	0.56
L7	6.3×10^{-4}	15.17	1.69	3.51	0.63	0.58
L8	3.9×10^{-4}	15.44	1.70	3.54	0.60	0.57
L9	4.8×10^{-4}	15.63	1.76	3.63	0.56	0.53
T0	6.3×10^{-4}	15.72	1.87	3.79	0.52	0.46
T1	6.4×10^{-4}	15.74	2.03	4.00	0.48	0.35
T2	3.6×10^{-4}	15.71	2.22	4.24	0.44	0.21
T3	3.6×10^{-4}	15.69	2.41	4.47	0.41	0.03
T4	5.6×10^{-4}	15.74	2.58	4.64	0.40	-0.16
T5	7.1×10^{-4}	15.93	2.69	4.73	0.40	-0.36
T6	4.0×10^{-4}	16.32	2.75	4.73	0.42	-0.52
T7	2.1×10^{-3}	16.98	2.78	4.72	0.44	-0.64
T8	7.5×10^{-4}	17.95	2.84	4.80	0.46	-0.65

Using the modelling of the high-z QSO population (Sect. 3.3.1) and the MLT population described in this section, we can now derive the ratio of the brown dwarf to high-z QSO density on the *Euclid* footprint. This is shown in Fig. 3.4 for two different apparent magnitudes. The variation of this ratio as a function of apparent magnitude is also demonstrated in Fig. 3.11 for two positions with different galactic coordinates in the *Euclid* footprint.

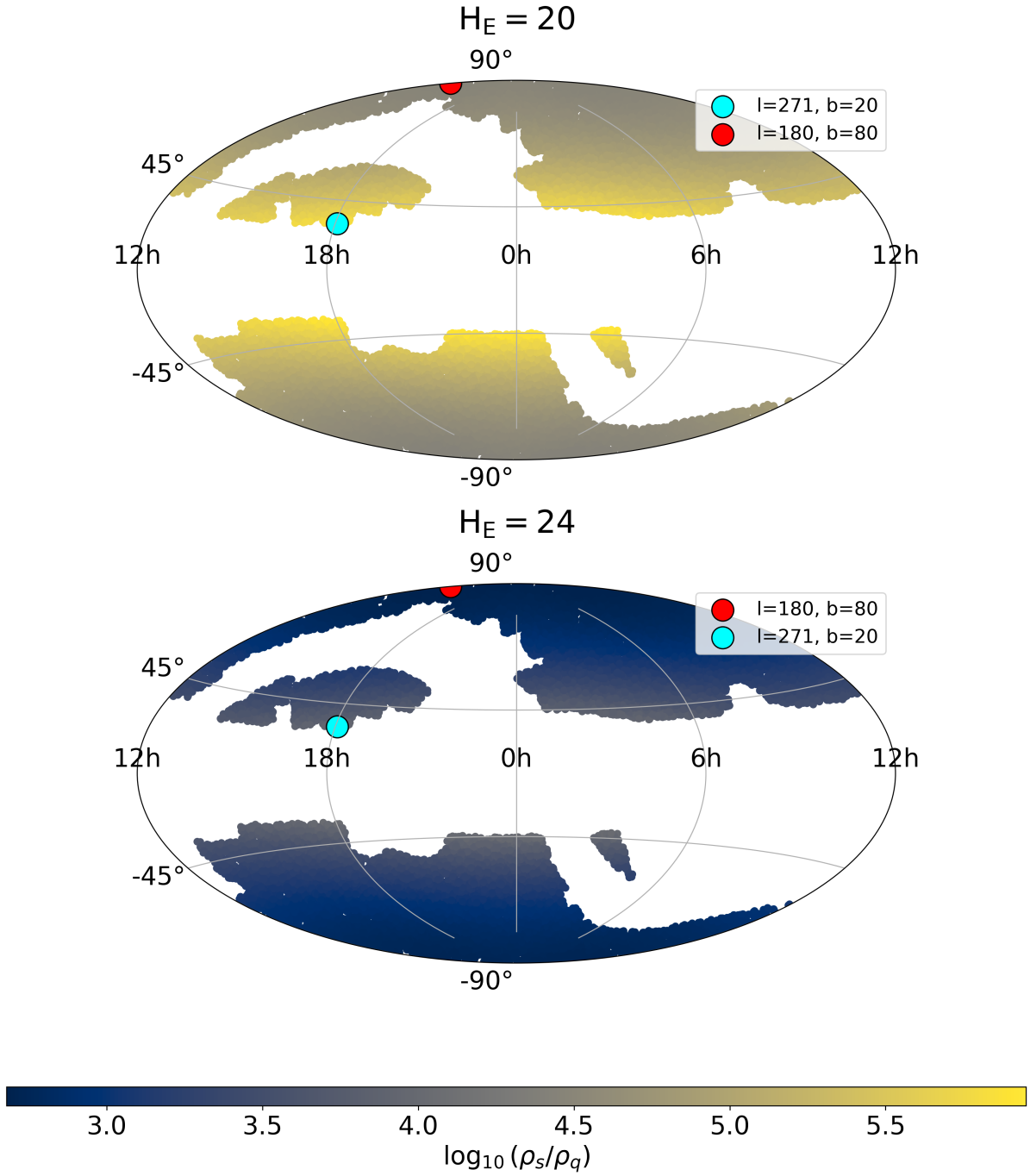


Figure 3.4 – Ratio of the surface density of brown dwarfs of spectral type M6 to T8 to the surface density of $z > 7$ quasars over the *Euclid* footprint shown in Galactic coordinates and galactic projection. Top: for magnitude $H_{\rm E}$= 20. Bottom: for magnitude $H_{\rm E}$= 24.

3.3.2.2 Model templates

For the purpose of our Bayesian analysis, we need to model the spatial distribution of MLT dwarfs, as well as their luminosity and spectral energy distributions from the optical to the near-infrared. For spectral energy distributions, we use the stellar dwarf Luyten Half-Second (LHS) library (Bakos et al. 2002) for M3-M5 brown dwarfs and the brown dwarf spectral library SpeX prism (Burgasser 2014; Burgasser &Splat Development Team 2017) for other spectral types. The use of spectral data enables magnitudes and colours to be correctly determined and modelled in the *Euclid* filters or any other filters, as long as they overlap the spectral range of the spectra. Using spectral data rather than tabulated colours in the literature allows magnitudes and colours to be correctly determined and modelled in *Euclid* or other filters, provided they overlap the spectral range of the spectra. Moreover, there are between 5 and ≈ 50 spectra available for each spectral type, adding spectral diversity to the model.

We also add to our model WISE photometric data for the W1 and W2 bands, anchored to the average JHK photometry of brown dwarfs as tabulated in Best et al. (2017). This allows us to include in our model the WISE photometry that cannot be calculated from the MLT spectra that do not extend beyond the K band. Fig. 4.3 shows the $W1 - W2$ colours for all MLT types as indicated in Best et al. (2017), showing that T dwarfs, especially late types, have very strong $W1 - W2$ colours that can be easily discriminated from the colours of high-redshift quasars.

3.3.3 Contaminant galaxy model

The contamination by galaxies at intermediate redshift poses a significant difficulty in the search for high-redshift quasars. These galaxies can exhibit colours that closely resemble quasars, especially in the near-infrared and optical bands. As a result, when conducting surveys aimed at identifying high-redshift quasars, some galaxies at intermediate redshift can easily be misclassified as quasars, leading to an overestimation of the quasar population and skewing the results of the study (e.g. Hewett et al. 2006).

There are two different types of galaxies susceptible to contaminating the quasar samples because they display red colours that mimic those of high-z quasars, in particular at low SNR: early-type galaxies, on the one hand, and dusty starburst galaxies, on the other hand. The former exhibit an important 4000Å break because an old stellar population dominates them, whereas the presence of dust reddens the latter.

We have carried out two different tasks to assess the extent of this contamination by intermediate-z galaxies. The first one consists of precisely determining the contaminant population using straightforward simulations. The second one is to quantify the contamination in the real universe based on the COSMOS2022 field (Weaver et al. 2022) by computing the Luminosity Function of the contaminant population (hereafter LF). The LF will be used to model the abundance of contaminant galaxies.

3.3.3.1 Understanding the nature of contaminant galaxies

In previous works, the nature of contaminant galaxies is often associated with early-type galaxies without a precise justification. Before adopting this usual assumption, we conducted an exhaustive study to determine the exact nature of the contaminants.

For these needs, we run `Owl-z` on a simulated catalogue containing 10^5 galaxies uniformly sorted in redshift (between $0 \leq z \leq 6$), spectrophotometric templates, and extinction values (between $0 \leq A_V \leq 3$ magnitudes), using the Calzetti et al. (2000) extinction law. The template library includes two different evolutionary synthesis libraries:

— Eight models issued from the Bruzual & Charlot code (Bruzual &Charlot 2003) (hereafter BC03), with Chabrier IMF (Chabrier 2003) and solar metallicity, matching the observed colours of local galaxies from E to Im types (namely a delta burst Single Stellar Population -SSP-, a constant star-forming system, and six τ-models with exponentially decaying SFR),

— 10 Starburst99 templates (Leitherer et al. 1999), including e-lines, for single bursts and constant Star Formation Rate (SFR) models, each one spanning five metallicities (Z =0.04, 0.02, 0.008, 0.004 and 0.001) and 37 ages for the stellar population (between 0 and 1 Gyr).

The apparent magnitudes and associated errors have been calculated using these templates, sorted to uniformly sample the actual range of magnitudes expected for the EWS in the reference H-band, that is $22 \leq H \leq 25$. Photometric error bars and noise are scaled to apparent magnitudes, assuming a Gaussian distribution, according to the expected SNR in the different bands used in the EWS. The probability threshold for the selection as a quasar is set to $P_q = 0.1$ for this exercise, following Mortlock et al. (2012) and Euclid Collaboration: Barnett et al. (2019). The impact of this precise value will be discussed later in Chapter 4.

The initial objective of this exercise is to identify the templates (corresponding to galaxy types) that result in misclassification as quasars. The results are presented in Figure 3.5, where we see which templates contribute the most to the contamination for BC03 and Starburst99 models as a function of redshift. It can be observed that the initial two templates in the Starburst99 population are responsible for the greatest number of erroneous classifications. The majority of contamination (63%) originates from the initial template, while the remaining 27% stems from the second. These two templates correspond to single bursts of star formation with metallicities Z =0.04 (double solar) and solar respectively. In contrast, the remaining templates contribute only a negligible amount of contamination. The BC03 templates exhibit the greatest misclassification rates for the first template, that is the single stellar population model. The contaminant galaxies which are well represented by this template make up 88% of the total sample. For both libraries, Starburst99 and BC03, the primary contamination comes from single stellar population models. Note that the distribution used for this exercise has a homogeneous distribution in redshift, star-formation history, age and reddening. To understand the nature of the contaminant in the true universe, a comparison with the true distribution in a cosmological field is needed, as discussed below.

To gain further insight into the extent of contamination from different galaxy types, we examine the physical parameters of galaxies selected as quasars, with a particular focus

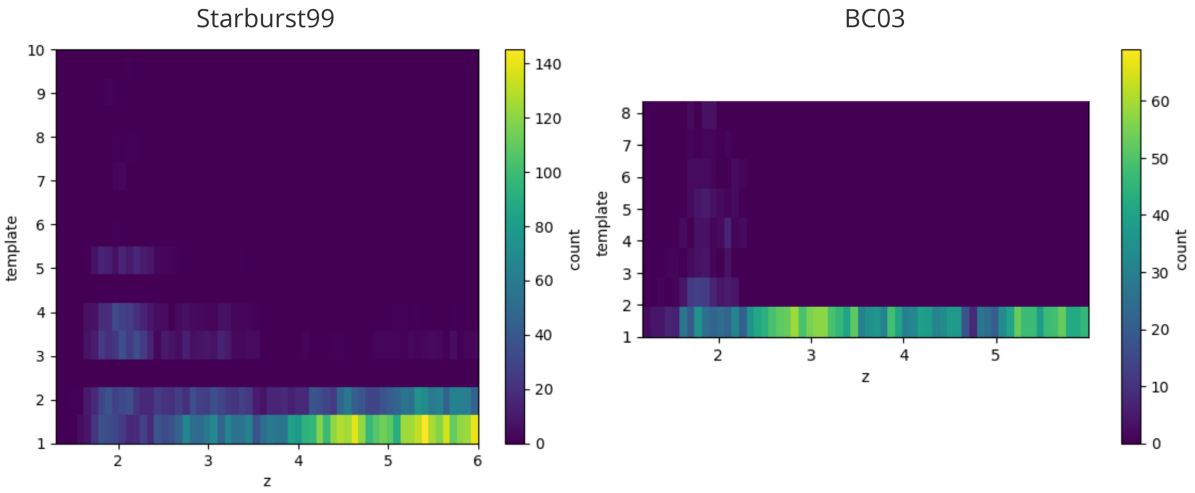


Figure 3.5 – Templates as a function of redshift colour-coded by the density of galaxies that are selected as quasars (i.e. $\mathbb{P}$>0.1). On the left are the Starburst99 template results, and on the right are the BC03 template results.

on the templates for both libraries. We start by plotting a colour-colour diagram and comparing the colours of the selected galaxies with those of the actual COSMOS2020 galaxies as if they were observed by Euclid at the WIDE and DEEP survey depths. As illustrated in Figure 3.6, we observe that there is only a small selection of galaxies (green rectangle) with colours compatible with quasars. We then isolate this selection for both BC03 and Starburst99 libraries.

Following the isolation of this galaxy sample exhibiting realistic and COSMOS-compatible colours (green rectangle in Figure 3.6), the extinction and redshift distributions of these galaxies are presented in order to see which galaxies may potentially contaminate the data set in real observations. It is evident from Figure 3.7 that the galaxies selected for analysis using Starburst99 have realistic colours. However, the only galaxies that survive the colour processing are those at the highest redshifts. Therefore, these galaxies are high-z objects anyway and do not qualify as "contaminants".

In contrast, in Figure 3.8, we show the same results for the BC03 library. Our selection includes only SSP models, as discussed above. They exhibit a distinct redshift distribution, comprising two populations: one at low/intermediate redshift ($1 < z < 2$) and one at high redshift. Using the same considerations as for the Starburst99 library, only the low-redshift sample qualifies as "contaminants" and will be used for the modelling.

In conclusion, these tests show that young Starbursts, represented by the Starburst99 models, only contaminate the sample at low SNR values (dispersed in the colour diagrams) for extremely high values of $A_V \geq 1.5$, spanning a relatively broad domain in redshift but are more present at high redshifts ($z > 4$). Acknowledging the significant influence of SNR values in the reference magnitude on our selection process is important. Indeed, a low SNR in the reference filter H induces a noisy determination of the $J - H$ colour, leading to a higher degree of confusion in the probability calculation and a greater likelihood of misidentifying high-redshift quasars. This general remark affects all contaminant galaxies

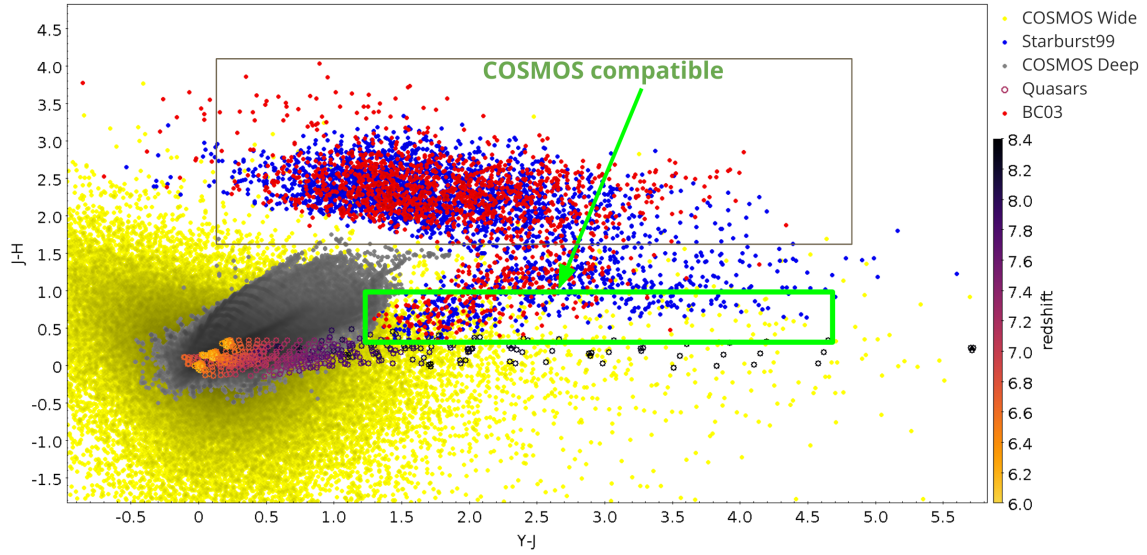


Figure 3.6 – The *Euclid* $Y - J$, $J - H$ colour-colour scatter diagram illustrates the various populations involved. The scatter in yellow represents data from the Simulated COSMOS catalogue, with *Euclid* WIDE depth and *Euclid* filter transmissions. The same data set is presented in grey for comparison purposes, but with *Euclid* DEEP depth. (*red*) BC03 galaxies that satisfy the $\mathrm{SNR_{VIS}} < 3$ criterion and are selected as quasars ($\mathbb{P} > 0.1$). The same selection process was applied to Starburst99 (*blue*), and the resulting data set is shown in blue. The colours have been coded by redshift, and the rectangles in green indicate the Simulated BC03 and Starburst99 galaxies that satisfy the colour selection criteria of quasars. Any galaxy with a $J - H > 1$ has been eliminated in this case.

but is particularly important for dusty starbursts. The main contaminants among this population are galaxies at $z \gtrsim 4$ with $A_V \leq 1.5$, and galaxies at $z \leq 3$ with $A_V \geq 1.5$. In the two cases, such extremely reddened galaxies are relatively rare in the real universe at these redshifts.

For BC03 models, we retain only the SSP model (template 1). It corresponds to the typical stellar population for an early-type galaxy. We can see in the histograms in Figure 3.8 that this population of contaminants is represented by either an old stellar population with low/null reddening at 1<z<2 or a high redshift population at z>5 and high A_V, the two populations having similar colours (degenerate solutions in z/A_V). Early-type galaxies at mid-z are expected to be more abundant in the Euclid Survey than extremely reddened high-z star-burst. In this regard, our results join the previous works in considering that the main contamination comes from early-type galaxies at intermediate redshift $1 < z < 2$ (e.g. Euclid Collaboration: Barnett et al. 2019).

The templates that better represent this contamination by early-type galaxies are SSP (Single Star Population) models with ages above 1 Gyr, at intermediate redshifts, $1 < z < 2$. Indeed, the spectral energy distribution of early-type galaxies, such as elliptical

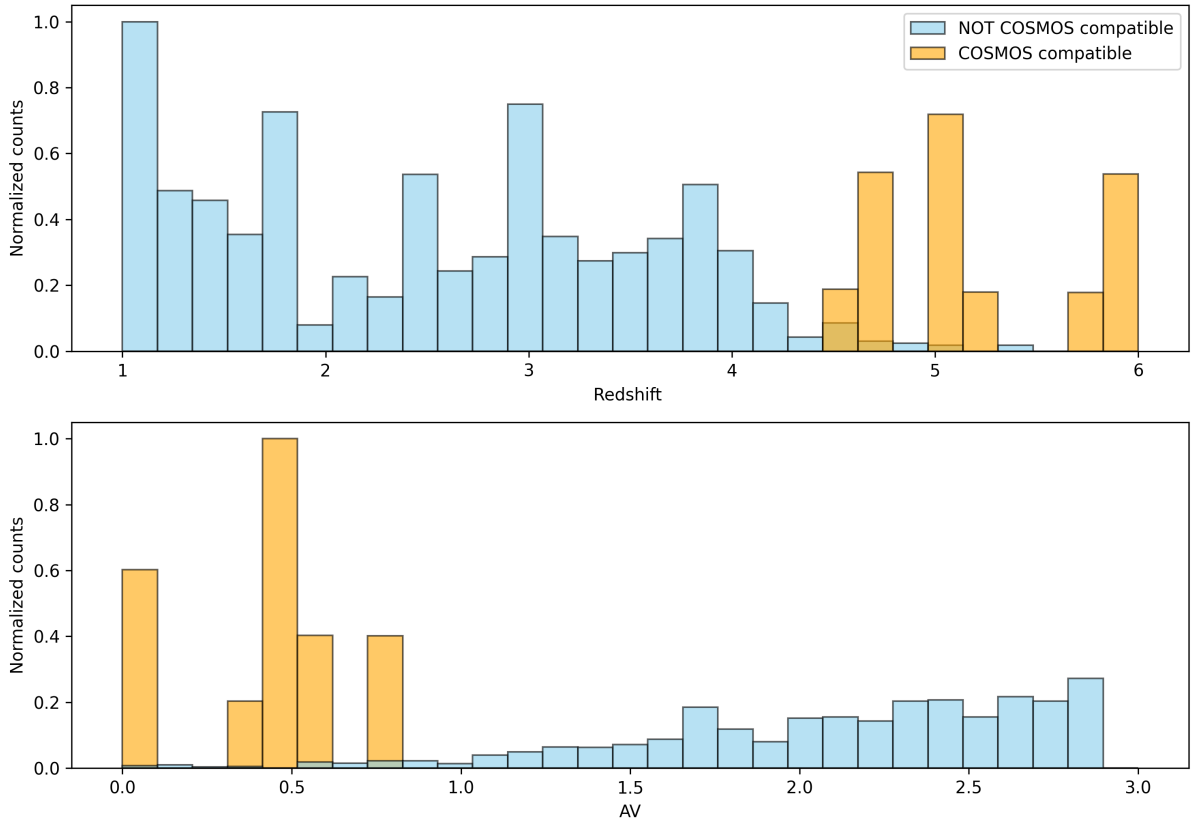


Figure 3.7 – The top panel depicts the distribution of redshift for the Starburst99 first template, with two populations distinguished by their colours. The compatible COSMOS refer to the galaxies with quasar and COSMOS galaxy-like colours. In contrast, the non-COSMOS compatible ones refer to the ones that are unrealistic. The bottom panel presents the same information but for extinction.

and lenticular galaxies, are characterised by an old stellar population, which is well represented by an initial burst of star formation followed by a rapid decline (see, e.g. Ali et al. 2024).

In conclusion, the templates retained to represent the contaminant galaxies are the early-type galaxies from the BC03 library. In the BC03 models (Template 1), this template exists with a large array of ages. We chose 38 templates with different ages (Age>1Gy). We did not vary the extinction as extinction and age are degenerate and have the same effect on reproducing the colours in the Optical and near-infrared domains. We now have a variety of early-type galaxies to include in the model to account for their diversity.

3.3.3.2 Luminosity Function of contaminant galaxies

The abundance of early-type galaxies has been well studied at $z \sim 1.5$, for instance by Zucca et al. (2006), who studied the evolution of the LF for different filters and spectrophotometric types of galaxies. We can use these previous findings to guide our

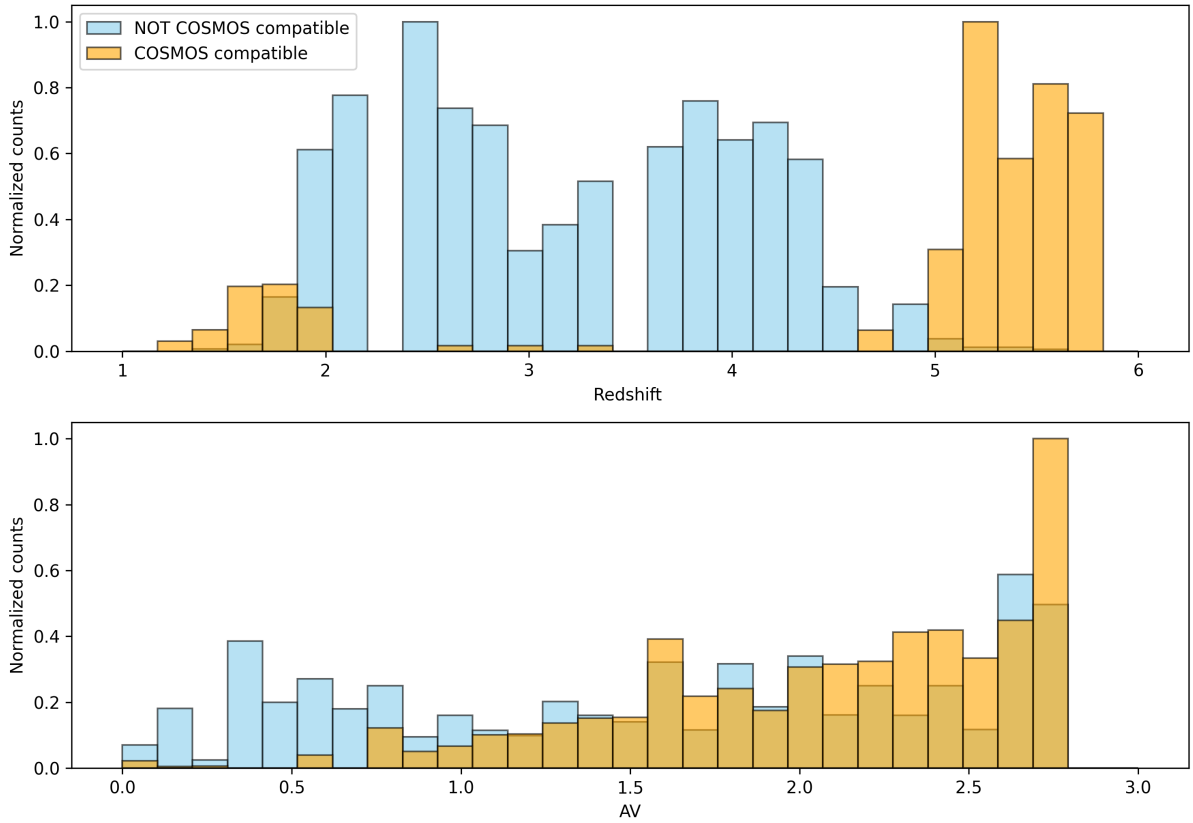


Figure 3.8 – The top panel depicts the distribution of redshift for the BC03 first template, with two populations distinguished by their colours. The compatible COSMOS refer to the galaxies with quasar and COSMOS galaxy-like colours. In contrast, the non-COSMOS compatible ones refer to the ones that are unrealistic. The bottom panel presents the same information but for extinction.

modelling, as shown below, to cover the sensitive magnitude domain.

After identifying the nature of the contaminant galaxies, we determine the LF of this population. We use the Euclidized COSMOS2022 catalogue (Weaver et al. 2022): EL-COSMOS catalogue for this need. This catalogue is described in more detail in Sect. 4.5.4. Note that for galaxies detected in the J-band at intermediate redshift, the relevant filter for the determination of the LF is the rest-frame B.

To represent the LF, we use the classical parameterisation provided by the Schechter (1978) function in terms of magnitude given by:

$$n(M) = (0.4\ \ln 10)\ \phi^*\ [10^{0.4(M^*-M)}]^{\alpha+1} \exp\left[-10^{0.4(M^*-M)}\right]. \tag{3.17}$$

In this function, $n(M)\ dM$ represents the number of galaxies per co-moving Mpc^3 with magnitudes between M and $M + dM$. The parameters are as follows:

— ϕ^*: the normalisation factor, representing the overall volume density of galaxies in units of Mpc$^{-3}\times$ mag^{-1}.

— M^*: the characteristic magnitude, representing the cut-off between the (bright)

luminosity regime dominated by the exponential function and the (faint) regime dominated by the power-law.

— α: the faint-end slope of the luminosity function, describing the distribution of faint galaxies.

I start by running `Owl-z` on the EL-COSMOS catalogue, with only one contamination source: MLT stars, and the same threshold as before for the selection of high-z quasars ($P_q > 0.1$). For consistency, we check the spectrophotometric type and redshift of all galaxies that have been selected as quasars. This test shows that the contamination by galaxies comes majorly and as expected from Early-type galaxies (16 galaxies in total) at intermediate redshift ($z \sim 1.5$).

In order to model the LF for this population of contaminants, we extract from the EL-COSMOS catalogue a subset of early-type galaxies within the sensitive redshift domain $1 < z < 2$. To compute absolute magnitudes, we model the spectral energy distribution with a template well suited to represent this population, namely a short, exponentially decaying model with characteristic star-formation time $\tau = 0.1$ Gyr, age = 10 Gy and solar metallicity ($Z = 0.02$). It is worth mentioning that the precise choice of this model does not change the results. The LF points and their associated error bars are then computed; error bars in the LF include Poisson noise and field-to-field variance added in quadrature. Field-to-field variance is derived using the Trenti &Stiavelli (2008) method and calculator. We then fit the Schechter function in Eq. (3.17) to the LF points using a χ^2 minimisation.

Table 3.3 – Comparison of the Schechter function parameters obtained in this work with the results of Zucca et al. (2006) for Early type galaxies at redshift $1 < z < 1.2$. The value of the slope has been fixed to $\alpha = -1.23$ in this fit.

	α	M^*	ϕ^* $\times 10^{-4}$ Mpc^{-3}
Zucca et al. (2006)	-1.23	$-21.5^{+0.48}_{-0.57}$	$3.0^{+1.4}_{-1.2}$
This work	-1.23	$-21.7^{+0.17}_{-0.17}$	$2.9^{+0.6}_{-0.6}$

Towards the faint end, I applied a magnitude cut to $M_B < -19.7$ because beyond this limit, our sample suffers from incompleteness. The parameters were chosen to enable an easy comparison with the work of Zucca et al. (2006). This motivated us to fix the same value for the parameter α. For comparison, a value of $\alpha = -1.0$ was adopted by Euclid Collaboration: Barnett et al. (2019) to represent the same population. Note that the precise choice of α is irrelevant here because, on the one hand, only the faintest part of the LF is dominated by the power law, and, on the other hand, the population of contaminant galaxies lies preferentially in the bright end.

Fig. 3.9 shows that our LF points agree with Zucca et al. (2006). Our fit parameters are shown in Table 3.3. The LF obtained above is then used to model the prior distribution of early-type galaxies $\rho_{g_i}(H_{\rm mod}, z)$ given in mag$^{-1}\times$ deg$^{-2}\times$ dz^{-1} units by:

$$\rho_g(H, z_g) = \frac{1}{4\pi} \times \frac{dV_c}{dz_g} \times \Phi_g \left[H_{\rm mod} - \mu - K_{\rm corr}(z_g), z_g\right] \tag{3.18}$$

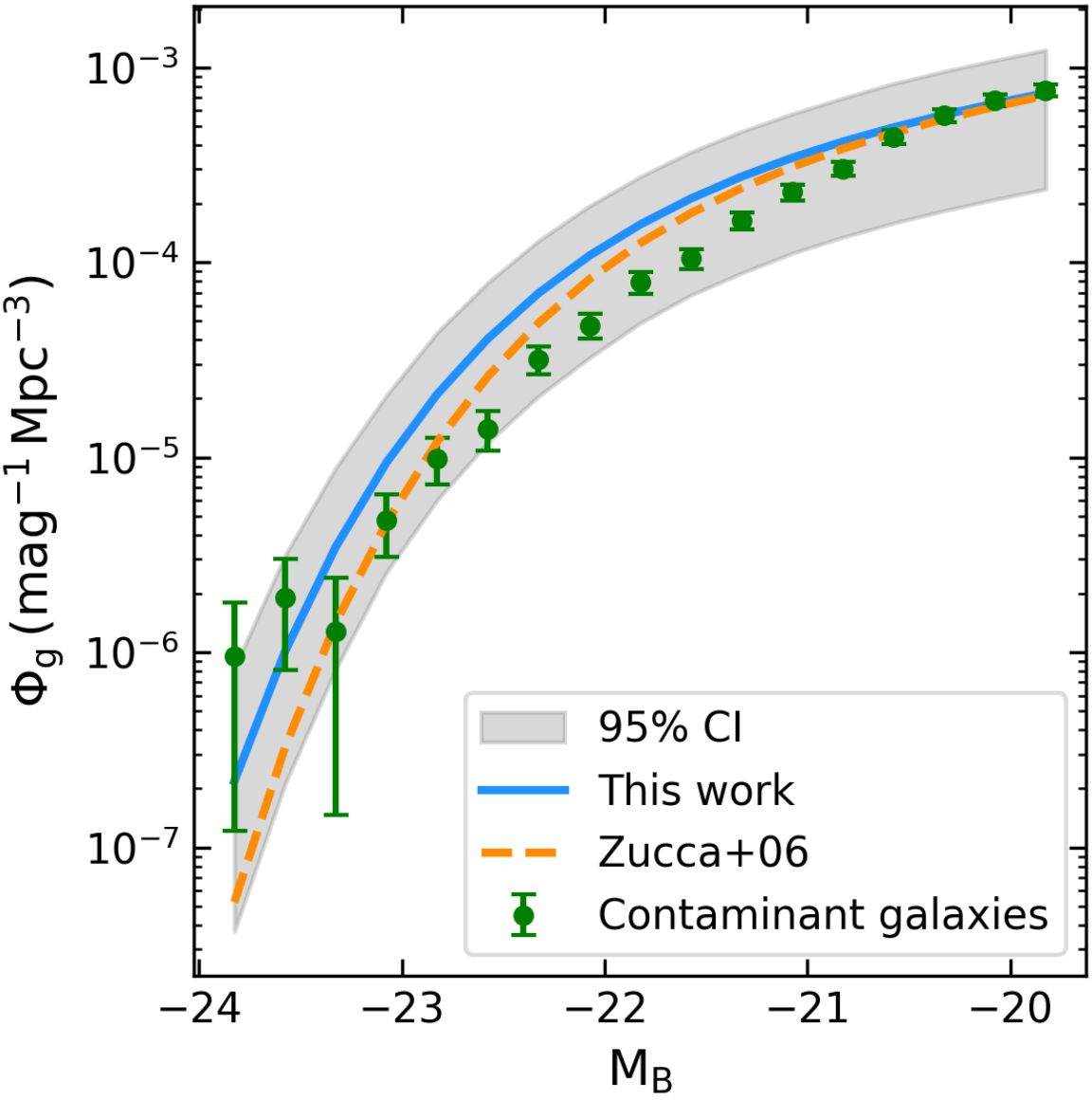


Figure 3.9 – Luminosity function fit in the B-band (blue) in addition to its 95% confidence interval (grey) for the population of early-type galaxies at $1 < z < 2$ in the E-COSMOS catalogue data (green), compared to the early-type galaxies (dashed orange) luminosity function fit by Zucca et al. 2006.

where z_g is the redshift range of the contaminant galaxies $1 < z < 2$ and Φ_g is the luminosity function fit obtained in this work.

For a given set of parameters $\{z_g, H_{\mathrm{mod}}\}$ and an early-type galaxy model g_i, the weighted evidence can be quantified as:

$$W_{g_i} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho_{g_i}(H_{\mathrm{mod}}, z_g) P(\mathrm{bands}, H_{\mathrm{obs}} | H_{\mathrm{obs}}, g_i) dH_{\mathrm{mod}} dz_g. \tag{3.19}$$

3.3.4 Unbalanced prior problem

In the context of Bayesian selection, where contamination is the dominant factor in terms of numerical abundance (see figures 3.10 and 3.11), it is of paramount importance to give careful consideration to the priors involved. In such cases, the prior probabilities frequently reflect surface densities, which represent the spatial or numerical distribution of both the contaminant and the target populations. In light of the fact that the contaminants outnumber the targets, the prior associated with the contaminants will be considerably higher due to their larger surface density. This imbalance presents a challenge in detecting true targets, as their smaller surface densities translate into lower prior probabilities, initially making them less likely to be identified. A Bayesian approach addresses this issue by updating the priors with observed data (impossible in our case), thereby adjusting the posterior probabilities to identify true targets despite their rarity. However, this process requires precise modelling of the likelihoods to ensure that the data can effectively distinguish the few targets from the dominant contamination. This allows the Bayesian framework to provide a statistically sound mechanism for selecting targets, even when contaminants significantly outnumber their presence.

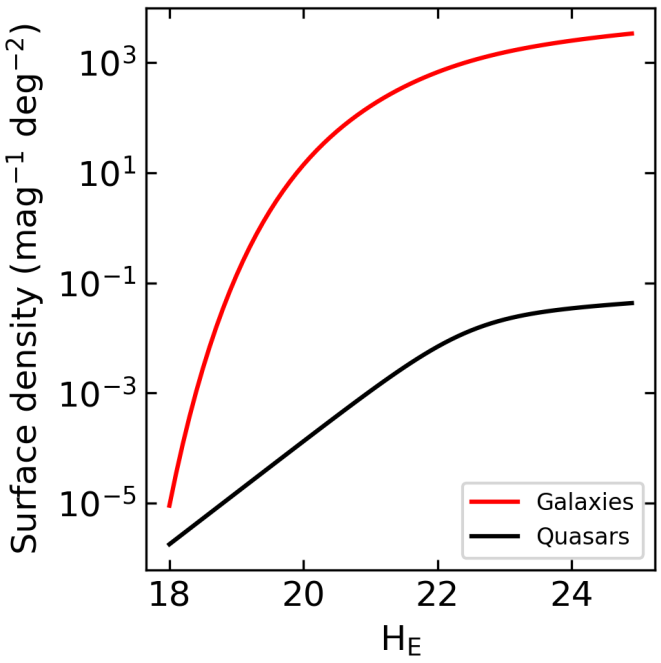


Figure 3.10 – Surface densities comparison between galaxies and quasars.

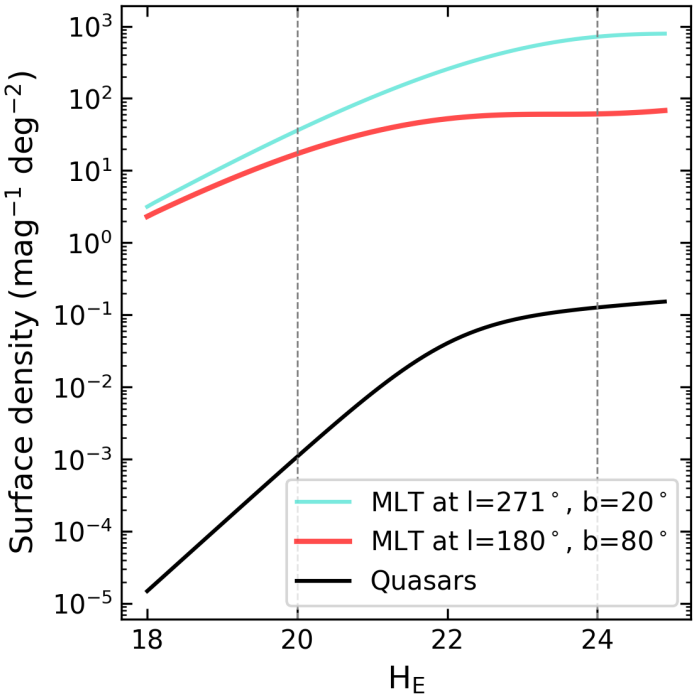


Figure 3.11 – Ratio of MLT to $z > 7$ quasar density as a function of apparent magnitude in two different locations of the *Euclid* footprint.

4 `Owl-z`: A flexible tool to select high redshift quasars

Contents

4.1 `Owl-z`: Ingredients and outputs . . . 83
4.2 `Owl-z`: how it works . . . 85
4.2.1 Configuration . . . 85
4.2.2 Model building . . . 86
4.3 Performance estimators . . . 86
4.3.1 Completeness . . . 86
4.3.2 Purity . . . 87
4.3.3 F-measure . . . 87
4.4 Re-identifying known $z > 7$ quasars . . . 88
4.5 Performance of `Owl-z` based on simulated catalogues . . . 89
4.5.1 Data . . . 89
4.5.1.1 Completeness catalogues . . . 89
4.5.1.2 Purity catalogues . . . 91
4.5.2 Completeness . . . 92
4.5.3 Redshift estimation . . . 95
4.5.4 Purity and F-measure . . . 96
4.5.4.1 Global performance . . . 97
4.6 Discussion . . . 98
4.6.1 Influence of the selection threshold . . . 98
4.6.2 Optimising the identification of $z > 7$ quasars . . . 100
4.6.3 Influence of the position in the EWS footprint . . . 101
4.6.4 Influence of thick disc MLT . . . 103
4.6.5 Comparison with previous works . . . 103
4.7 Conclusions . . . 104

The main objective of this chapter is to introduce the `Owl-z` tool. This Bayesian comparison model code builds models of the populations to be classified. Then, it compares them using Bayesian processes with photometry, giving the probability that each object belongs to a different population. In this case, the code is applied to the selection of z>7 quasars in the EWS (see Chapter 2), using the modelling presented in Chapter 3 for both the population of high-z quasars and the contaminants. The goal is not only to describe the code but also to assess its performance in this particular representative case.

For these needs, the chapter starts with a brief description of the ingredients used by the code and the output provided. As part of the validation, `Owl-z` was able to reidentify all spectroscopically confirmed quasars at $z \geq 7$, demonstrating the code's versatility regarding its application to photometric catalogues other than the EWS. The performance of `Owl-z`, based on a metric combining completeness and purity called F-measure, is analysed in the case of the EWS using simulated catalogues in a wide range of redshifts ($7 \leq z \leq 12$) and H-band *Euclid* magnitudes ($18 \leq H_{\rm E} \leq 24.5$). This chapter presents and discusses the results obtained by `Owl-z` and its expected performance in selecting $z \geq 7$ quasars in the EWS.

4.1 `Owl-z`: Ingredients and outputs



Figure 4.1 – `Owl-z` logo

The ingredients needed to run `Owl-z` can be divided into two principal categories: those pertaining to modelling the targets to be studied and those pertaining to the priors used in the Bayesian classification.

The first category includes the information needed to generate the model fluxes for the libraries of templates used to simulate the target data, in this case, high-z quasars, MLT stars and contaminant galaxies. This category includes:

— Template libraries, providing the Spectral Energy Distributions (hereafter SED) of the different astrophysical objects. They could be theoretical libraries (as for the BC2003 library used in Chapter 3), or empirical libraries. Although other formats can be used and adapted for `Owl-z`, in this thesis, we have used ASCII files containing wavelengths (in Angstroms, Å) versus fluxes in F_λ units (i.e., F_λ [erg s^{-1} cm^{-2} Å^{-1}].

— Filter transmissions. The current implementation contains ASCII files containing the filter transmissions in % of photons versus wavelength.
— The parameter space covered by the modelling: magnitudes, redshifts, types...
— The reference magnitude.
— The survey depth, given by the limiting magnitudes in all the photometric bands.

The ingredients needed to define the priors are related to the computation of the surface densities for the different populations, in this case, high-z quasars, MLT stars, and contaminant galaxies, as explained in Chapter 3. To calculate the surface densities of stars, it is necessary to derive the local densities for all spectral types; we use the ones in table 3.2 and the absolute magnitude provided. The model magnitude in the reference filter is given a vector value for the purpose of applying the model to Euclid data; for example, $18 < H_{\rm E} < 25$. Finally, the prior surface densities for stars depend on galactic longitude and latitude; therefore, the target to be analysed using `Owl-z` must include this information. In the case of extra-galactic sources (quasars and galaxies), only the model magnitude and redshift (model parameters) are required.

Input catalogues to be explored using `Owl-z` exclusively include fluxes and associated errors, along with the coordinates utilised in the prior for the stars. The fluxes and errors on fluxes are currently given in F_ν units, i.e., F_ν [erg s^{-1} cm^{-2} Hz^{-1}]. The coordinates are galactic. A crude colour or magnitude selection prior to running `Owl-z` is highly recommended to reduce the size of the input catalogue while keeping only the sources that have a chance to belong to the target category, in this case, z>7 quasars.

`Owl-z` outputs an ASCII file containing the following information for each object in the input catalogue:
— A list of probabilities of belonging to each class. The user then interprets the data and decides on the selection.
— A list of helpful output parameters, including the optimal redshift for galaxies and quasars and the optimal spectral type for the best-fit stellar model.
— The SEDs of the galaxies and quasars that yield the MAP (see Chapter 3).

4.2 `Owl-z`: how it works

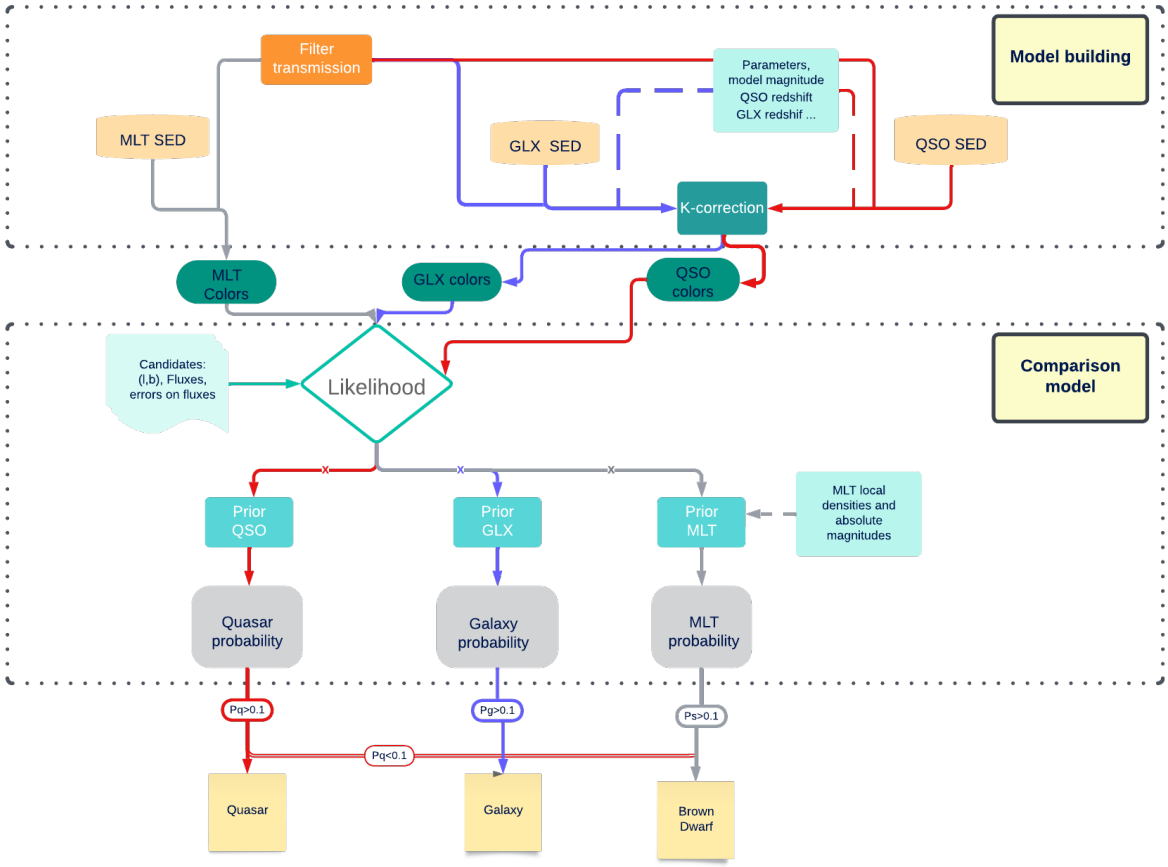


Figure 4.2 – Flowchart describing the `Owl-z` Bayesian model comparison code.

The flowchart of `Owl-z` is given in Figure 4.2. The general use of `Owl-z` can be done in two main steps. The first step is where the photometric models are built, and the second is where these models are used to compare using a likelihood and multiplied by the priors to calculate the probability. However, a configuration file must be carefully filled in for correct use.

4.2.1 Configuration

The configuration file provided is simple (YAML format), asking the user to specify the desired paths for all input SEDs. The package already contains template libraries providing SEDs for all the populations, but more could be added, provided they are in the correct format. The configuration files allow the user to choose a redshift range for the quasars, another for the galaxies, and a magnitude (observed) range. This magnitude will be used as a model magnitude and corresponds to the magnitude in the chosen reference filter (the user also gives this information). The only requirement for the reference magnitude choice is that, in the quasar redshift range to explore, all quasar candidates have a detection in

this reference magnitude. The user is also required to provide a list of filters; again, an extensive library of filters is provided in the correct format. An example of a configuration file looks like the one provided in Annex B.

4.2.2 Model building

In this step, `Owl-z` uses the SEDs provided by the template libraries for each population to calculate the magnitudes in all the specified filters. For clarity, I will explain a use case for the Euclid filters $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$.

For quasars and galaxies, `Owl-z` stars by reading the model parameters for the given (quasar or galaxy) model: the redshift range and the reference magnitude. It then proceeds to take each SED, applying the redshift, taking into account the K-correction, and for each redshift value, it uses the filter transmissions to calculate the magnitudes using the equation 0.9. It then calculated the colours using the reference magnitude such as :

$$\text{colour}_i = m_{ref} - m_i \tag{4.1}$$

where m_{ref} is the magnitude in the reference filter and m_i is the magnitude in a given band, that is $i = (1, 2, 3...N_{\rm bands} - 1)$ where $N_{\rm bands}$ is the number of filter transmissions used. In total, $N_{\rm bands} - 1$ colours are calculated. Finally, the colours are stored in one file per population for each redshift for each model in an ASCII file to be used in the next step.

`Owl-z` also calculates its own K-corrections and absolute magnitudes at specific wavelengths (1450Å for quasars and 1700Å for galaxies), which are used in the calculations of the priors for the extragalactic sources: galaxies and quasars. These values are calculated for each SED and redshift and stored in ASCII files for each of the two populations.

In the case of MLT stars, models are built using only the SEDs in the template library by directly integrating and calculating the colours. They are then stored in the same way as quasars and galaxies.

4.3 Performance estimators

In classification tasks, it is important to evaluate a model's performance using appropriate metrics to understand how well it makes predictions. Two important concepts are completeness and purity, which can be combined into the F-measure for a balanced assessment of performance. These metrics provide a more comprehensive view than simply using accuracy, especially when dealing with imbalanced data.

4.3.1 Completeness

In classification, completeness is analogous to recall. It measures the proportion of actual positive instances that the model correctly predicted. Completeness focuses on how much of the relevant data (i.e. true positives) the classifier has captured. Mathematically,

completeness (or recall) is defined as:

$$C = \frac{TP}{TP + FN}. \tag{4.2}$$

Where:
— TP is the number of true positives (correctly predicted positive cases).
— FN is the number of false negatives (actual positives that were incorrectly predicted as negatives).
A high completeness score indicates that the classifier is good at identifying most of the actual positive cases, but it may miss precision if it also predicts many false positives.

4.3.2 Purity

In classification, purity corresponds to precision. It reflects how many of the instances predicted as positive by the model were positive. Purity tells us the quality of the positive predictions the model made.

Mathematically, purity (or precision) is defined as:

$$P = \frac{1 + TP}{1 + TP + FP} \tag{4.3}$$

Where FP is the number of false positives (instances incorrectly classified as positive).

A high purity score indicates that it is likely to be correct when the model predicts something as positive, though it might miss out on some positives (lower completeness). Using this definition of purity, purity of 100% can either mean that all quasars have been correctly classified with zero contamination or that no source has been classified as a quasar; since purity and completeness give us two necessary pieces of information. We have chosen to combine purity and completeness in the F-measure as it removes the confusion from the previous definition of purity.

4.3.3 F-measure

To balance completeness and purity, the F-measure (the F1 score) combines both metrics into a single score. It is the harmonic mean of precision and recall, ensuring that both are considered equally important.

The F-measure is defined as:

$$F = 2\,\frac{CP}{C + P} \tag{4.4}$$

Where:
— C is completeness (recall),
— P is purity (precision).
The F-measure provides a balanced assessment of the classifier's performance, especially when there's an imbalance between positive and negative cases. It penalises the model

when either purity or completeness is low, making it useful for cases where both false positives and false negatives are critical. The F-measure is a statistical measurement usually employed to assess the performance of a classifier (Baeza-Yates &Ribeiro-Neto 2011).

The interpretation of F-Measure reads as follows:

— F = 1: This is the best possible value, indicating perfect precision and recall.

— F = 0: This is the worst value, meaning the model failed to make correct predictions.

A high value of F-measure suggests a well-performing classifier that balances completeness and purity, making it effective for real-world tasks with imbalanced data. This is a perfect metric for quasar selection.

4.4 Re-identifying known $z > 7$ quasars

As part of the validation process, I analysed the performance of my code in recovering all known quasars at $z > 7$. `Owl-z`'s versatility enables its application to different optical and near-infrared data sets used in the discovery of these quasars. The objective here is to determine if `Owl-z` can identify them as quasars and compare their spectroscopic redshift with the redshift obtained by the code.

Table 4.1 lists the known quasars, the photometric bands used for their discovery, their measured spectroscopic redshifts, and the depth reached in a reference photometric band. The photometric data include data from the Dark Energy Camera Legacy Survey (DECaLS; Dey et al. 2019), the UKIRT Infrared Deep Sky Survey (UKIDSS; Warren et al. 2007), and the Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010a). Given the shallow depth and non-constraining photometry of the WISE bands $W3$ and $W4$, I only use photometric data from bands W1 and W2 in the following.

Table 4.1 – Known spectroscopically confirmed quasars at $z > 7$. The photometric data used for their discovery and spectroscopic redshift ($z_{\rm spec}$) are indicated. The columns $z_{\rm out}$ and P_q refer to the redshift and probability, respectively, of being a quasar calculated by `Owl-z`.

QSO name	Photometry	$z_{\rm spec}$	$z_{\rm out}$	Depth in ref band	Reference
J1342+0928	$z_{\rm DECam}$, Y,J,H,K, $W1$,$W2$	7.54	7.8	J 20.52 5σ	Bañados et al. (2018)
J1120+0641	$z_{\rm DECam}$, Y,J,H,K,$W1$,$W2$	7.08	7.05	J 20.52 5σ	Mortlock et al. (2011)
J0313-1806	$z_{\rm PS1}$,$y_{\rm PS1}$,J,K	7.64	7.7	J 20.52 5σ	Wang et al. (2021)
J0038–1527	$z_{\rm DECam}$,$y_{\rm PS1}$, Y,J,H,K, $W1$,$W2$	7.02	6.95	J 20.52 5σ	Wang et al. (2018)
J1007+2115	$z_{\rm DECam}$,$y_{\rm PS1}$, Y,J,H,K, $W1$,$W2$	7.51	7.65	J 20.52 5σ	Yang et al. (2020)
J0252–0503	$z_{\rm DECam}$, Y, J, $W1$, $W2$	7.02	6.9	J 20.52 5σ	Yang et al. (2019)
J1243+0100	$g_{\rm HSC}$, $r_{\rm HSC}$, $i_{\rm HSC}$, $z_{\rm HSC}$, $y_{\rm HSC}$	7.07	9.9	$y_{\rm HSC}$ 24.4 5σ	Matsuoka et al. (2019b)
J2356+0017	$i_{\rm HSC}$, $z_{\rm HSC}$, $y_{\rm HSC}$	7.01	6.75	$y_{\rm HSC}$ 24.4 5σ	Matsuoka et al. (2019a)

The pre-selection methods used to select these quasars were mainly based on colour selection and the identification of strong Lyman breaks ($z - J > 4$) around $\sim 1\mu$m. An additional criterion was used for the five cases out of eight where WISE photometry was available, consisting of a $W1 - W2 < 0.7$ colour selection enabling the rejection of contamination by T-dwarfs (see Fig. 4.3).(see figure 4.3).

The results of `Owl-z` are also presented in Table 4.1, particularly the output redshifts and P_q values returned by the code. The primary potential contaminants expected for the

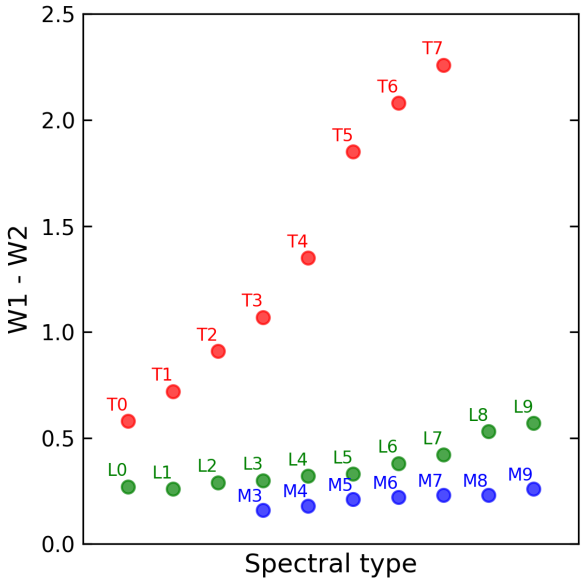


Figure 4.3 – WISE W1 − W2 colours of MLT dwarfs from Best et al. (2017).

QSOs in this study are L brown dwarfs. `Owl-z` yields a probability $P_q > 0.99$ for all of the aforementioned QSOs, showing the high performance of the code in re-identifying the whole sample. Fig. 4.4 shows the photometric data, including error bars, for all quasars and the best-fit quasar and MLT dwarf solutions returned by the code.

Some results deserve a specific comment. `Owl-z` has provided an excellent result for all quasars, including those detected in a single band with non-detection constraints in the others (J2356+0017 and J1243+0100) because the photometric bands in HSC are very deep, and the Lyα break much easier to detect. These quasars have broad $Ly\alpha$ emission and are low-luminosity quasars, matching well the properties of the quasar population model (Sect. 3.3.1), unlike the SEDs of the MLT dwarfs (Sect. 3.3.2). As expected, the best-fit MLT solutions for quasars with WISE photometry are late M or L dwarfs, given the aforementioned colour selection criterion ($W1 - W2 < 0.7$) that excludes T dwarfs from photometric selection (Fig. 4.3). Conversely, best-fit solutions by T-dwarfs are more likely when WISE data are not available, as is the case for J1120+0641. WISE data proved crucial for J0252-0503, without which there would have been only one detection.

4.5 Performance of `Owl-z` based on simulated catalogues

4.5.1 Data

4.5.1.1 Completeness catalogues

To estimate the completeness of the `Owl-z` selection method, we create a mock catalogue of 500,000 quasars. Three different scenarios have been developed to gain a deeper

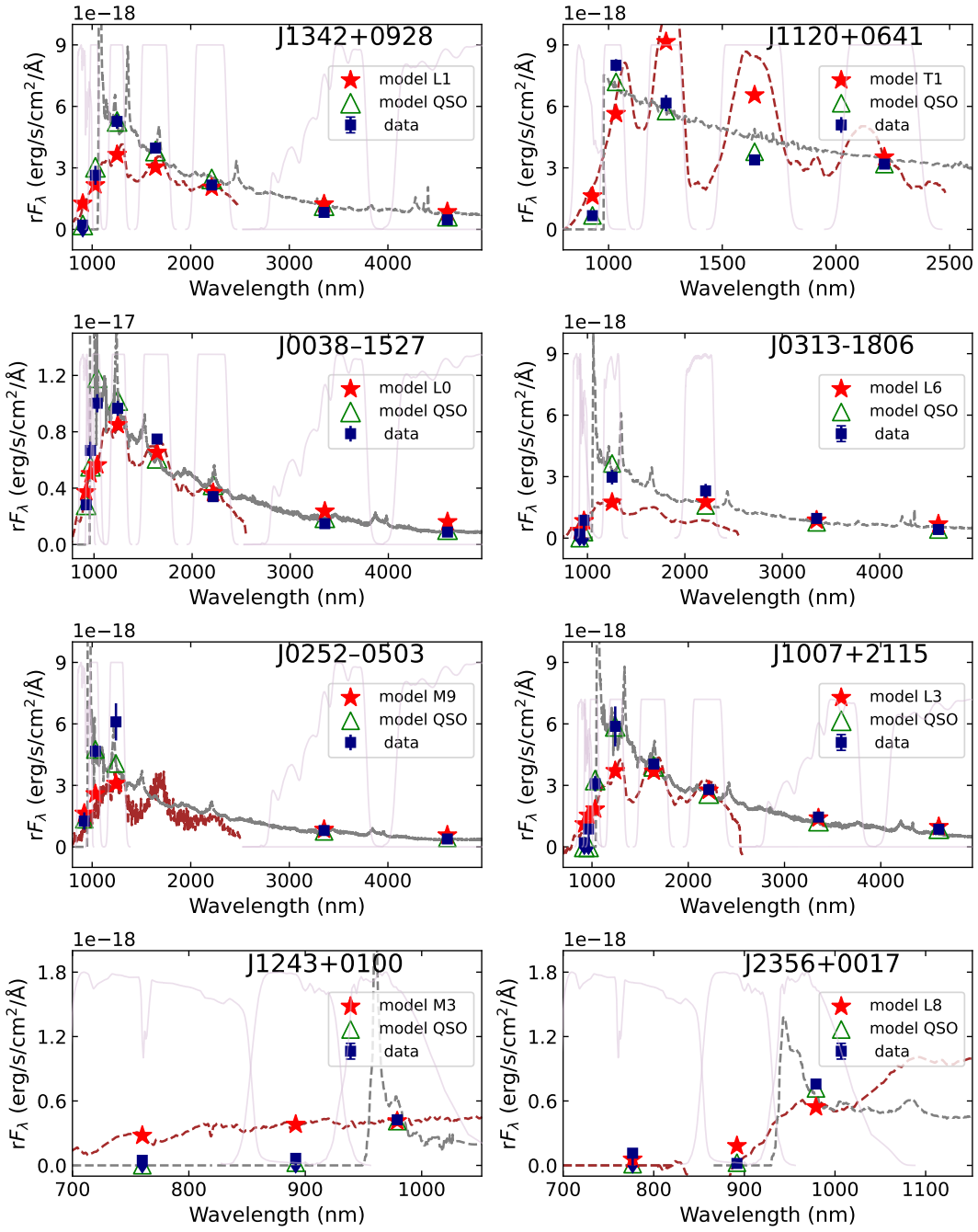


Figure 4.4 – Photometric data for the eight spectroscopically confirmed quasars at $z > 7$ and the `Owl-z` best-fit solutions for quasars and MLT dwarfs. Blue squares represent the original photometric data points. The best-fit quasar model is in dashed grey, the best-fit model for an MLT dwarf is in dashed brown, and the filter transmission curves are shown in solid light purple. Model photometry for best-fit solutions is represented by green triangles for quasars and red stars for MLT dwarfs.

understanding of the impact of using only Euclid data and of replacing or supplementing the existing data with other bands. The first scenario involves using only Euclid data. In the second scenario, the optical band $I_{\rm E}$ is replaced by the optical z-band of LSST. In the third scenario, in addition to the Euclid optical and NIR data, the photometry is extended to the WISE coverage in the W1 and W2 bands (see Table4.4). The sensitivities of these additional bands can also be found in Table 4.3 The completeness is calculated on the EWS DR6 (Euclid Collaboration: Scaramella et al. 2022) footprint, which covers 15,000 deg^2. Quasars are randomly sorted with a flat distribution in redshift in the range $7 \leq z \leq 12$, a flat distribution in absolute magnitude M_{1450} in the range $-29 \leq M_{1450} \leq -22$, a flat distribution over all the quasar templates in our library, and finally, a flat spatial distribution within the 15,000 deg^2 of the EWS footprint. For each object in the catalogue, we calculate apparent magnitudes in the *Euclid* bands, in the LSST z band, and in the WISE W_1 and W_2 bands. For photometric errors, we adopt those corresponding to the DR1 release of the LSST data (Ivezić et al. 2019b) and those for the WISE data; we use the W1 and W2 bands and their corresponding errors from (`wise2.ipac.caltech.edu`). For *Euclid* photometric errors, we use the signal-to-noise ratio maps across the EWS footprint as described in Euclid Collaboration: Scaramella et al. 2022; all of this information can be found in Table 4.3.

Table 4.2 – Scenarios explored in the simulations for the completeness estimation.

Survey	**Depth in NIR**	**Depth in Optical**
Euclid	YJH* 24	O* 24.5
Euclid +LSST	YJH* 24	z(5 σ)
Euclid +WISE	YJH* 24 W1 19.8 W2 18.9 (5σ)	O* 24.5

* The SNR at the reference magnitude is variable, calculated from the Zodiacal light maps as presented in Euclid Collaboration: Scaramella et al. (2022).

4.5.1.2 Purity catalogues

To estimate the purity, we need to simulate the content of the input photometric catalogues, including the contaminant populations of stars and galaxies, as accurately as possible, in a realistic way. Bright high-z quasars are rare, according to the current LF. For this reason, we force the inclusion of such sources in the simulated catalogues, as explained below, to determine how difficult it will be to identify and study these sources if they ever exist. In order to achieve this, the simulations are conducted on a surface that is 10 times larger than that of the EWS. Subsequently, scaling back to the original surface (1000 deg^2) is performed, whereby only those sources that the LF predicts are accounted for. Note that the definition of purity as given in Sect. 4.3accounts for the possible detection of these unique, bright and rare quasars.

The COSMOS2020 field (Weaver et al. 2022) represents an ideal base for this exercise, given the extended wavelength coverage and the excellent quality of the photometric

redshifts that can be used as spectroscopic redshifts for our needs. All sources in this field have been fitted using LePHARE (Arnouts &Ilbert 2011) and a library of reference templates, and this provided us with a best-fit template and a redshift. We use these results to compute the *Euclid* Optical and NIR photometry via a process we call "Euclidization" (additional details will be provided in *Euclid* preparation, EC & Garnett et al.). This process has also allowed us to simulate the SNR and associated photometric errors in each filter. This new *Euclid*-like COSMOS catalogue will be called hereafter E-COSMOS.

The original COSMOS catalogue covers 2 deg^2. Multiple realisations of the parent E-COSMOS catalogue are needed to cover a statistically significant field of view. This is done by randomly sorting the coordinates of the field centre within the EWS footprint in such a way that a full (not overlapping) large catalogue is created. This catalogue covers 1000 deg^2 separated into six different areas. The rationale behind this choice is that it represents the smallest surface area over which we can account for a sufficient number of quasars. Furthermore, this surface has been divided across the region of interest to accommodate the various scenarios in galactic latitude/longitude and to account for the inevitable differences in resolution, however slight they may be. Fig. 4.5 and Table 4.5 display the location of these areas around the *Euclid* Footprint. It is worth noting that the location of the field is expected to have a direct impact on the performance, given the different distribution of contaminant stars on the one hand and the different SNR achieved across the EWS due to zodiacal light variations on the other hand (Euclid Collaboration: Scaramella et al. 2022). These effects are discussed below. The E-COSMOS catalogue has been cleansed of all stellar objects and is situated far from the Galactic plane. To properly account for the presence of contaminant stars, we randomly inject additional MLT dwarfs into the catalogues, following the distribution presented in Sect. 3.3.2, reproducing the number densities expected at the galactic coordinates where the simulated field is located. Since E-COSMOS does not contain any spectroscopically identified high-redshift quasar, this population has been randomly injected in the simulated fields, following the prescriptions presented in Sect. 4.5.2 to account for the SNR in the EWS properly. High-z quasars are randomly sorted according to their LF. Obviously, the number of bright $z > 7$ quasars is expected to be very small in the EWS field. In particular, the current LF does not predict any quasar brighter than $M_{1450} < -23$ in the entire field. If these rare bright objects exist, we force the measurement of purity as explained above. To better capture this population, 10 realisations of the entire *Euclid* footprint are performed, allowing us to retrieve some quasars towards the brightest luminosities and up to redshifts $z > 10$. However, in order to facilitate consistent comparison of the statistical performance estimators C and P, it is necessary to normalise all populations to the same surface area (15,000 deg^2).

4.5.2 Completeness

The performance of `Owl-z` on the EWS is estimated in terms of completeness and purity in order to ascertain the sensitivity to confusion and contamination by MLT dwarfs and early-type galaxies. To this end, a series of catalogues is simulated in the following sections using *Euclid* filters (optical $I_{\rm E}$ and NIR $Y_{\rm E}J_{\rm E}H_{\rm E}$) and sensitivities detailed in Table 4.3.

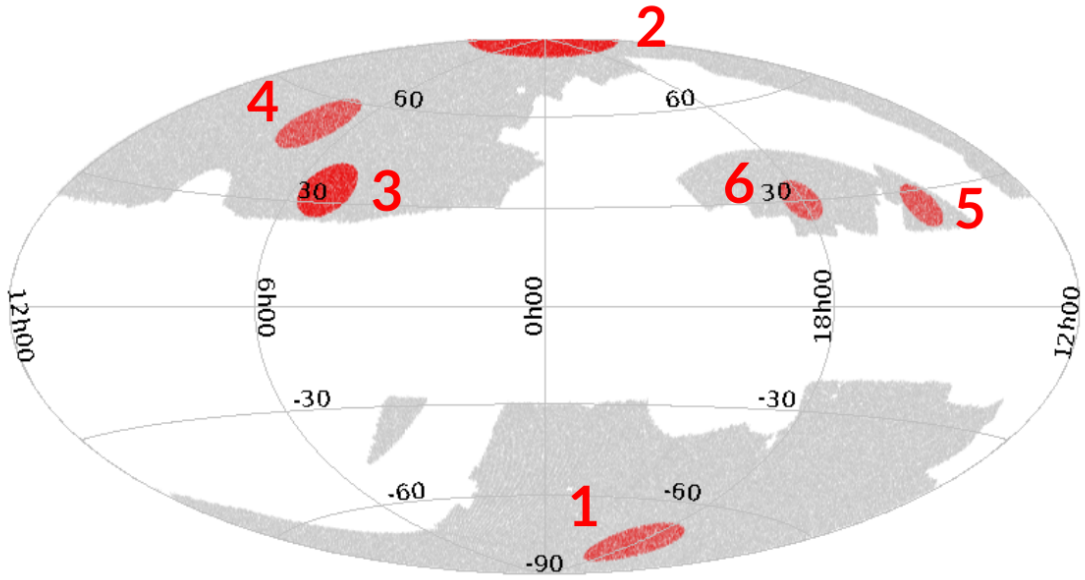


Figure 4.5 – (*red*) The location of the regions of focus used in the computation of the purity how in Galactic coordinates and galactic projection. Details on these surfaces can be found in Table 4.5. (*grey*) the EWS DR6 footprint.

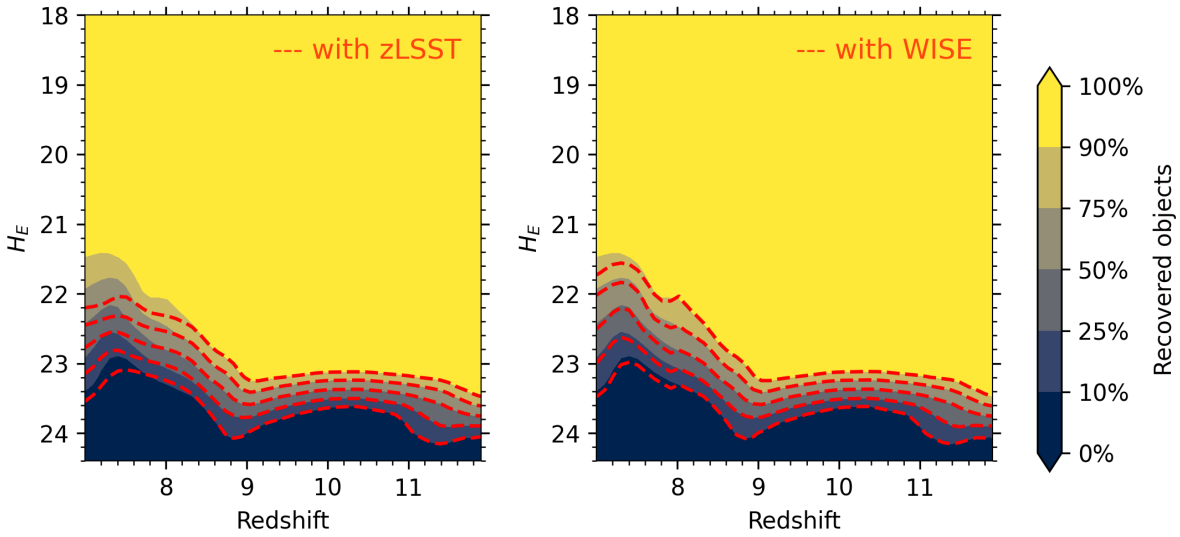


Figure 4.6 – The selection functions of high redshift quasars determined using `Owl-z` for *Euclid* $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ data. The selection function designs the level of completeness per redshift bin; the completeness of a bin is calculated by counting the percentage of quasars in the bin with a probability of $\mathbb{P} > 0.1$. Several cases are shown: *(left)* In red, the optical band O of *Euclid* is replaced by the z-band from LSST. *(right) Euclid* $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ data and in red *Euclid* $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ in addition to WISE (W1 and W2 bands).

Table 4.3 – Photometric bands and depth. The *Euclid* bands characteristics are a duplicate of Table 7 in Euclid Collaboration: Scaramella et al. 2022. Statistics are provided for each channel: optical $I_{\rm E}$ and LSST z-band and NIR $Y_{\rm E}$, $J_{\rm E}$, $H_{\rm E}$, and WISE channels W1 and W2. Column three provides the median SNR, whereas column four gives the depth (in AB magnitudes) for a 5σ point-like source

band	λ_{eff}	median SNR	median depth
$I_{\rm E}$	6869.74	15.9	26.2
z	8678.90	5	25.9
$Y_{\rm E}$	10688.60	6.5	24.3
$J_{\rm E}$	13422.62	7.8	24.5
$H_{\rm E}$	17410.63	7.2	24.4
W1	33526.00	5	19.8
W2	46028.00	5	18.9

Table 4.4 – Scenarios explored in the simulations for the completeness estimation.

Survey	**NIR bands**	**optical bands**
Euclid	$Y_{\rm E}J_{\rm E}H_{\rm E}$	$I_{\rm E}$
Euclid +LSST	$Y_{\rm E}J_{\rm E}H_{\rm E}$	z
Euclid +WISE	$Y_{\rm E}J_{\rm E}H_{\rm E}$W1 W2	$I_{\rm E}$

For all the objects in our mock quasar catalogue, we calculate the probability P_q, P_s and P_g that the quasar is identified as a quasar, star and galaxy, respectively, as defined in Eq. (3.1). Additionally, each category's templates that allow for a maximum posterior are retrieved.

For the sake of consistency with respect to previous works, we define a successful classification as a high-z quasar when $P_q > 0.1$, and we will examine in Sect. 4.6 how a different definition affects the results.

Results are presented as a function of redshift and apparent $H_{\rm E}$ magnitude in Fig. 4.6, for a selection from *Euclid*'s $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ data. The colours correspond to different increments of the completeness value. We also evaluate how completeness varies when the LSST z band (top-hand figure) and the WISE $W1$ and $W2$ bands (bottom-hand figure) are added to the photometric dataset. This is represented by the dotted red contour lines corresponding to the completeness increments.

To achieve maximum completeness, quasar magnitudes need to be 2 to 3 magnitudes brighter than the EWS detection limit, depending on the redshift. The closer to the detection limit, the lower the SNR, the greater the confusion between quasars and brown dwarfs or galaxies at intermediate redshifts, preventing reliable identification.

Between redshift 7 and 8, completeness is lower than at higher redshift, up to $H_{\rm E} \approx 21$, due to more significant confusion with late-type dwarfs. Due to the substantial increase in late-type dwarf flux in the reddest part of the optical band and the width of the *Euclid* $I_{\rm E}$ band, any late-type dwarf will appear significantly fainter in the $I_{\rm E}$ band than

in any z-band image of comparable depth (see Fig. 4.7) as a result, z-band data from LSST or from the UNIONS survey (https://www.skysurvey.cc/) provide significantly improved discrimination against L- and T- dwarfs (dashed red curves in the left-hand plot of Fig. 4.6).

Fig. 4.6 also shows a loss in completeness around redshift 10 and between magnitudes 22 and 23 of $H_{\rm E}$, compared with its value at lower redshift values. We attribute this loss in completeness mainly to an increase in confusion with early-type galaxies as their colours are very similar (see Fig. 4.7).

Finally, the analysis of completeness, including WISE data (bottom panel in Fig. 4.6) shows no significant improvement, in contrast to the improvement mentioned in Sect. 4.4 for the re-identification of known quasars at high redshift. Only a modest improvement in the decrease of the confusion with late-type dwarfs is perceptible for the brightest quasars at the lowest redshift. This is due to the fact that the depth of the WISE data (AB magnitude ~ 19 at 5σ in the $W1$ and $W2$ bands) does not match that of the *Euclid* data, making them unconstrained and allowing no significant improvements over the *Euclid* data alone, either above or below the WISE detection limit.

4.5.3 Redshift estimation

Following Chapter 3, the redshift estimated by `Owl-z` corresponds to the maximum posterior value of the parameter z, the maximum posterior $\int W_q(z_{out}, H_E) dH_E$. We refer to it as "photometric redshift" or z_{out} hereafter.

Photometric redshift accuracy has been estimated by comparing the true redshift injected in the simulation z_{true} to z_{out}. Fig. 4.8 displays this comparison for the redshift interval explored in the simulations presented in Sect. 4.5.2. To evaluate the accuracy of the photometric redshift, the sample has been limited to objects brighter than $H_E < 24$, identified as quasars using the same criterion as in Sect. 4.5.2, that is $P_q > 0.1$.

As shown in Fig. 4.8, an excellent correlation between z_{true} and z_{out} is found, irrespective of the photometric dataset. For the $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ filter set, the Pearson correlation coefficient is found to be 0.97, whereas the Spearman correlation coefficient is 0.98, indicating a strong linear correlation. The same results are found for the $zY_{\rm E}J_{\rm E}H_{\rm E}$ dataset. The classical quality estimators used for photometric redshift yield the following results, with $\Delta z = z_{out} - z_{true}$: $\sigma(\Delta z/(1+z)) = 0.030$, the median of $(\Delta z/(1+z)) =$ -0.007, and the normalised median absolute deviation, which is less sensitive to outliers: $\sigma_{z,MAD} = 1.48 \times$ median $(|\Delta z|/(1+z)) = 0.021$. Regarding the fraction of outliers, defined in a conservative way as sources with $\Delta z > 0.1(1+z_{true})$, it is found to be only 2.45% (2.23%) for the $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}(zY_{\rm E}J_{\rm E}H_{\rm E})$ datasets. Only 13.7% (13.3%) of the sample exhibit $|\Delta z| > 0.05(1+z_{true})$ for the $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}(zY_{\rm E}J_{\rm E}H_{\rm E})$ datasets.

These results demonstrate the quality and the reliability of the photometric redshift obtained by `Owl-z` for sources identified as high-z quasars. This is of utmost importance, given the use expected for the output redshift in the selection of samples for spectroscopic follow-up. It is also essential in the determination of the purity, as shown in Sect. 4.5.4.

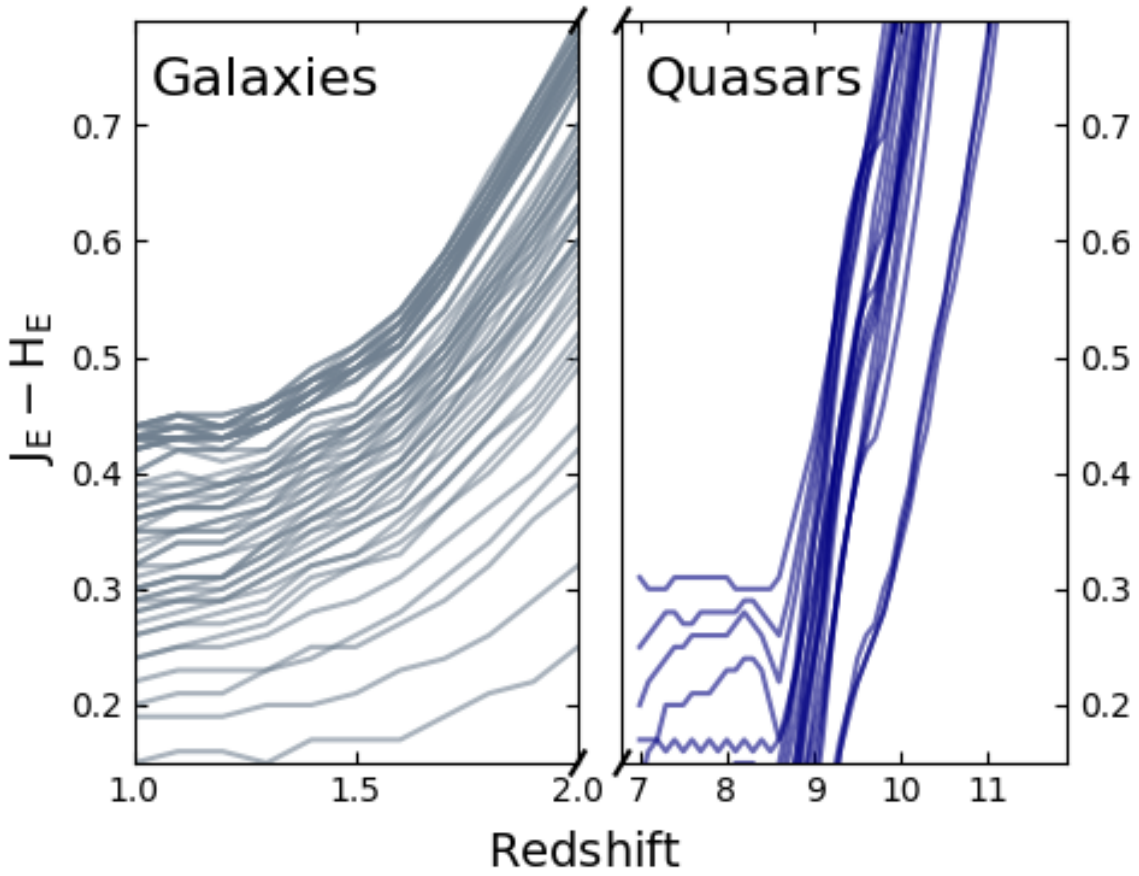


Figure 4.7 – Tracks of $J_{\rm E}$-$H_{\rm E}$ colours as a function of redshift of Early-Type galaxies (left panel) and high redshift quasars (right panel). For the early-type galaxies, the lower the colour of the track, the younger the galaxy.

4.5.4 Purity and F-measure

The purity of the sample expresses the expected reliability of `Owl-z` in extracting true quasars from photometric observations of real fields, which is applied to the EWS in this case. Here, we use the *Euclid* bands $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$, the photometric depths and the variable SNR maps of *Euclid*. With the definition in Sect. 4.3.2, a purity of 100% can either mean that no object(quasar or contaminant) has been identified as a quasar or all the identified objects are quasars.

`Owl-z` has been run on these simulated catalogues presented in Sec. 4.5.1.2. The pre-selection criteria and $P_q > 0.1$ threshold applied are identical to those used for the completeness in Sect. 4.5.2. We study the variation of the purity as a function of redshift (z_{out}) and apparent magnitude in the reference filter. The results presented in Fig. 4.9 provide a summary of the performance of `Owl-z` in terms of completeness and purity. In each bin, the figure displays the values of C and P. Additionally, it illustrates another performance estimator, the F-measure, which will be defined in Sect. 4.5.4.1. As shown

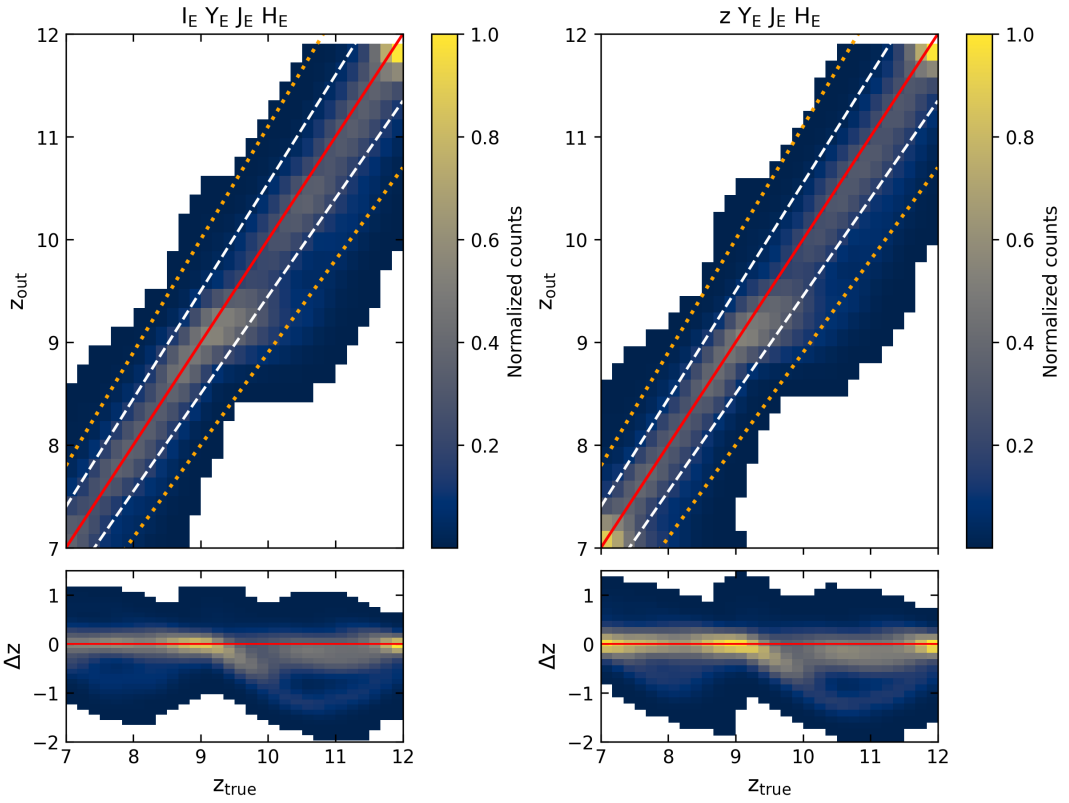


Figure 4.8 – Top panel shows the comparison between the true redshift and the redshift estimated by `Owl-z` with both *Euclid* $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ dataset *(left)* and z$Y_{\rm E}J_{\rm E}H_{\rm E}$ dataset *(right)*. Red solid lines display the 1:1 bisector; the thresholds corresponding to $|\Delta z| > 0.1(1+z_{true})$ and $|\Delta z| > 0.05(1+z_{true})$ are also displayed to guide the eye as orange dotted and white dashed lines, respectively. The bottom panel shows the residual $\Delta z = z_{out} - z_{true}$ as a function of redshift, with the red line corresponding to $\Delta z = 0$

in this figure, the code is expected to achieve a high purity for the brightest sources, irrespective of the redshift. The influence of the precise choice for the threshold in P_q is discussed in Sect. 4.6.1.

4.5.4.1 Global performance

To evaluate the global performance of `Owl-z` by combining both the completeness and purity measurements, we use the F-Measure value introduced in Sect. 4.3.

The classification performance—specifically, the code's ability to identify high-z quasars—is evaluated using the F-measure, which indicates the quality of the classification. A higher F-measure indicates a superior classification performance. We use F-measure hereafter as a performance metric.

The results summarising the performance of `Owl-z` in terms of F-measure are also provided in Fig. 4.9. As shown in this figure, high values of F-measure are obtained for bright ($H_{\rm E} < 21$) sources at all redshifts between $7 < z < 11$, indicating that the

performance in the spectroscopic confirmation of bright rare quasars is expected to be high. On the contrary, the F-measure drops significantly towards fainter magnitudes and high redshifts, following the drop in the purity value. Note that satisfying results can still be achieved for $H_{\rm E} < 23$ and $z < 8$. Also, the performance depends on the choice of P_q threshold, as discussed in Sect. 4.6.1.

4.6 Discussion

In this section, we study the influence of different parameters on the performance of `Owl-z`, such as the threshold used to select quasars as a function of redshift and magnitude. We also give some guidelines to optimise the selection of high-z quasars and their spectroscopic follow-up. The discussion in this section is based on the performance expected on the EWS.

4.6.1 Influence of the selection threshold

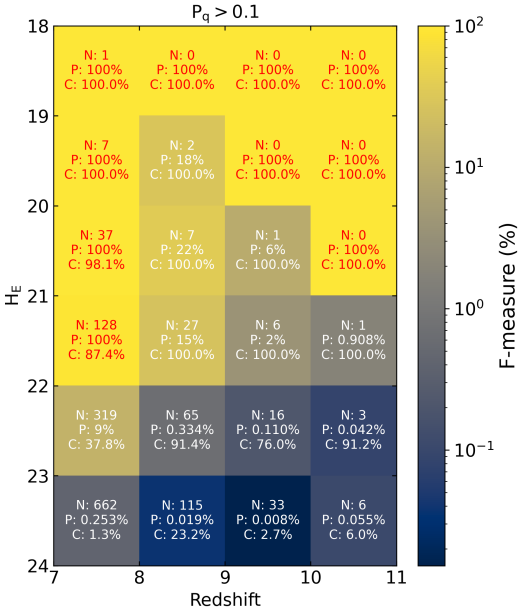


Figure 4.9 – Classification performance of `Owl-z` determined using the threshold $\zeta = 0.1$. In each bin, we report N: the number of quasars injected in the bin, P: the purity and C: the completeness. The colour code indicates the performance measurement F-measure in each bin

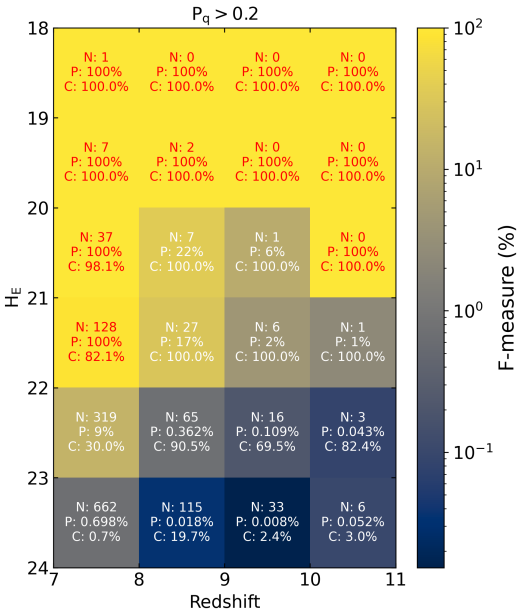


Figure 4.10 – Same is in Fig 4.9, for a threshold $\zeta = 0.2$

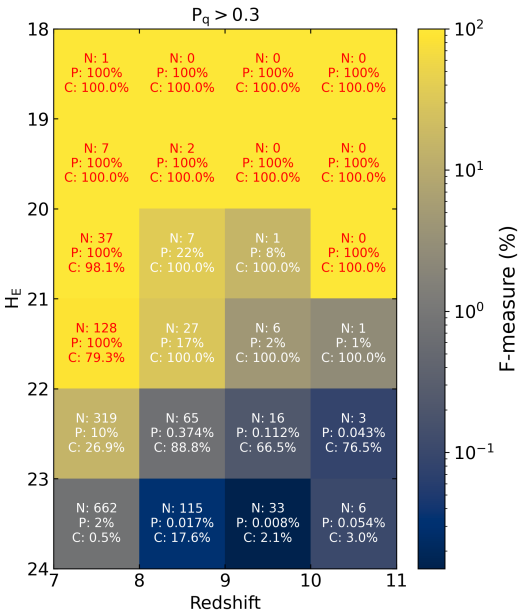


Figure 4.11 – Same is in Fig 4.9, for a threshold $\zeta = 0.3$

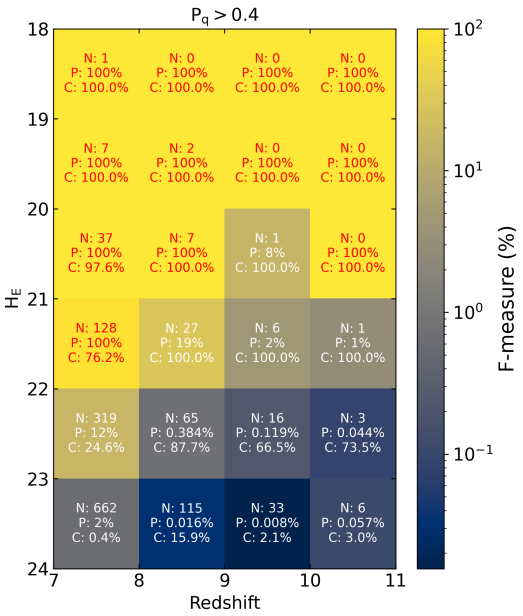


Figure 4.12 – Same is in Fig 4.9, for a threshold $\zeta = 0.4$

In the previous section, we consider an object as a quasar candidate when the posterior probability $\mathbb{P}$ is above the threshold $\zeta = 0.1$. See the results in Figures 4.9 to 4.17. This parameter clearly impacts the performance because lower values will improve the completeness while inducing worse purity values and vice-versa. There is a trade-off to find to optimise the performance. Here, we discuss the influence of ζ on F-measure.

To this end, we use the data discussed in Sect. 4.5.4, change the value of ζ to select quasar candidates and recalculate the F-measure in each bin. The results are presented in Fig. 4.9 for $\zeta = 0.1$ and in Fig. 4.17 for $\zeta = 0.9$.

Two main trends are observed for sources respectively brighter/fainter than $H_{\rm E}$= 22.

Increasing the threshold improves purity, albeit at the expense of completeness for faint sources ($H_{\rm E} > 22$). For bright sources ($H_{\rm E} < 22$), the completeness is almost unaffected by increasing the classification threshold, thanks to the ability of the code to compute high P_q values for bright quasars.

For faint sources, the performance of `Owl-z` drops rapidly towards the faintest magnitudes above $H_{\rm E} = 22$, irrespective of the redshift bin and ζ due to low SNR values, which contribute to a rapid decrease in purity, even when completeness is still good. The analysis of the results shows that early-type galaxies are the main contributors to contamination at these faint magnitudes ($H_{\rm E} > 22$). Conversely, at bright magnitudes ($H_{\rm E} < 22$ and brighter), the SNR is high and enables adequate classification between stars, galaxies and quasars with good completeness, and higher values of ζ improve the purity and the values of the F-measure increase.

In Fig. 4.9, the $8 < z < 10$ redshift bins for bright candidates ($19 < H_{\rm E} < 22$) are clearly affected by low purity values. The main sources of contamination in this area are early-type galaxies at redshifts $1 < z < 2$ due to the confusion between the 4000

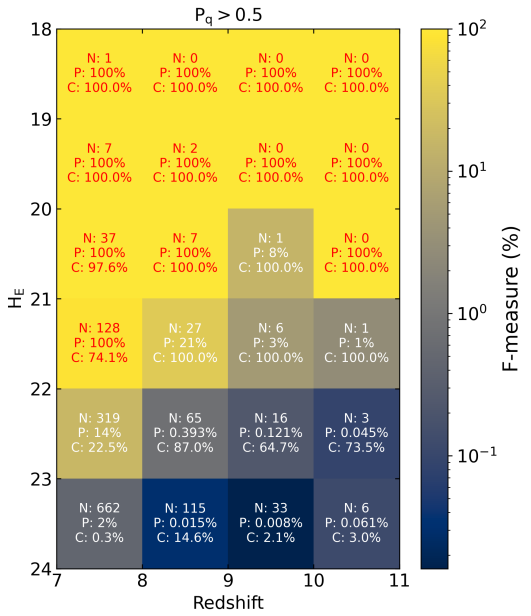


Figure 4.13 – Same is in Fig 4.9, for a threshold $\zeta = 0.5$

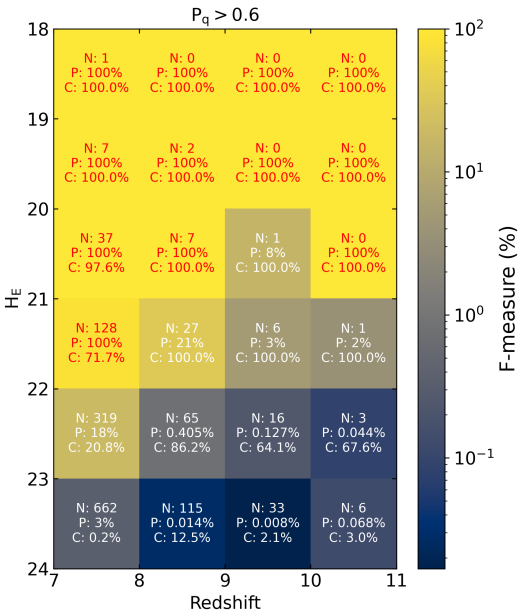


Figure 4.14 – Same is in Fig 4.9, for a threshold $\zeta = 0.6$

Å break at intermediate redshift and the Lyman-α break at high redshift, as explained in Sect. 3.3.3. There is also a minor contribution from contamination by late L-type dwarfs. Increasing the classification threshold has the immediate effect of reducing this contamination and, therefore, improving the purity. As completeness remains high, this results in a significant improvement in $F-$measure values.

This analysis suggests that an adaptive threshold can be used to optimise the $F-$measure values in the magnitude/redshift parameter space. This is the subject of the next section.

4.6.2 Optimising the identification of $z > 7$ quasars

We now investigate if the value of the F-measure can be further increased by adjusting the selection threshold ζ in each redshift and magnitude bin. To this end, we use the data discussed in Sect. 4.4, change the value of ζ in increments of 0.1 and select the value that maximises F in each bin. The result is summarised in Fig. 4.18, showing the value of ζ that maximises F in each bin of magnitude $H_{\rm E}$ and redshift z_{out}. The results show that, for bright magnitudes ($H_{\rm E}< 21$), a low threshold value $\zeta = 0.1$ is sufficient to maximise F-measure. This is due to the robustness of the selection for high SNR values irrespective of the redshift, as shown previously. On the contrary, for faint magnitudes ($H_{\rm E}> 21$), the value of ζ needs to be increased to compensate for the decline in purity, requiring ζ values as high as 0.9 towards the faintest magnitudes.

In conclusion, adopting a variable threshold enables the optimisation of the F-measure and, consequently, the effectiveness of photometric and spectroscopic follow-up campaigns based on `Owl-z` selected candidates. Finding the best threshold value ζ for a quasar

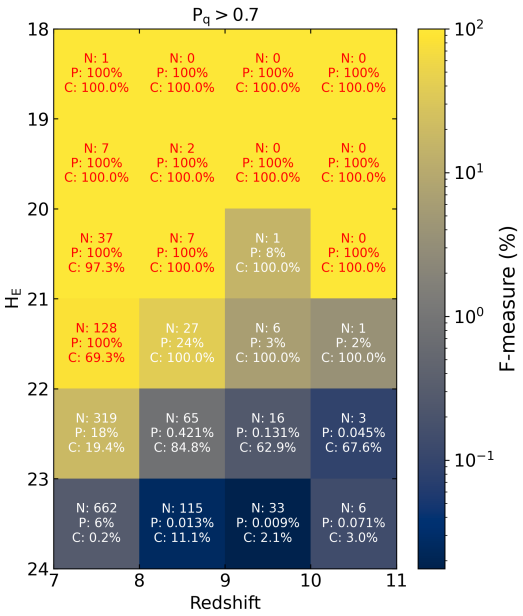


Figure 4.15 – Same is in Fig 4.9, for a threshold $\zeta = 0.7$

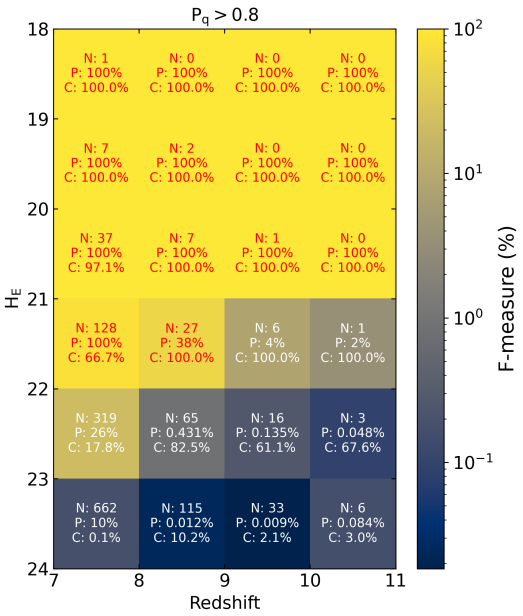


Figure 4.16 – Same is in Fig 4.9, for a threshold $\zeta = 0.8$

candidate with magnitude $H_{\rm E}$ and a redshift z_{out} from any other survey data can be done by following the same prescription presented in this article in the case of EWS.

4.6.3 Influence of the position in the EWS footprint

In this section, we evaluate the effect of different locations in the EWS footprint. As described in Euclid Collaboration: Scaramella et al. 2022, different zodiacal light levels across the footprint modify the SNR for a given magnitude in each filter, and different sky positions and Galactic coordinates introduce varying stellar densities, as described in Sect. 3.3.2. To this end, we select six regions in the *Euclid* footprint, representing different zodiacal light levels and Galactic coordinates. These regions are shown in Fig. 4.5, and Table 4.5 lists their Galactic coordinates and SNR values at reference magnitudes in all Euclid filters.

According to the results presented in Sect. 4.5, the confusion by MLT dwarfs is only sensible in the redshift interval $z = [7, 8]$, where there is also contamination by early-type galaxies. Therefore, we focus the analysis on this redshift bin because it is more susceptible to being affected by the precise distribution assumed for MLT stars. In addition, this redshift bin is also the most populated by quasars, meaning that the results are expected to be statistically more significant than in the other bins. For each of these regions, we compute simulated catalogues following the prescriptions in Sect. 4.5.4, and we run `Owl-z` to derive the value of F-measure.

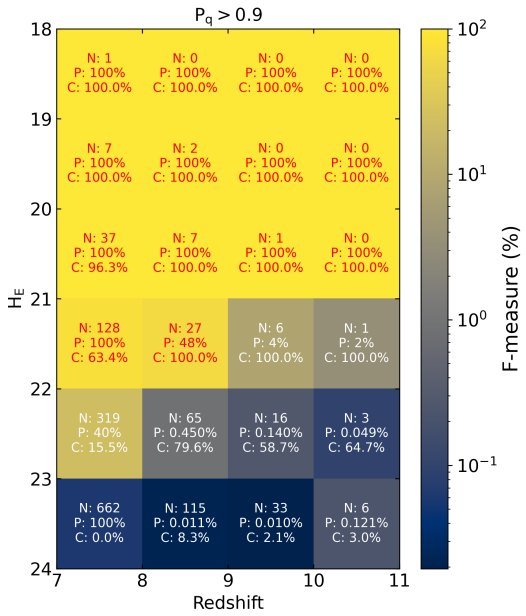


Figure 4.17 – Same is in Fig 4.9, for a threshold $\zeta = 0.9$

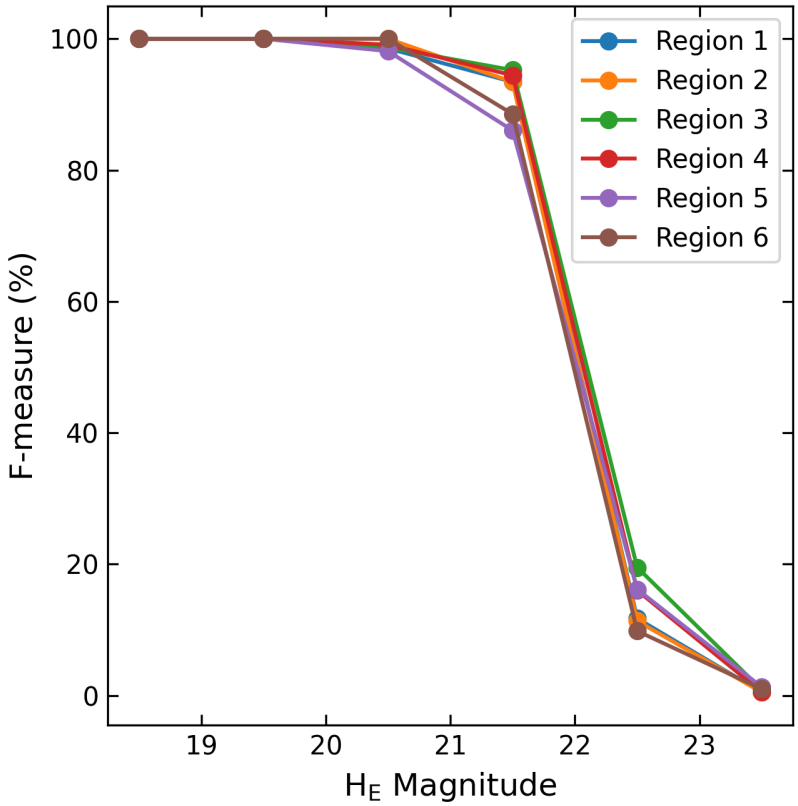


Figure 4.19 – F-measure evolution as a function of the H_E magnitude in each of the regions shown in Table 4.5, computed in the $7 < z < 8$ interval.

The results are shown in Fig. 4.19 and display minor variations in the values of F-measure. Regions 5 and 6 exhibit a slight decline in F-measure within the magnitude bin $[21, 22]$ compared to the others. This can be attributed either to the fact that these regions possess the lowest Galactic latitudes ($b = 26$ and $b = 30$) or to the most unfavourable SNR values. According to the results presented in Sect. 4.5, the Comparison with region 3, which is also at a low Galactic latitude but has higher SNR values, suggests that SNR variation is the dominant factor in variations in F-measure. In other words, brown dwarf contamination does not appear to depend significantly on Galactic latitude within the *Euclid* footprint, which is consistent with figure 3.3 and with the fact that up to magnitude ≈ 22 contamination is dominated by late-type L and T dwarfs at distances small compared to the scale height of the Galactic disc.

4.6.4 Influence of thick disc MLT

We aim to evaluate the robustness of our selection process when adding contamination by thick disc MLTs that are not included in the model. We assess the impact on the F-measure of adding a contribution of thick-disk brown dwarfs to the input catalogue, following the prescriptions described in Sect. 4.3, and ignoring differences in brown dwarf populations between the thin and thick discs. To do so, we employ the density function outlined in Chapter 3 to generate a population of MLT dwarfs from the thick disc, in addition to the previously added MLT dwarfs from the thin disc, and we analyse the impact on F-measure on the six regions introduced previously. The results are presented in Fig. 4.20, comparing two scenarios: one where only thin disc MLT dwarfs are injected in the input catalogue (shown by yellow lines) and another where thick disc MLT dwarfs are added (shown by blue lines). The trends on these graphs can be qualitatively explained by the relative density of stars between the thin and thick discs as a function of magnitude. At low magnitudes, the thin disk dominates at all Galactic latitudes, and there are no differences between the two scenarios. At magnitudes above 21, however, and at high Galactic latitudes only (see figure 3.3), the thick disk population begins to contribute to the counts and introduce a difference between the two scenarios (regions 1, 2, 4 and to a lesser extent region 3, which is at a relatively low Galactic latitude but closest to the Galactic centre in longitude). Overall, this analysis serves as a sanity check, confirming that the precise modelling of the thick disc is unlikely to significantly affect the results for the EWS but would be warranted for deeper surveys.

4.6.5 Comparison with previous works

The results obtained in this article can be compared to previous similar works, at least qualitatively. A proper comparison will need to run the various codes on the same simulated catalogues, which is currently impossible because not all codes are publicly available. The Bayesian classification approach is similar to that adopted by Pipien et al. 2018b for the analysis of the CFHQSIR survey (Pipien et al. 2018a) searching for quasars at $6.5 < z < 7.5$. Our results in the lowest redshift bin are comparable to Pipien et al. 2018b. Note that Pipien et al. 2018b do not include the population of contaminant

Table 4.5 – Properties of areas and catalogues simulated across the EWS footprint at reference magnitudes of 24 for $Y_{\rm E}J_{\rm E}H_{\rm E}$ and 24.5 for $I_{\rm E}$. The SNR values are averaged over each surface.

Region	Surface (deg^2)	l(deg)	b(deg)	SNR$_{I_{\rm E}}$	SNR$_{Y_{\rm E}}$	SNR$_{J_{\rm E}}$	SNR$_{H_{\rm E}}$
1	200	270	−75	17.5	7.0	8.45	7.9
2	200	180	85	16.5	6.6	7.8	7.4
3	200	78	35	17.8	7.0	8.5	7.7
4	200	107	55	18.6	7.4	8.8	8.22
5	100	227	26	13.8	5.6	6.6	6.1
6	100	270	30	14.3	5.9	7.1	6.5

galaxies, a choice that has no impact on their results because the dominant contamination in this bin comes mainly from LT stars, as shown in Sect. 4.5.

Our work also compares to and complements the study conducted by Euclid Collaboration: Barnett et al. 2019 on the EWS using a similar Bayesian classification method and the same contaminant populations of stars and galaxies. With respect Euclid Collaboration: Barnett et al. 2019, the work with `Owl-z` extends the study beyond $z > 9$ and introduces different performance metrics in addition to the completeness, namely purity and F-measure (combining completeness and purity). Our definition of purity has also been adapted to account for rare quasars beyond $z > 9$ that could be found in the EWS if they ever exist. We also use updated prescriptions for the quasar LF and the contaminant populations. It is worth noting that we have used the EWS SNR maps here, making our predictions closer to the real observations. The performance of `Owl-z` in terms of completeness in the common redshift range with Euclid Collaboration: Barnett et al. 2019 is found to be better than in this previous work. This improvement is mainly because the global SNR is expected to be better than assumed by Euclid Collaboration: Barnett et al. 2019. On the contrary, the effect of the detailed 2D modelling of the SNR in the EWS has little impact on the performance of `Owl-z` and similar codes, possibly because the depth is not dramatically different across the survey, as discussed in Sect. 4.6.3.

Possibly, the main advantage of `Owl-z` with respect to other existing codes is its versatility in adapting to different photometric surveys, as shown in Sect. 4.4. The code will be publicly available and open to access.

4.7 Conclusions

The conclusions of this chapter can be summarised as follows:

— This chapter briefly presents the ingredients, outputs and practical use of `Owl-z`, a public Bayesian code developed during this thesis aiming at identifying $z > 7$ quasars in massive photometric surveys. `Owl-z` is based on the classification of sources as high-z quasars or contaminants, the latter being MLT stars or early-type galaxies at intermediate redshifts. The populations of contaminants have been carefully modelled for the needs of the Bayesian model selection. Although the code has been tuned to be used on the EWS, it can be easily adapted to different photometric

surveys, as shown in this work.

— `Owl-z` has re-identified all spectroscopically-confirmed quasars at $z > 7$. This exercise has also demonstrated the versatility of `Owl-z` regarding its application to different photometric catalogues.

— The performance of `Owl-z` based on F-measure, a metric combining completeness and purity, has been estimated using simulations for the EWS in a wide range of redshifts ($7 < z < 12$) and reference magnitudes ($18 < H_{\rm E} < 24.5$). The results show that `Owl-z` reaches full performance for bright sources ($H_{\rm E} < 22$) irrespective of the redshift, meaning that the performance in the spectroscopic confirmation of bright rare quasars selected by `Owl-z` is expected to be high. The performance drops significantly towards fainter magnitudes and high redshifts.

— The threshold value for the selection of quasars can be applied in the post-treatment phase, meaning that this value can be tuned to optimise the selection of $z > 7$ quasars. For instance, a small/large value can be used for the brightest/faintest sources ($H_{\rm E} < 21$ and > 21, respectively) to optimise the performance. Given the steep change of the optimum threshold on the magnitude versus redshift plane, a relatively crude cut could be adopted to optimise the spectroscopic follow-up.

— A uniform performance is obtained with `Owl-z` across the EWS footprint, irrespective of the Galactic coordinates and relative SNR values.

— The results of the code remain robust despite inevitable uncertainties in the detailed modelling of the contaminating stellar populations, particularly the thin and thick discs.

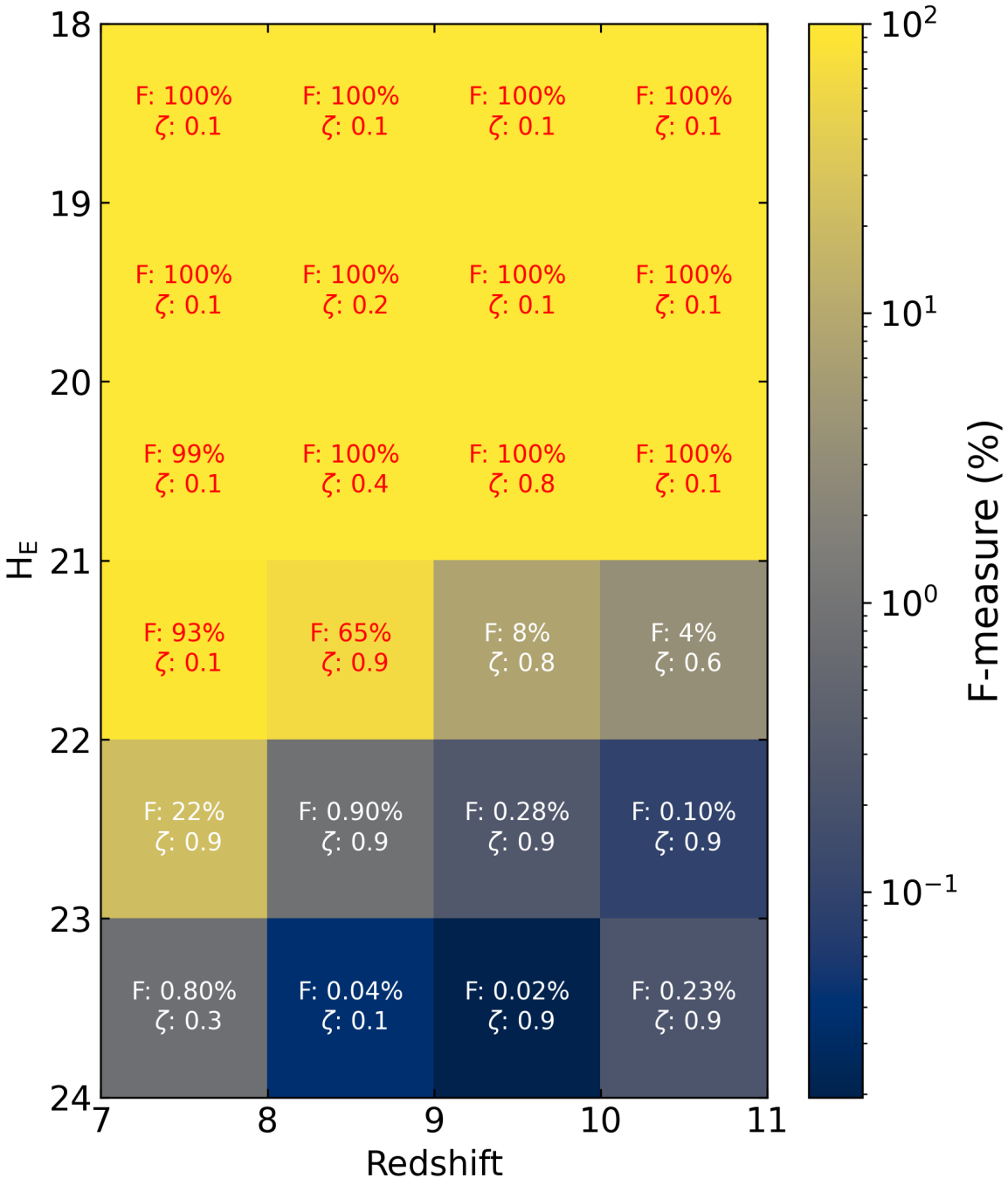


Figure 4.18 – Classification performance of `Owl-z` on the magnitude $H_{\rm E}$ and redshift z_{out} plane. The threshold value of ζ that maximises F-measure is given for each bin.

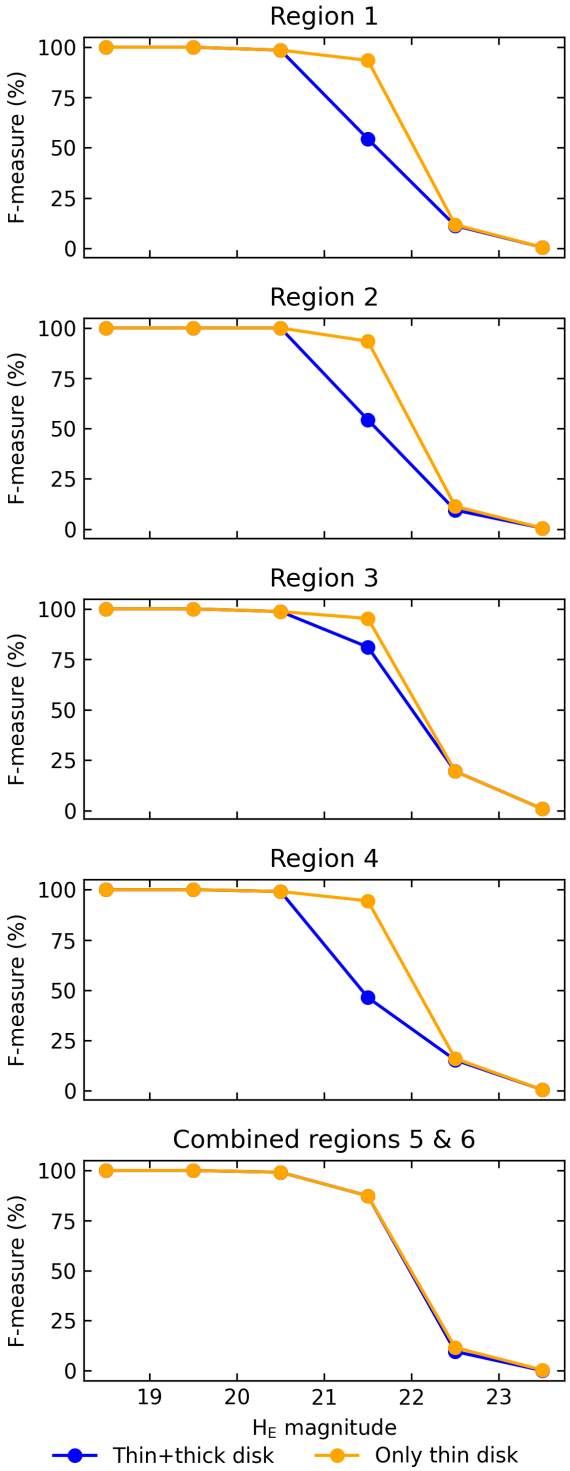


Figure 4.20 – F-measure evolution as a function of the $H_{\rm E}$ magnitude in each of the regions shown in Table 4.5, computed in the $7 < z < 8$ interval. Two cases are shown for the contribution of MLT stars. In yellow is presented the evolution when only the thin disc contribution is accounted for, both in the simulation and the Bayesian modelling. In blue, the same result is shown when the contribution of the thick disc is added to the simulations.

5 Application of `Owl-z` to Euclid data

Contents

5.1 The Euclid Early Release Observations Program 108
5.1.1 Observations 109
5.1.2 Photometric catalogue 110
5.1.3 Preselection of high-redshift quasar candidates 111
5.2 Results 112
5.2.1 Redshift estimation 114
5.2.2 Visual inspection 115
5.2.3 Spectroscopic confirmation of a Brown Dwarf 120
5.3 Discussion 123

During the last year of my thesis, in May 2024, the first reduced Euclid data were made available to the consortium. This is the "Quick Euclid Data Release", waiting for the first official "Early Release", which will take place in March 2025. A few months before finishing this manuscript, I was therefore able to perform tests with `Owl-z` on real Euclid data, with still preliminary results for reasons discussed below. I have contributed to several articles related to the Euclid Early Release Observations (ERO hereafter), as described in Weaver et al. 2024, Cuillandre et al. 2024 and Atek et al. 2024. This chapter is focused on the application of `Owl-z` to the ERO data program. I start by describing the ERO observations and data, a small but representative sub-sample of the data expected at the end of the survey. After that, I present the results obtained using `Owl-z` on these data for the selection of high-z quasars, and I discuss these results.

5.1 The Euclid Early Release Observations Program

The ERO program was designed to illustrate the capabilities of the Euclid space telescope across a diverse set of astrophysical environments before the main mission survey began. As part of this effort, six targeted projects were selected to demonstrate Euclid's observational advantages, including its ability to study large-scale structure, galaxy evolution, and the high-redshift universe (Cuillandre et al. 2024). Among these projects, the "Magnifying Lens" program (PI: H. Atek) focuses on two massive clusters of galaxies, Abell 2390 and Abell 2764, using both Euclid's instruments—VIS/I_E for high-resolution optical imaging, NISP/Y_E, J_E, and H_E for near-infrared photometry, and the NISP grism for slitless spectroscopy (Atek et al. 2024). Abell 2390 and Abell 2764 act as cosmic lenses, using their gravitational potential to magnify background galaxies, enabling an unprecedented census of high-redshift systems ($z \gtrsim 6$), including (rest-frame) UV-bright galaxies critical to understanding the epoch of reionisation (Weaver et al.

2024). Euclid's capacity to probe these galaxies over a wide 0.75 deg^2 field represents a substantial improvement over previous deep surveys conducted with the Hubble Space Telescope (HST) and JWST, which were limited to smaller areas (e.g., CANDELS, Duncan et al. 2014; Bouwens et al. 2015; Finkelstein et al. 2015). By dramatically increasing the number of detectable luminous galaxies from the reionisation epoch, these observations provide crucial constraints on the bright end of the UV luminosity function, an important diagnostic of early galaxy formation and evolution. In addition, the same dataset enables complementary analyses, including weak lensing studies, to probe the distribution of dark matter and investigations into their member galaxies' virial masses and assembly histories.

The ERO program is not only a demonstration of Euclid's imaging and spectroscopic capabilities but also an invaluable dataset for exploring some of the most pressing questions in extragalactic astronomy, from the role of galaxies and quasars in cosmic reionisation to the large-scale distribution of dark matter. Given the unprecedented scale of Euclid's near-infrared survey, these initial observations set the stage for identifying rare and extreme objects, such as massive dusty star-forming galaxies and high-redshift quiescent systems, which will be critical for constraining the evolutionary pathways of galaxies at early cosmic epochs.

5.1.1 Observations

The ERO "Magnifying Lens" project consists of intermediate-depth observations of two well-known clusters of galaxies, Abell 2390 and Abell 2764. While both are massive systems rich in lensed background sources, Abell 2390 ($z = 0.228$, Sohn et al. 2020) is significantly better studied than Abell 2764 ($z = 0.07$, Katgert et al. 1996). Indeed, Abell 2390 is one of the better-studied lensing clusters, showing several multiple imaged systems with known redshifts, providing a fair constraint of the mass distribution in its core (see, e.g. Pello et al. 1991; Frye &Broadhurst 1998; Pelló et al. 1999). See Atek et al. 2024 for a complete description of the field selection.

For each cluster, Euclid obtained three reference observing sequences (ROS, see Fig. 8 of Euclid Collaboration: Scaramella et al. 2022), each spanning 70.2 minutes and covering imaging in the VIS/I_E and NISP/Y_E, J_E, and H_E bands. A dithering pattern of $3' \times 3'$ was used between each ROS to ensure uniform coverage across the 36 chips. Individual exposures in the I_E, Y_E, J_E, and H_E bands were stacked to generate four deep mosaics using `SWarp` (Bertin et al. 2002), with images produced at two resolutions: a native pixel scale of $0.''1$ pixel^{-1} for VIS and $0.''3$ pixel^{-1} for NISP. The background light was modelled and subtracted using a 64-pixel mesh with a $3\times$ smoothing factor. The VIS background was particularly challenging due to bluer Galactic cirrus foregrounds. Although alternative reduction methods were considered, this work adopts the "Compact-sources" images optimised for photometry of faint, compact objects (Cuillandre et al. 2024).

Each mosaic spans 0.75 deg^2, making the VIS images among the deepest obtained at such high resolution over degree-scale fields, reaching a depth of $I_E = 27.1$ AB (5σ point source). This is comparable to the largest HST mosaic ever obtained (ACS/F814W over 1.7 deg^2 in COSMOS, Koekemoer et al. 2011). Similarly, the Y_E, J_E, and H_E bands reach depths of ~ 24.5 AB (5σ point source), making them some of the deepest NIR images covering degree-scales, rivalling the VIDEO (Jarvis et al. 2013) and UltraVISTA surveys

of COSMOS (McCracken et al. 2012). Although initially designed to identify strongly lensed galaxies, the large contiguous areas outside the cluster centres make these mosaics particularly valuable for identifying rare high-redshift rest-frame UV-bright objects that are difficult to detect in smaller NIR surveys.

5.1.2 Photometric catalogue

Sources were detected in each field using Pythonic Source Extractor (`SEP`, Bertin &Arnouts 1996; Barbary 2016) from a `CHI_MEAN` coadd of all three NISP Y_E, J_E, and H_E mosaics, constructed with `SWarp` (Bertin et al. 2002). To maximise completeness, `SEP` was configured to detect sources with at least three contiguous pixels above a 1.5σ threshold after smoothing with a $0.''45$ (1.5 pixels) full-width-half-maximum Gaussian kernel (see Euclid Collaboration et al. 2024b for relevant discussion). Since separating low-resolution NISP-only sources from line-of-sight neighbours is crucial, `SEP` was configured with aggressive deblending settings, using 2×10^5 thresholds and a contrast of 10^{-5}, tuned through visual inspection. Cleaning procedures were omitted to avoid removing real sources of interest. A total of 501,994 sources were recovered, with 98% exhibiting $S/N > 3$ in at least one of Y_E, J_E, and H_E bands. Approximately equal numbers of sources were identified in each 0.5 deg^2 field, with 252,675 in Abell 2390 and 249,319 in Abell 2764.

Photometry was extracted at source positions using circular apertures of diameters $0.''1, 0.''2, 0.''3, 0.''6, 1.''0, 1.''2, 1.''5$, and $2.''0$. Uncertainties were measured similarly using weight maps, and per-pixel error distributions were verified against direct noise estimates from the science mosaics. Given the $0.''3$ pixel^{-1} scale of NISP, aperture diameters below $0.''6$ are primarily suited for VIS photometry.

Due to the different resolutions between VIS and NISP, a naive approach would be to convolve all images to match the broadest PSF (H_E band). However, this work leverages VIS as a "dropout" band, where colour terms between I_E and the NISP bands (Y_E, J_E, and H_E) are secondary to the more critical SNR in VIS/I_E. Convolving VIS to lower resolutions would degrade its sensitivity. Instead, the VIS S/N is optimally determined from the original high-resolution images, using a carefully selected aperture that balances robustness against noise with minimal contamination from neighbouring sources. Since the PSFs of the three NISP bands are relatively similar and high-redshift Lyman break galaxies are expected to be compact ($< 1''$ radius, e.g., Bowler et al. 2017), NISP colours can be accurately determined without requiring PSF-matching.

Aperture flux estimates and uncertainties were corrected to total fluxes using multiplicative factors derived from observed PSFs. These PSFs were obtained by optimally stacking re-centred cutouts of point sources from the science mosaics following Weaver et al. (2024). Correction factors for VIS are 16×, 3×, and 2× for $0.''1, 0.''2$, and $0.''3$ aperture diameters, respectively, while for NISP, they are 1.2×, 1.3×, and 1.4× for $1.''2$ diameter apertures in Y_E, J_E, and H_E, respectively.

A combined photometric catalogue of all 501,994 sources is released alongside this work and the overview paper by Atek et al. 2024 to facilitate community engagement with these ERO data. The catalogue provides unique source identifiers, field-specific IDs, sky coordinates, elementary shape estimates, aperture and total fluxes, flux uncer-

tainties, aperture-to-total corrections, and `Source Extractor`-like flags. Photometry is pre-corrected for Galactic extinction following the extinction maps of Schlafly &Finkbeiner 2011, using the extinction curves of Fitzpatrick &Massa 2007. A significant fraction of artefacts have been flagged, particularly those near mosaic edges where only a single dither contributes and along the diffraction spikes of bright stars. These affected regions, covering approximately 15% of the total catalogue, are identified using a hand-built region mask (`use_phot` = 0) that excludes sources near bright stars, diffraction spikes, and within $1'$ of the mosaic edges.

For full details, consult the README document accompanying the catalogue at https://zenodo.org/doi/10.5281/zenodo.11151975.

5.1.3 Preselection of high-redshift quasar candidates

To identify potential high-redshift ($z \geq 6$) quasar candidates, I applied a series of selection criteria to the photometric catalogue described in Weaver et al. (2024), which comprises a total of 501,995 sources. Given the presence of artefacts near mosaic edges and bright diffraction spikes, I first restricted our sample to sources with `use_phot` = 1 (16% of the overall catalogue), ensuring that only reliable detections within well-calibrated regions of the survey were considered. This step eliminates spurious detections that could otherwise contaminate the selection.

To further refine the candidate list, I imposed colour cuts based on near-infrared photometry, leveraging the expected spectral properties of high-redshift quasars. First, I required all sources detected in the Y_E-band to satisfy a stringent $(J_E - H_E)$ colour cut, keeping only those within the range $-1 < (J_E - H_E) < 1$. This criterion effectively removes sources with extreme red or blue colours inconsistent with the expected properties of high-redshift quasars while allowing for some intrinsic diversity in spectral slopes. For sources that were not detected in the Y_E-band, no $(J_E - H_E)$ constraint was applied to avoid biasing against faint quasar candidates that may fall below the detection threshold in Y_E.

Additionally, I imposed a cut on the $(Y_E - J_E)$ colour to exclude objects with anomalously blue colours that are inconsistent with high-redshift quasar templates. Specifically, I required all sources to satisfy $(Y_E - J_E) > -1$, ensuring that I retain objects with spectral energy distributions (SEDs) consistent with the expected Lyman-break signature at $z \geq 6$. This selection preferentially isolates quasars while reducing contamination from lower-redshift interlopers, including Galactic brown dwarfs and lower-redshift galaxies.

Last but not least, to ensure robustness in the selection, I enforced a minimum SNR threshold in the H_E-band, requiring that all sources have $\mathrm{SNR}_{H_E} > 2$, which retained approximately 90% of the catalogue. This criterion ensures that candidates have reliable flux measurements in H_E, reducing contamination from spurious low-SNR detections while maintaining sensitivity to faint high-redshift objects. Additionally, this requirement satisfies the condition for running OWL-z, which mandates that all sources must be detected in the reference band, H_E in this case. Furthermore, as expected for high-redshift ($z \geq 6$) quasars, all preselected sources exhibit no significant detection in the optical VIS band, with $\mathrm{SNR}_{\mathrm{VIS}} < 2$ (13% of the sources in the catalogue), ensuring they are consistent with the expected Lyman-break signature at these redshifts.

Applying all these preselection criteria resulted in a final sample of 11,963 sources, which were then processed using OWL-z.

These selection criteria are designed to preselect high-redshift quasar candidates while efficiently balancing completeness and contamination. The final sample provides an optimised starting point for further analysis.

5.2 Results

I ran OWL-z on the full catalogue, combining sources from Abell 2390 and Abell 2764. The model parameters were set for the redshift range $6 \leq z \leq 12$, with the reference magnitude chosen as H_E in the interval $18 \leq H_E \leq 26.5$. The results show consistent trends: faint-magnitude objects tend to have lower probabilities (as demonstrated in sect. 4.5.2), while bright sources—though fewer in number—were selected with high probabilities.

A total of 292 sources ($\approx 2\%$) were selected with $P_q > 0.1$, while 187 sources met the stricter criterion of $P_q > 0.9$. This indicates that increasing the probability threshold did not significantly alter the selection, as most chosen sources are on the brighter end of the magnitude range (see Figure 5.1), where the F-measure remains high even at lower thresholds, as expected from the discussion in Chapter 4. Given this, I adopt the lower threshold of $P_q > 0.1$ for subsequent analysis.

Additionally, 78% of the selected sources had galaxy probabilities greater than 0.1 ($P_g > 0.1$), while only 3% had a probability of being brown dwarfs ($P_s > 0.1$).

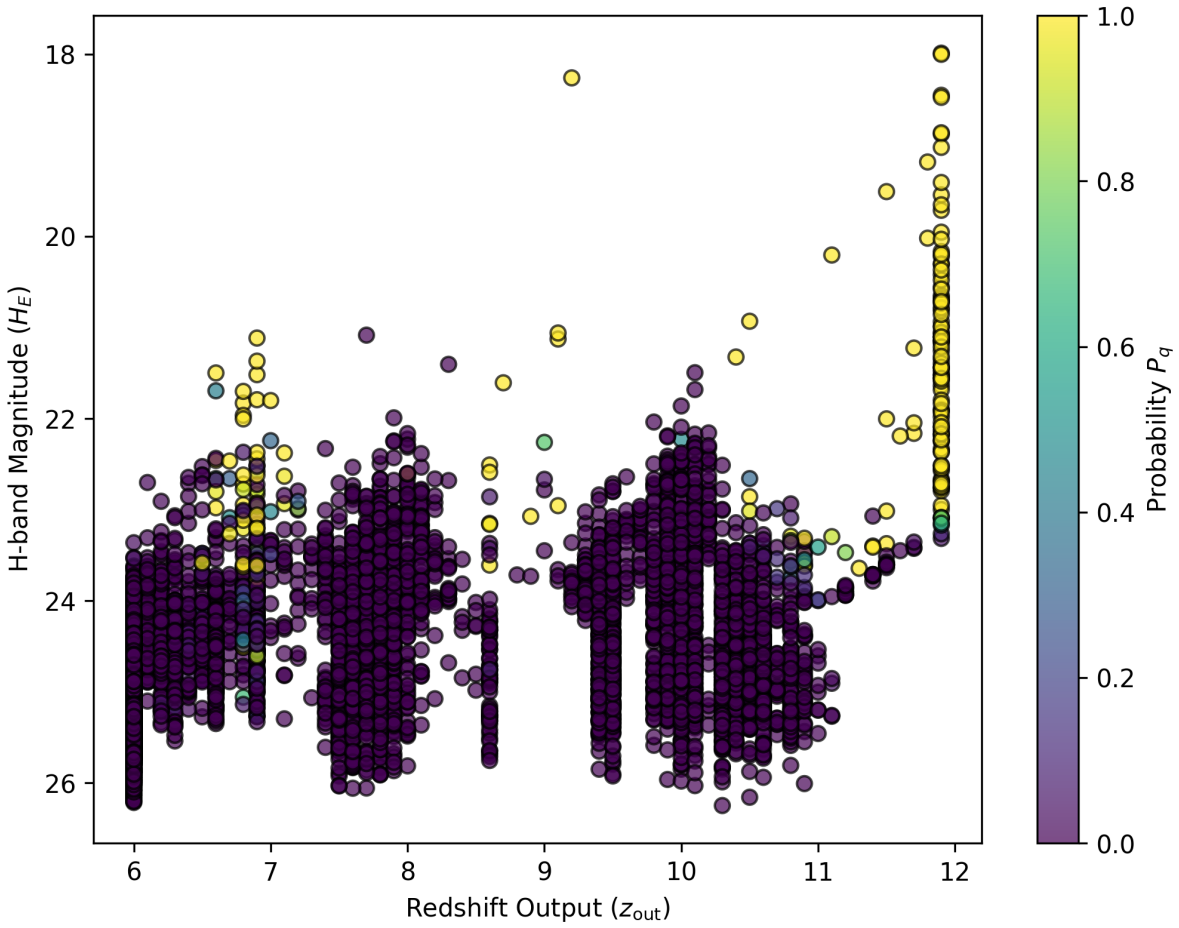


Figure 5.1 – The output redshift (z_{out}) versus H-band magnitude (H_E) for preselected high-redshift quasar candidates. The colour scale represents the quasar probability (P_q), as computed by OWL-z. The y-axis is inverted to follow the conventional magnitude representation, where brighter sources appear at the top. This selection is based on Euclid Y_E, J_E, H_E photometry, ensuring robust detection in H_E and adherence to colour selection criteria for identifying $z \geq 6$ candidates. The binned appearance is due to the large bins used in redshift (dz=0.1)

5.2.1 Redshift estimation

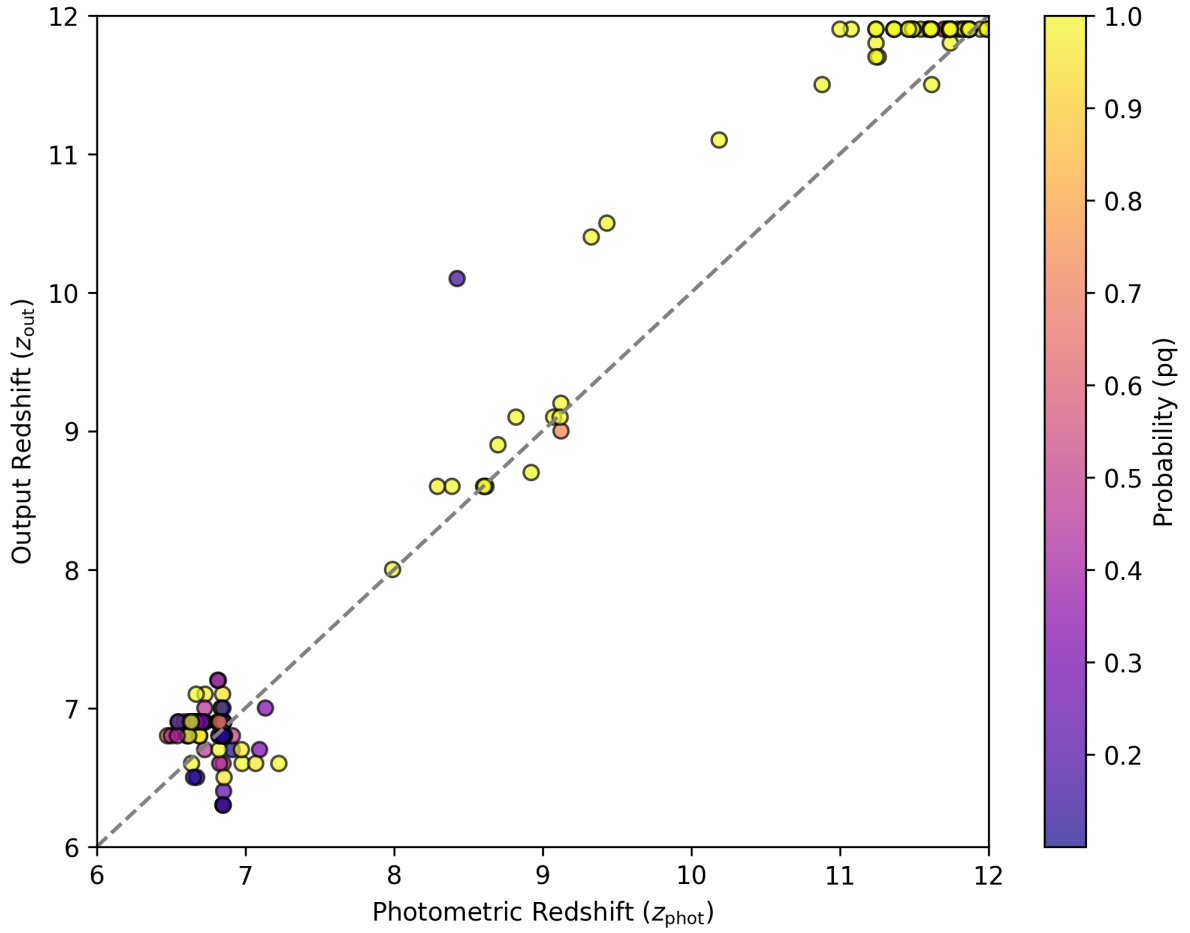


Figure 5.2 – Comparison of photometric redshift ($z_{\rm phot}$) computed by Weaver et al. (2024) with the output redshift ($z_{\rm out}$) from OWL-z, this figure only shows quasar candidates with $P_q > 0.1$. The colour scale represents the quasar probability (P_q). The dashed red diagonal line corresponds to $z_{\rm in} = z_{\rm out}$.

Figure 5.2 indicates good agreement at $z_{\rm out} < 9$. Deviations from the $z_{\rm in} = z_{\rm out}$ line highlight sources where OWL-z assigns significantly different (higher) redshifts compared to the initial photometric estimate. This occurs at higher redshifts, where the photometry is much fainter, making it more challenging to determine a well-constrained maximum posterior estimate.

Regarding the population with assigned $z_{\rm out} \sim 12$, these sources were further investigated through visual inspection and revealed to be mostly artefacts. I consider only quasar candidates with $P_q > 0.1$ for the subsequent analysis. This results in 292 candidates in the Abell 2390 and Abell 2764 fields. Since the images of the two clusters are distinct, I conduct the analysis separately for each cluster. And I only present the visual inspection results on Abell 2390 as the process is similar.

5.2.2 Visual inspection

In this section, I classified the sources based on visual inspection into two main categories: **sources** and **artefacts**. The **sources** category includes all objects resembling real (and red) astronomical sources. In contrast, the **artefacts** category consists of observational artefacts such as persistence, moving objects, and miscellaneous artefacts of unknown origin.

The **artefacts** category also includes sources that appear bright exclusively in the Y_E-band (see Figure 5.5 source Id: 1299), a recurring issue in the image processing pipeline. This phenomenon persists even in newer data sets (post-ERO), and I do not have a definitive explanation.

I now present all the selected sources that appear to be real astronomical objects (17/158 sources in Abell 2390), along with their individual NIR magnitudes and composite colour images created using the three NIR bands. All these sources are not detected in the I_E band. Among them, 12 can be qualified as "secure" regarding the reliability of the detection. They are shown in Fig. 5.3.

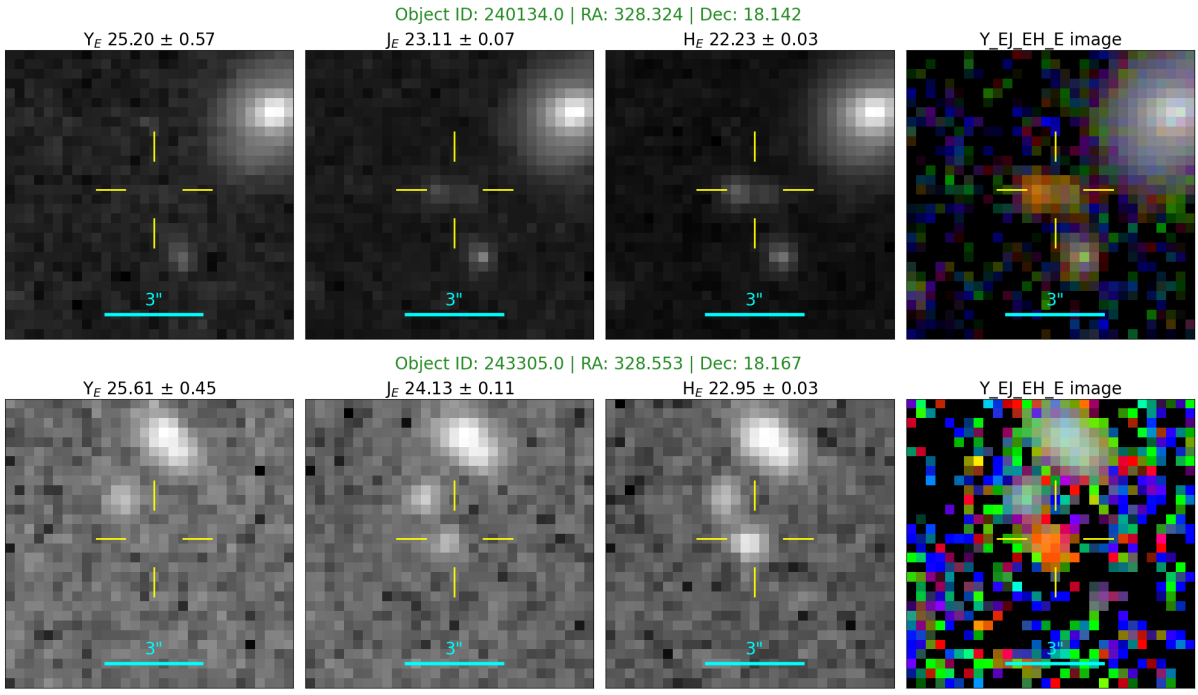
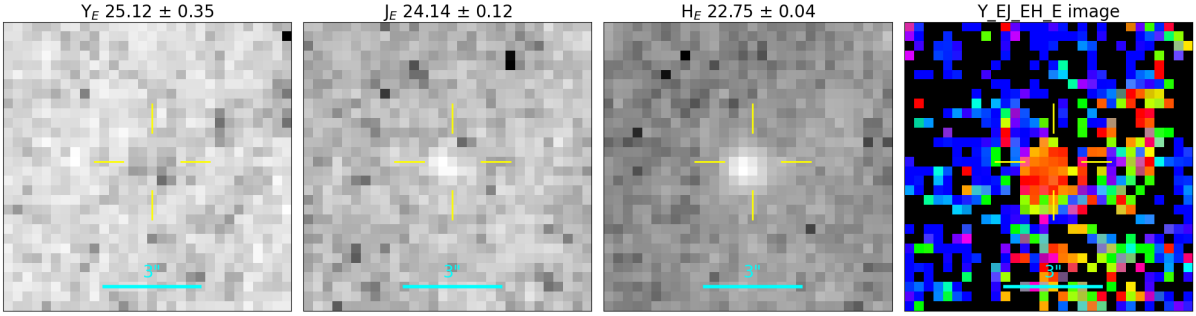
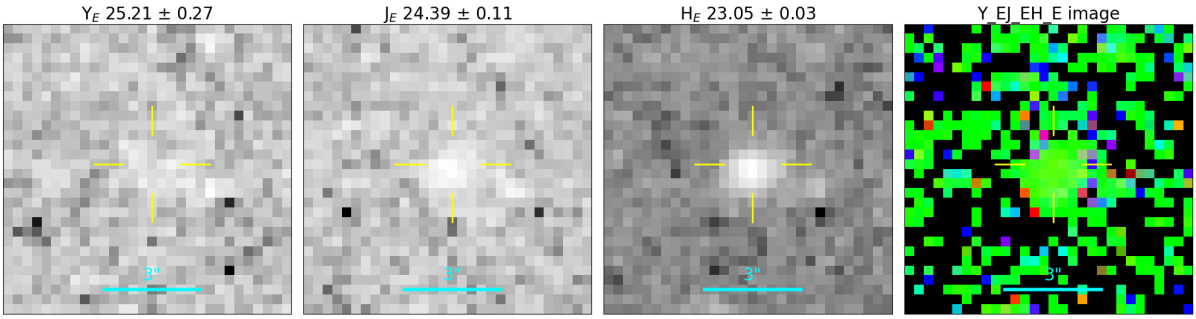


Figure 5.3 – Cutout images (30 × 30 pixels) of sources selected as z>6 quasars by `Owl-z`, and visually resembling real red sources, extracted from Euclid NISP Y_E, J_E and H_E bands. Each panel presents grayscale cutouts for Y_E, J_E and H_E bands, along with a composite RGB image using scaling by Lupton et al. 2004. Yellow markers indicate the source position, and a cyan scale bar represents a length of 3 arcseconds for reference.

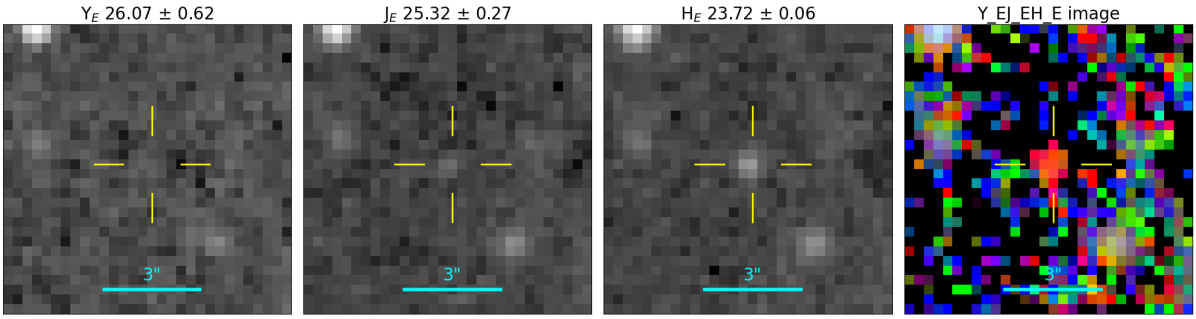
Object ID: 33385.0 | RA: 328.117 | Dec: 17.371



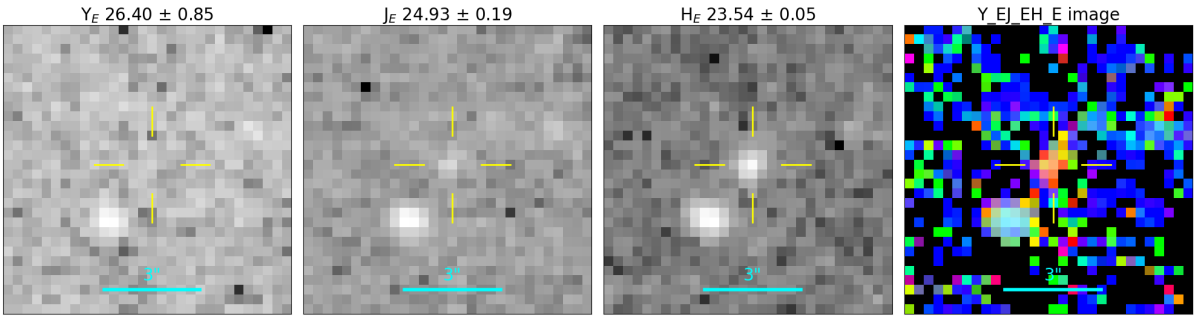
Object ID: 36779.0 | RA: 328.458 | Dec: 17.386



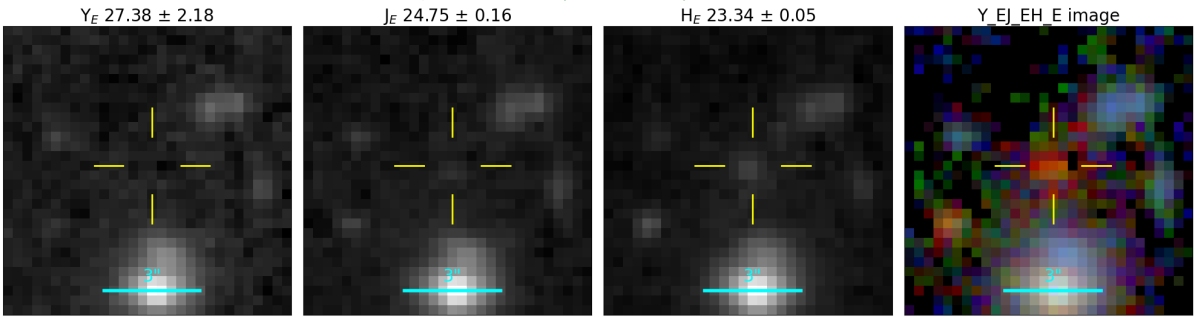
Object ID: 53958.0 | RA: 328.322 | Dec: 17.458



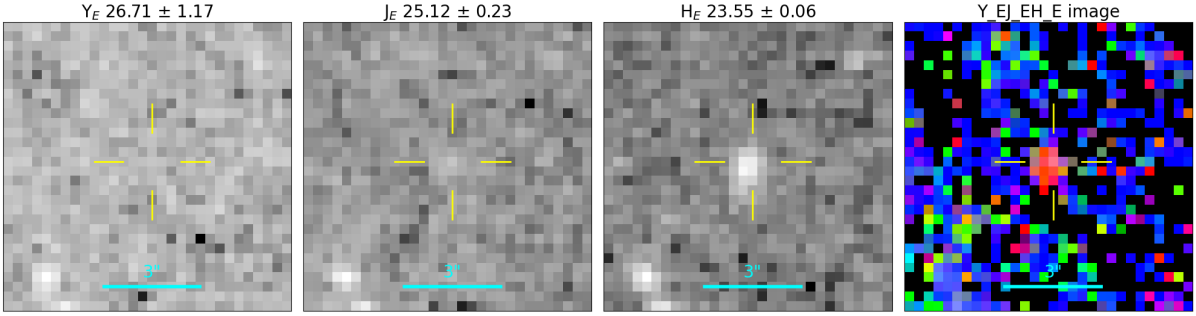
Object ID: 63917.0 | RA: 328.316 | Dec: 17.493



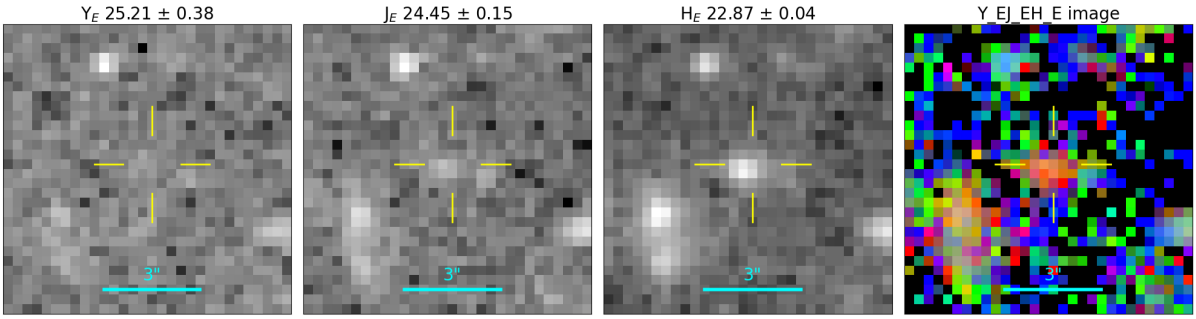
Object ID: 86772.0 | RA: 328.460 | Dec: 17.568



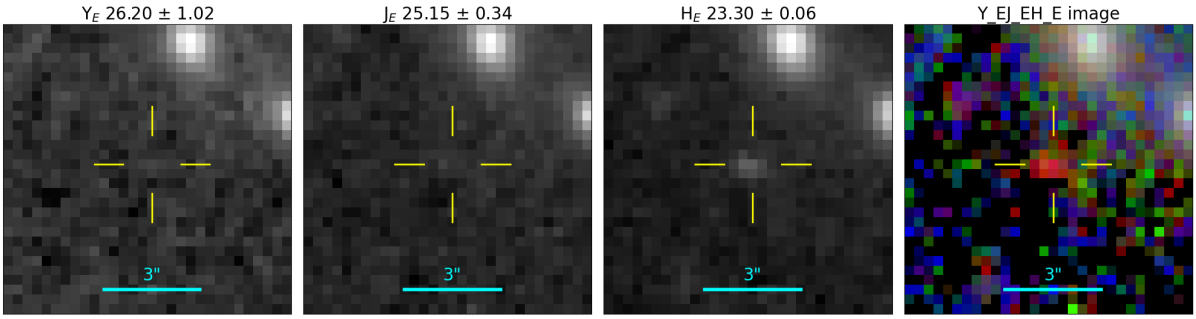
Object ID: 113553.0 | RA: 328.331 | Dec: 17.654



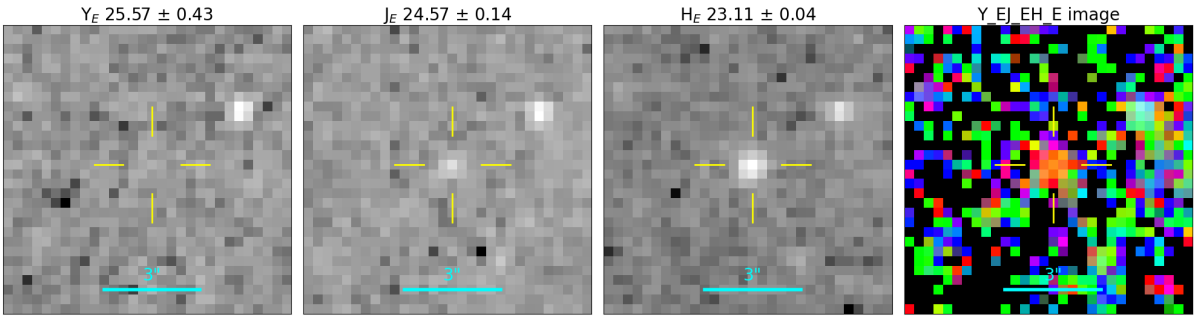
Object ID: 132996.0 | RA: 328.791 | Dec: 17.714



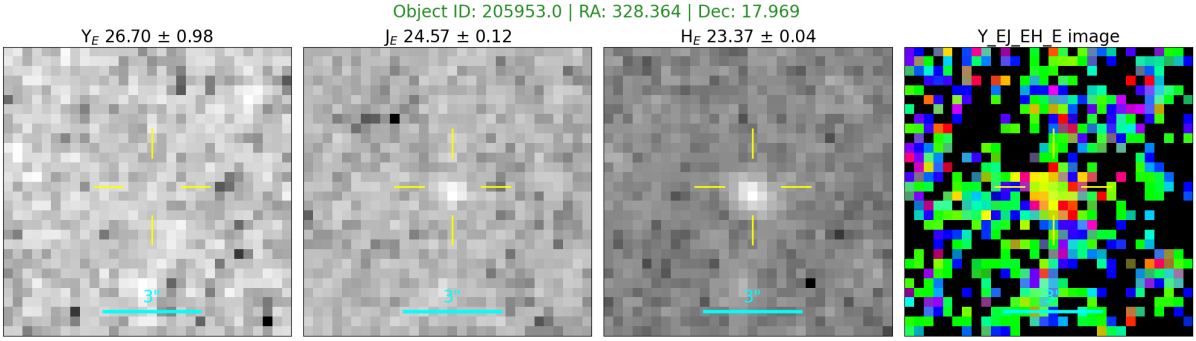
Object ID: 199819.0 | RA: 327.966 | Dec: 17.942



Object ID: 199913.0 | RA: 328.526 | Dec: 17.943



Object ID: 205953.0 | RA: 328.364 | Dec: 17.969



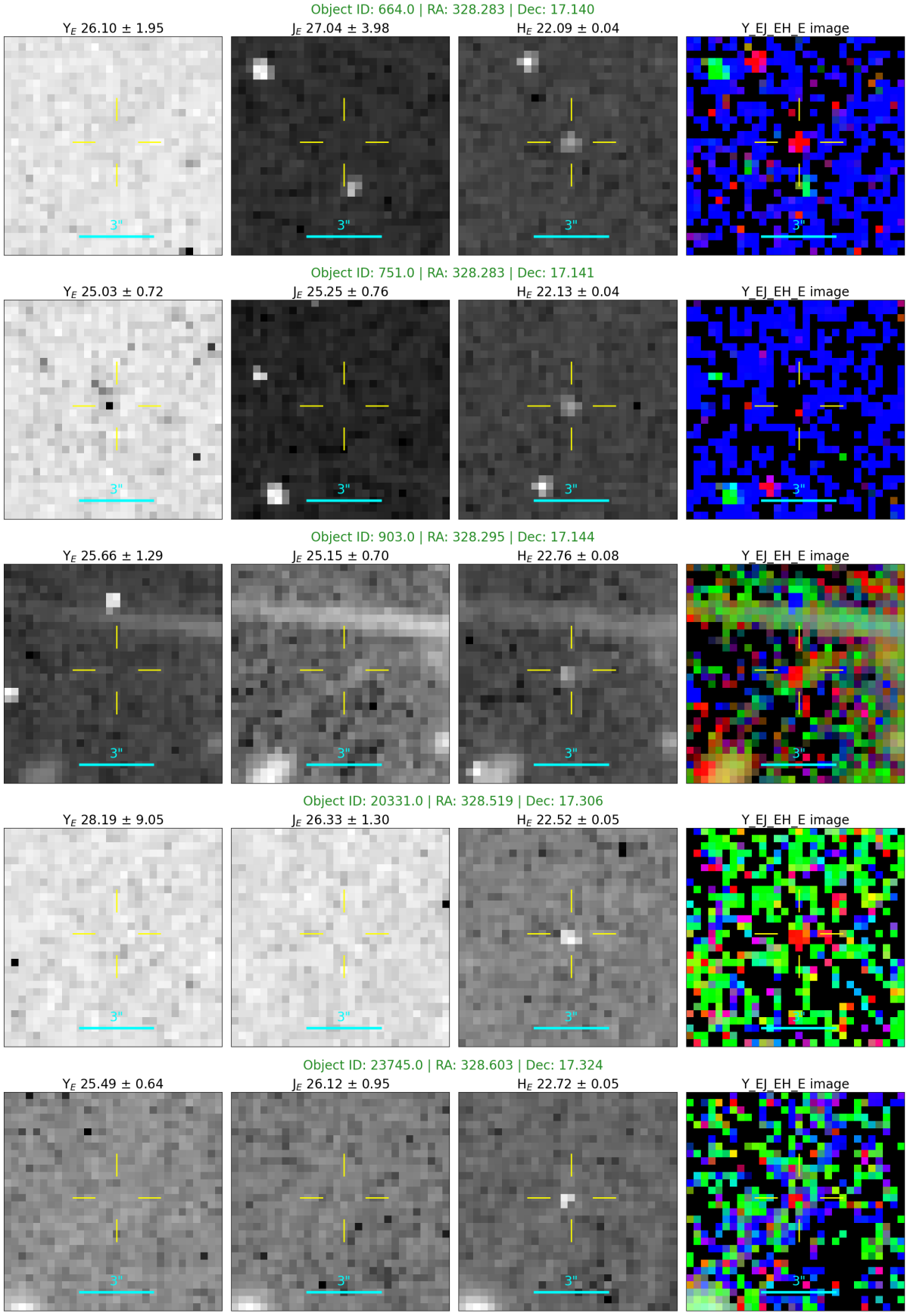


Figure 5.4 – Same as in Fig. 5.3, for dubious sources displaying a weird profile or suffering from contamination.

Five objects could not be securely identified as real astronomical sources and are dubious regarding either their profile (four of them) or the fact that they suffer from contamination (one source). They are shown in Fig. 5.4. In the following, we keep all these 17 sources in the "validated" sample.

Some artefacts are also shown below in Fig. 5.5. These are a few examples, and a small sample of them was used to illustrate the case. They are subsequently ignored as they are only contaminations

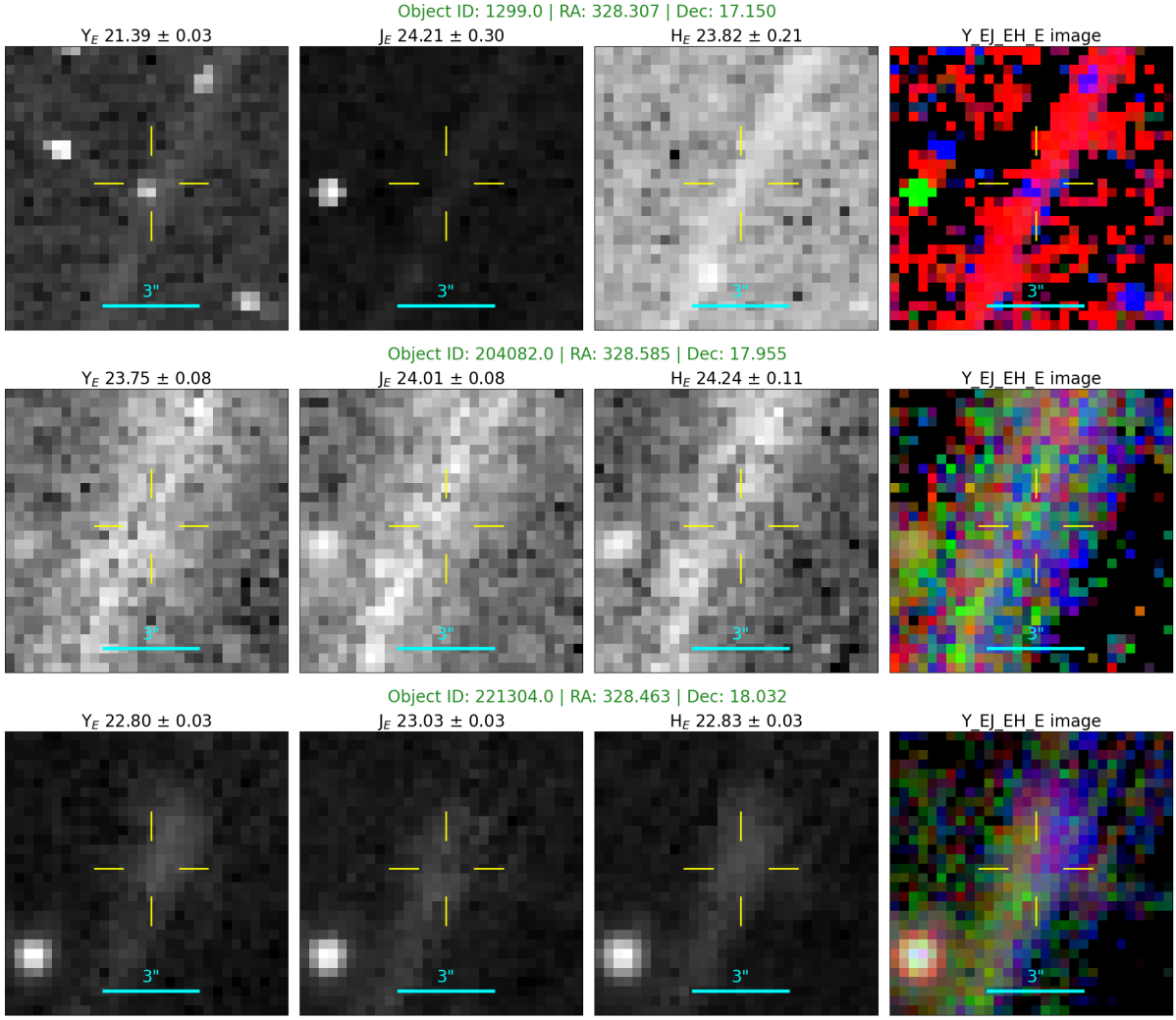


Figure 5.5 – Cutout images (30 × 30 pixels) of artefacts identified in the Euclid NISP images. These artefacts were initially selected as quasar candidates using `Owl-z` but were later classified as false detections upon visual inspection. The artefacts arise from persistence effects and other instrumental systematics. Each panel displays grayscale cutouts for Y, J, and H bands and a composite RGB image using Lupton et al. 2004. Yellow markers indicate the artefact position and a cyan scale bar represents a length of 3 arcseconds for reference.

In summary, as shown above, after visual inspection (results displayed in Figures 5.3 to

5.5), only 17 out of 158 sources in the Abell 2390 cluster passed the verification process. Some sources appear blended (e.g., ID 240134), while others seem extended, such as ID 132996 and ID 199913. Additionally, some sources like source ID 751 (qualified as "dubious") appear unusually small, likely due to a PSF issue or an RGB combination failure. Because we cannot exclude these dubious sources without additional analysis, they are kept in the sample. The overwhelming number of artefacts classified as quasars by `Owl-z` results from the input catalogues, which still needs some additional work before being reliable. Also, the difficulty in confirming the reality of the detections highlights one of the biggest shortcomings of the code—visual inspection remains essential to distinguish real sources from artefacts. Needless to say artefacts will pollute the results of any code selecting extreme objects, not only `Owl-z`. The main origin of contamination by spurious sources comes from persistence. If persistence-free images could be made available in the future, this will significantly reduce this contamination.

5.2.3 Spectroscopic confirmation of a Brown Dwarf

The source ID_222452, identified as euc328+18, was selected from the Euclid Early Release Observations (ERO) field of the Abell 2390 cluster. This extreme source, shown in Fig. 5.6, was initially identified by Weaver et al. (2024) as a high-z source based on photometric redshifts. It was also suggested to be an artefact due to persistence. When analysed with `Owl-z`, it is clearly identified as a brown dwarf (see below).

The source euc328+18_id222452 was observed using the Keck Low-Resolution Imaging Spectrometer (LRIS) on July 31, 2024, with a total integration time of 40 minutes (eight exposures of five minutes each). As shown in Fig. 5.7, the resulting spectrum has a low SNR, making the identification challenging. To assess the reliability of the identification as a brown dwarf, the spectrum has been processed using a Savitzky-Golay filter from the SciPy library to smooth the data, with three levels of smoothing applied—window sizes of 100, 50, and 25 pixels. Corresponding flux error estimates have also been provided. This analysis aims to determine whether a stellar template can be fitted to the observed spectrum, aiding in the characterisation of the source.

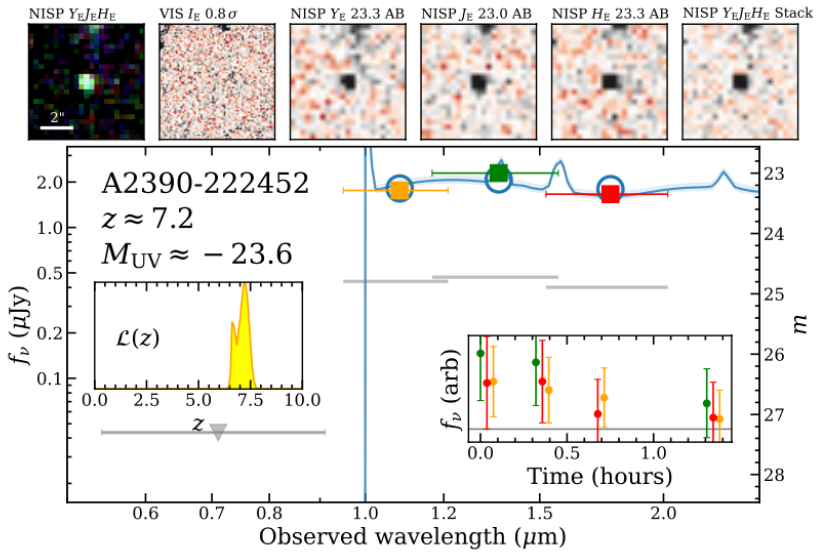


Figure 5.6 – The euc328+18_id222452 source, initially identified in Weaver et al. (2024) and suggested to be an artefact due to persistence, was later confirmed as a genuine astronomical object. Subsequent analysis using `Owl-z` classified it as a T7 brown dwarf, a result further validated as a brown dwarf through spectroscopic follow-up.

I employed a multi-step approach involving wavelength calibration, flux normalisation, and least-squares minimisation to determine the stellar template that best fits the observed spectrum. First, the observed spectrum was loaded from the Euclid Early Release Observations (ERO) data, consisting of wavelength (λ), intensity (F_λ), and associated uncertainties (σ_F). The flux values were normalised to their maximum to ensure consistent template comparison.

Given that the spectral templates had different wavelength resolutions and coverage than the observed spectrum, a calibration step was required. First, I restricted the template wavelength range to the region overlapping with the observed spectrum and applied linear regression to calibrate the template wavelengths on the observed scale. A best-fit linear relation defined this transformation:

$$\lambda'_{\text{template}} = m\lambda_{\text{template}} + b, \tag{5.1}$$

Where m and b are the slope and intercept obtained from a linear regression fit between the observed and template wavelengths.

Following wavelength calibration, I interpolated the template flux values onto the observed wavelength grid using linear interpolation. The interpolated template flux was then normalised to unity to allow for direct comparison. To account for differences in absolute flux scaling, a multiplicative scaling factor α was introduced and optimised using least-squares minimisation:

$$\alpha_{\text{best}} = \arg\min_{\alpha} \sum \left(F_{\lambda,\text{obs}} - \alpha F_{\lambda,\text{template}}\right)^2. \tag{5.2}$$

This optimisation was performed for each template in our sample, and the best-fit template was identified as minimising the least-squares residual.

Templates with insufficient wavelength overlap were automatically excluded from the fitting process. The final best-fit template was visually compared to the observed spectrum, and the calibrated fit was saved for further analysis. This method ensures a robust comparison, mitigating systematic errors due to wavelength mismatches and flux inconsistencies.

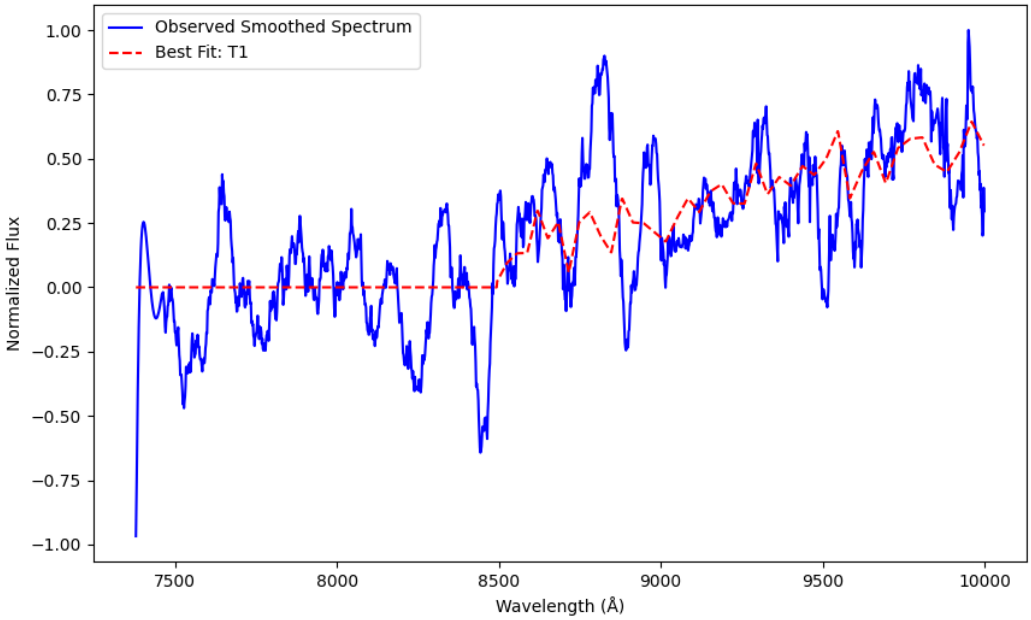


Figure 5.7 – Observed smoothed spectrum of euc328+18_id222452 (blue) compared to the best-fit template (red dashed line). The best-fit template, identified through wavelength and flux calibration followed by least-squares minimisation, corresponds to a T7 brown dwarf. The match confirms the spectral classification of the source, initially identified in the Euclid Early Release Observations (ERO) field.

The euc328+18_id222452 source was initially identified as an artefact due to persistence by Weaver et al. (2024) and collaborators, while members of the Quasar WQ team classified it as a high-redshift ($z > 6$) quasar candidate. Through careful analysis with `Owl-z`, I independently identified the object as a brown dwarf—a classification later confirmed spectroscopically. The spectral fit suggests a T1 dwarf, while `Owl-z` classifies it as a T7, likely due to the spectrum's low signal-to-noise ratio (SNR) and limited resolution. Despite these constraints, the results strongly indicate that the source is a brown dwarf rather than a quasar, demonstrating `Owl-z`'s robustness in distinguishing between faint brown dwarfs and high-redshift quasars.

5.3 Discussion

The Euclid Early Release Observations (ERO) dataset provided the first opportunity to apply `Owl-z` on real data, testing its robustness in selecting astrophysical sources. The pipeline successfully identified a variety of red objects, as confirmed by cutout images, demonstrating its ability to isolate faint, cool sources in the near-infrared. Among these, `Owl-z` identified euc328+18_id222452 as a brown dwarf, a classification later supported by spectral fitting. This result underscores the potential of `Owl-z` in distinguishing brown dwarfs from high-redshift quasars, validating its application to large-scale surveys.

Despite these successes, the analysis of ERO data also revealed certain limitations. The dataset itself is limited in sky coverage, preventing a comprehensive statistical assessment of selection efficiency or the derivation of any statistically significant quantity like an LF for z>7 quasars. Indeed, we do not expect more than one genuine quasar at z>7 in this field and less than ∼10 at z>6, roughly consistent with our findings, but it isn't easy to push the conclusions further.

Furthermore, `Owl-z` relies on the quality of the used input catalogue, and we are currently struggling to distinguish between real astrophysical sources and instrumental artefacts, as demonstrated by cases where sources initially selected as candidates were later found to be persistence residuals or spurious detections. This highlights the necessity of visual inspection as a final verification step. While `Owl-z` is an effective tool, human oversight is still needed to mitigate contamination.

The upcoming EWS data will provide a more extensive testing ground, allowing for a more refined evaluation of the pipeline's performance. A key priority will be reducing artefact contamination while maintaining sensitivity to genuine astrophysical objects. With improvements in classification algorithms and additional validation through spectroscopic follow-up, I anticipate further enhancing `Owl-z`'s reliability in future applications.

6 Conclusions and Perspectives

Contents

6.1 Conclusions 124
6.2 Technical perspectives 125
6.3 Scientific perspectives 127

6.1 Conclusions

This thesis was conducted within the framework of the Euclid Consortium. I present the code `Owl-z`, a Bayesian model comparison tool for selecting high-redshift quasars ($z > 7$) in large photometric surveys. I have optimised and used this tool in preparation for the Euclid Wide Survey (EWS). Still, it can be easily adapted to similar photometric surveys, as shown in the thesis. The performance of `Owl-z` has been estimated based on the classification efficiency using Euclid simulated data and real spectroscopically confirmed quasars. I summarise below the key conclusions of this work:

— **`Owl-z` is a robust and flexible tool:** It effectively classifies sources as high-redshift quasars or contaminants. The Bayesian framework allows for a probabilistic classification, making it adaptable to datasets and configurations beyond the Euclid Wide Survey and the search for high-redshift quasars.

— **Re-identification of known high-redshift quasars:** `Owl-z` successfully re-identified all spectroscopically confirmed quasars at $z > 7$, demonstrating its capability to work with different photometric catalogues and proving its reliability for selecting high-redshift quasar candidates.

— **Performance assessment:** Using Euclid simulated datasets, we evaluated `Owl-z` performance across redshifts $7 \leq z \leq 12$ and magnitudes $18 \leq H_E \leq 24.5$ in the EWS. The tool achieves high completeness and purity for bright quasars ($H_E < 22$), ensuring a high success rate for spectroscopic confirmation of rare, luminous quasars. However, due to contamination and lower signal-to-noise ratios, performance decreases at fainter magnitudes and higher redshifts.

— **Optimisation of the selection thresholds:** `Owl-z` allows for flexible post-processing selection of quasar candidates by tuning the probability threshold. We found that lower thresholds maximise completeness for bright sources ($H_E < 21$), while for fainter sources ($H_E > 21$), a higher threshold is recommended to increase purity. Implementing an adaptive threshold strategy optimises the F-measure, optimising the performance metrics.

— **Uniform performance across the EWS footprint:** The performance of `Owl-z` remains consistent across different Galactic coordinates and varying signal-to-noise

ratios within the EWS footprint. While minor variations exist due to sky position, these do not significantly impact the performance of the quasar selection.
— **Robustness to uncertainties in contaminant populations:** We tested `Owl-z` performance under different models of stellar contaminants, including variations in the MLT star population from thin and thick Galactic discs. The results of this validation test indicate that `Owl-z` remains robust against uncertainties in modelling these populations, especially regarding the faintest and less abundant MLT dwarfs, reinforcing confidence in its application to real surveys.
— The application of `Owl-z` to the **Euclid Early Release Observations (ERO)** was an important step for validating the code on real astrophysical data. By leveraging the first reduced Euclid data, this study demonstrates the capabilities of `Owl-z` in identifying a reasonable sample of high-redshift quasar candidates and its natural limitations in handling observational artefacts. These artefacts are susceptible to contaminating all methods applied to EWS data, not only `Owl-z`.
— The selection process on the ERO data yielded a subset of sources consistent with high-redshift quasars but also highlighted `Owl-z` weakness against separating real sources from artefacts, emphasising the need for careful data reduction and visual inspection a posteriori.
— Visual inspection remains an essential step in verifying candidates due to the presence of observational systematics such as persistence and spurious detections.
— The ERO data exercise also demonstrated robustness in distinguishing high-redshift quasars from brown dwarfs, as evidenced by the spectroscopic confirmation of euc328+18_id222452 as a brown dwarf.

The small sky coverage of the ERO dataset prevents a definitive statistical assessment of selection performance; a larger dataset is needed for more reliable conclusions. The future EWS will supply a more comprehensive testing ground, allowing for a better understanding of `Owl-z`'s strengths and limitations.

6.2 Technical perspectives

The results and performance of `Owl-z` are overall satisfactory. However, several aspects could be improved, opening several perspectives for future research and implementation. The following points highlight key areas for further developments and potential applications:

— **Expansion to additional photometric datasets:** While `Owl-z` has been tested primarily on Euclid data, its framework allows for application to other surveys such as LSST, Roman, and JWST data. However, the spectral templates of brown dwarfs limit their use to the spectral domain after the K band (centred at 2.2 microns (2190 nm). Extending the analysis to these datasets would provide a more comprehensive assessment of its performance and enhance quasar selection capabilities across multiple instruments. One could achieve this improvement by including theoretical brown dwarf templates or, even better, real spectral templates from other instruments.
— **Integration of machine learning techniques:** Bayesian classification can be

complemented with deep learning approaches to refine the selection of high-redshift quasars. By incorporating convolutional neural networks or other AI-driven techniques, `Owl-z` could improve classification accuracy and optimise completeness and purity trade-offs.

— **Morphology**: `Owl-z` operates without making any assumption about the morphology of the sources. We could naively expect that quasar candidates would be point-like. Therefore, a morphological selection/prior is expected to improve the identification. However, the results using ERO data show that some of the selected sources are extended in the EWS data, possibly because the high spatial resolution achieved from space allow to resolve the host galaxies. Adding a step to check morphology could be beneficial, provided it is not too stringent and accounts for native morphology.

— **Improved contaminant modelling:** The current model accounts for MLT dwarfs and early-type galaxies as contaminants. Future refinements could explore additional sources of contamination, such as other types of galaxies, such as dusty star-forming galaxies or transient astrophysical phenomena, and brown dwarfs from the thick disk. As the surveys deepen, these contaminating populations will be harder to ignore.

— **Spectroscopic follow-up strategies:** The effectiveness of `Owl-z` could be further evaluated through dedicated spectroscopic campaigns targeting selected candidates. Observational validation would allow us to refine probability thresholds and improve selection confidence progressively.

— **Adaptation for faint quasar selection:** The current implementation of `Owl-z` performs optimally for bright quasars ($H_E < 22$). Exploring ways to enhance its capability at fainter magnitudes, potentially through hybrid statistical and machine learning methods, could extend its reach to lower-luminosity sources.

— **Incorporating additional physical constraints:** Future versions of `Owl-z` could integrate priors based on theoretical models of quasar evolution, reionisation studies, or cosmological simulations. This would improve selection accuracy by leveraging astrophysical constraints.

— Optimising photometric preselection criteria to balance completeness and purity more effectively in the presence of noise and systematic uncertainties.

— Regarding data from the Euclid Survey, it is clear that the correction for persistence artefacts and spurious detections in the catalogues would be greatly appreciated in the short term. As for now, nothing has been done to reduce contamination by non-astrophysical sources except visual inspection. A way to do that would be to develop automated procedures for identifying and flagging artefacts in input catalogues before processing with `Owl-z`.

— Closely related to the previous item, using a range of multi-wavelength cross-matching with external datasets (e.g., JWST, ground-based NIR surveys) would improve confidence in quasar classifications.

— Conducting simulations to understand better `Owl-z`'s behaviour under different noise conditions and dataset characteristics.

— Investigating improvements in redshift estimation techniques to minimise uncertain-

ties in photometric redshifts. `Owl-z` has not been designed to provide photometric redshifts despite a relatively good quality in the estimations. This could be improved by coupling a classical method to `Owl-z` in the post-process of the candidates.

Overall, `Owl-z` provides a strong tool for high-redshift quasar selection, and its continued evolution will play a key role in the upcoming large photometric surveys. The EWS will provide an unprecedented opportunity to test `Owl-z` and improve its performance, offering a more statistically robust evaluation of its efficiency and reliability. Additionally, ongoing collaboration with other large-scale surveys and spectroscopic follow-ups will further consolidate `Owl-z`'s role in extragalactic astronomy.

6.3 Scientific perspectives

Currently, the number of quasars (intrinsically bright, type 1 AGNs) spectroscopically confirmed beyond $z > 7$ is less than ten. In contrast, AGN signatures are found in low-luminosity sources discovered by JWST beyond z$\sim$10. The discovery of hundreds of (bright) quasars at redshift $z > 7$, as it is possible using EWS and similar data, would transform our understanding of the early universe and complement the findings with JWST. Below, we discuss some of the perspectives opened by this opportunity, including insights into early black hole formation, galaxy evolution, cosmic reionisation, and potential modifications of the standard cosmological model.

A large sample of high-z quasars would enable astronomers to constrain early supermassive black hole (SMBH) growth statistically. More precisely, the existence of many early quasars would challenge current **black hole seed formation theories**, mainly whether SMBHs originate from Population III star remnants (light seeds, $\sim 100 M_\odot$) or direct-collapse black holes (heavy seeds, 10^4–$10^5 M_\odot$). If quasars at $z > 7$ are too massive to be explained by Eddington-limited accretion, alternative models such as **super-Eddington accretion** or exotic scenarios (e.g., primordial black holes) would need to be considered.

Quasars at $z > 7$ provide a direct probe of early galaxy evolution. Studying their **host galaxies** would reveal how early SMBHs and galaxies co-evolve and whether the $M_\bullet - \sigma$ relation holds at high redshifts. Quasar-driven **feedback mechanisms** could be studied to determine their role in regulating early star formation and galaxy growth. Quasar spectra would provide insight into **chemical enrichment** and the presence of metals in the interstellar medium (ISM), offering clues about early star formation. All these aspects would need supplementary multi-wavelength observations to be conducted, for instance, with ALMA, SKA, and ELT/MOSAIC.

High-redshift quasars serve as probes of the intergalactic medium (IGM) during reionisation, as discussed in Sect. 1.1.6. The **Gunn-Peterson trough** in their spectra would help trace the timeline and spatial structure of reionisation, in particular for z$\sim$6-7 quasars, because the GP effect quickly saturates beyond x(HI)$\sim 10^{-3}$. A large quasar sample would enable **3D tomography** of reionisation, identifying ionised regions and their evolution over time, for instance, by combining SKA and JWST data. The proximity zones of these quasars would allow measurements of **neutral hydrogen fractions** at different epochs.

An unexpectedly high number of luminous quasars at $z > 7$ could challenge ΛCDM predictions. It may indicate the need for **new physics** in structure formation, such as

modified dark matter scenarios. Quasar clustering at high-z, only, only accessible through wide surveys (Euclid, Roman, LSST) could provide constraints on **initial density fluctuations** and the formation of large-scale structures. Also, high-resolution spectroscopy on these high-z quasars using, for instance, ELT/ANDES, could help catch variations in the fine-structure constant and might hint at **new fundamental physics**.

High-z quasar spectra offer a unique tool for mapping the early universe. Absorption lines (e.g., C IV, Mg II) reveal **early metal enrichment** and the chemical evolution of the IGM. Lyman-α forest observations allow for **mapping large-scale structure** and the distribution of neutral hydrogen (although this can also be performed with quasars and galaxies at lower redshifts). The detection of molecular gas and dust in quasar spectra could provide constraints on **early star formation environments**. For these needs, follow-up observations with ALMA and the ELT would be needed.

A large sample of $z > 7$ quasars would shape future observing strategies differently. Instruments like **JWST, ALMA, and the ELT** could perform deep spectroscopic follow-ups to study black hole and galaxy co-evolution. The 21-cm cosmology experiments (e.g., **SKA**) could cross-correlate quasar locations with neutral/ionised hydrogen maps, revealing the topology of reionisation. This will be particularly interesting if quasars are more significant contributors to the reionisation than previously expected, as JWST observations seem to indicate. If quasars extend beyond $z = 10$, future X-ray surveys like New-Athena could shed light into **early supermassive black hole growth mechanisms**.

The discovery of hundreds of quasars at $z > 7$ would be paradigm-shifting, providing crucial insights into SMBH formation, galaxy evolution, cosmic reionisation, and fundamental physics. This dataset would be a goldmine for understanding the universe in its infancy and refining our models of early structure formation. This could be within reach five years after Euclid has completed its EWS. Exciting and evolutionary times ahead.

ANNEXES

Annexe A

Cosmological concepts and tools for studying High-redshift quasars

To understand these distant quasars, cosmologists and astrophysicists employ a variety of tools, including redshift measurement, spectroscopy, and calculations of luminosity distance. Additionally, phenomena like the Gunn-Peterson trough and the CMB serve as essential markers in this research.

Redshift as a Measure of Distance and Time

One of the primary tools in studying high-redshift quasars is *redshift* itself. In cosmology, redshift (z) is a measure of how much the wavelength of light emitted by a distant object is stretched due to the expansion of the universe. The redshift-distance relation allows us to use high-redshift quasars as tracers of the early universe.

The redshift z is defined by the equation:

$$1 + z = \frac{\lambda_{\text{observed}}}{\lambda_{\text{emitted}}}, \tag{0.1}$$

where $\lambda_{\text{observed}}$ is the wavelength of the light received on Earth, and λ_{emitted} is the wavelength of the light when it was emitted from the source (e.g. quasar in the case of this study).

The higher the redshift, the farther back in time we are observing, with $z = 6$ quasars corresponding to a time when the universe was less than a billion years old. Redshift measurements allow us to place quasars in the context of cosmic history, gaining insight into early structure formation, the ionisation state of the universe, and more.

Spectral Analysis

Spectroscopy: By dispersing the quasar's light (emitted flux) into its constituent wavelengths, we can extract information about its composition, orientation,dynamics, and environment. The spectral lines in a quasar's spectrum provide key insights into its properties, such as:

— **Emission Lines:** High-redshift quasars exhibit strong emission lines from elements such as hydrogen, carbon, magnesium, and iron. These lines can be redshifted by the expanding universe, making them indicators of the quasar's redshift and velocity.

— **Absorption Lines:** Absorption lines, particularly those from intergalactic hydrogen, provide information about the gas between us and the quasar.

— **The Lyman-α Line:** The Lyman-alpha (Ly-α) line, corresponding to transitions in neutral hydrogen at a rest wavelength of 1216 Å, is important for studying the quasar's environment and the intergalactic medium.

Luminosity Distance

The *luminosity distance* is used to quantify the distance to high-redshift quasars. The relationship between the observed flux and luminosity (intrinsic property) of a quasar is given by the luminosity distance D_L, which is defined as:

$$D_L = \sqrt{\frac{L}{4\pi F}}, \tag{0.2}$$

where L is the luminosity of the quasar, and F is the observed flux.

The luminosity distance is related to redshift through the geometry of the universe and can be affected by factors such as the universe's expansion rate (Hubble parameter), dark energy, and matter density. By measuring the luminosity distance to high-redshift quasars, researchers can test cosmological models and refine parameters like the Hubble constant H_0.

The Comoving Distance

The comoving distance, D_c, between two objects in the universe can be expressed as:

$$D_c = c \int_{t_e}^{t_0} \frac{dt}{a(t)}, \tag{0.3}$$

where:
— c is the speed of light,
— t_e is the time of emission of light,
— t_0 is the current time,
— $a(t)$ is the scale factor, a function describing the expansion of the universe.
The scale factor is related to the redshift z by:

$$a(t) = \frac{1}{1+z}. \tag{0.4}$$

Comoving Volume

The comoving volume V_c is often used in cosmological studies to describe the fixed coordinate volume that expands with the universe. In a flat universe, it is given by:

$$V_c = \frac{4\pi}{3} D_c^3. \tag{0.5}$$

For non-flat geometries, curvature factors need to be incorporated into the calculation of D_c, altering the volume relation.

Magnitude and K-correction

In observational astronomy, the concepts of magnitude and K-correction are fundamental for understanding the brightness of celestial objects as observed from Earth. These measurements help us compare the intrinsic luminosities of objects across different distances and redshifts. Here, we provide a detailed overview of apparent magnitude, observed magnitude, and the K-correction, following the formalism described by Hogg et al. 2002.

Apparent Magnitude

The *apparent magnitude*, m, quantifies the brightness of an astronomical object as observed from Earth. It is defined relative to a reference flux F_0 by:

$$m = -2.5 \log_{10} \left(\frac{F_{\text{obs}}}{F_0} \right), \tag{0.6}$$

where:

— F_{obs} is the observed flux of the object, measured in units such as erg s^{-1} cm^{-2}.

— F_0 is the zero-point flux, a calibration constant chosen to define the magnitude scale.

Apparent magnitude depends on both the intrinsic luminosity of the source and the effects of distance and redshift.

Absolute Magnitude

To compare intrinsic brightness, we use the *absolute magnitude*, M, which represents the apparent magnitude the object would have if it were placed at a standard distance of 10 parsecs. The relationship between m and M is:

$$m - M = 5 \log_{10}(D_L) - 5, \tag{0.7}$$

where D_L is the luminosity distance in parsecs. The luminosity distance accounts for the effects of cosmic expansion and redshift as seen in section 6.3. Using the formula give in Hogg et al. 2002, the absolute magnitude in the emitted-frame bandpass Q can be calculated using:

$$M_Q = -2.5 \log_{10} \left[\frac{\int \frac{d\nu_e}{\nu_e} \frac{L_\nu(\nu_e)}{4\pi (10\,\text{pc})^2} Q(\nu_e)}{\int \frac{d\nu_e}{\nu_e} g_Q^\nu(\nu_e) Q(\nu_e)} \right] \tag{0.8}$$

where:

— M_Q: The absolute magnitude of the source measured in the emitted-frame bandpass Q.

— $L_\nu(\nu_e)$: The spectral density of luminosity (energy emitted per unit time per unit frequency) of the source at frequency ν_e.

— $Q(\nu_e)$: The bandpass function describing the response of the detector in the emitted-frame bandpass Q at frequency ν_e. It is analogous to $R(\nu_o)$ but applies in the emitted frame.

— $g_Q^\nu(\nu_e)$: The spectral density of flux for the zero-magnitude or standard source in the Q band.

— ν_e: The emitted frequency (rest-frame frequency) of the source.
— $\int \frac{d\nu_e}{\nu_e}$: Integration over the emitted frequency range weighted by $1/\nu_e$.
— $4\pi(10\,\mathrm{pc})^2$: The surface area of a sphere with a radius of 10 parsecs (pc). This corresponds to the distance at which the absolute magnitude M_Q is defined.

Observed Magnitude and Redshift Effects

At high redshifts, the observed magnitude is influenced not only by the luminosity distance but also by the shift of the source's spectrum. This shift causes photons to move into different observational filters, the higher the redshift the redder the source which implies a shift to the right of the spectrim to higher wavelenghts. This affects the measured flux. This effect is corrected using the *K-correction*. Using the the equation formulated in Hogg et al. 2002, the observed magnitude can be calculated as follows:

$$m_R = -2.5 \log_{10} \left[\frac{\int \frac{d\nu_o}{\nu_o} f_\nu(\nu_o) R(\nu_o)}{\int \frac{d\nu_o}{\nu_o} g_R^\nu(\nu_o) R(\nu_o)} \right] \tag{0.9}$$

where:
— m_R: The apparent magnitude of the source measured through the photometric bandpass R.
— $f_\nu(\nu_o)$: The spectral density of flux (energy per unit time, per unit area, per unit frequency) observed at frequency ν_o.
— $R(\nu_o)$: The bandpass function describing the detector's response at each frequency ν_o. It represents the mean contribution of a photon of frequency ν_o to the detector's signal.
— $g_R^\nu(\nu_o)$: The spectral density of flux for the zero-magnitude or standard source in the R band. For Vega-relative magnitudes, this is Vega (or a weighted sum of A0 stars). For AB magnitudes, it is a hypothetical constant source with $g_{\mathrm{AB}}^\nu(\nu) = 3631\,\mathrm{Jy}$.
— ν_o: The observed frequency at which the measurements are taken.
— $\int \frac{d\nu_o}{\nu_o}$: Integration over the observed frequency range weighted by $1/\nu_o$.

K-correction

The K-correction accounts for the differences between the rest-frame and observed-frame photometric systems due to redshift. It adjusts the observed apparent magnitude to the rest-frame magnitude. The corrected apparent magnitude m_{rest} is given by:

$$m_{\mathrm{rest}} = m_{\mathrm{R}} - K(z), \tag{0.10}$$

where $K(z)$ is the K-correction, a function of redshift z, and depends on the spectral energy distribution (SED) of the source and the filter response. The K-correction is calculated as:

$$K(z) = 2.5 \log_{10} \left(\frac{\int R(\nu) S(\nu)\, d\nu}{\int R(\nu/(1+z)) S(\nu)\, d\nu} \right), \tag{0.11}$$

where:
— $R(\nu)$ is the response function of the observational filter,
— $S(\nu)$ is the spectral energy distribution (SED) of the source,
— ν is the frequency of light.

Significance in Astronomical Observations

The K-correction enables us to compare the intrinsic luminosities of galaxies and other sources across a range of redshifts. Without this correction, differences in observed fluxes would combine intrinsic properties with redshift-induced effects. This is particularly important in large-scale surveys and studies of galaxy evolution.

Hogg et al. 2002 emphasizes the importance of precise flux calibration and detailed knowledge of the SED when applying K-corrections, especially for high-redshift sources. These K-corrections are essential for ensuring consistency in cosmological analyses and understanding the history of star formation and galaxy growth.

Gunn-Peterson Trough

The *Gunn-Peterson trough* is a feature in the spectra of high-redshift quasars that reveals critical information about the epoch of reionisation Gunn &Peterson (1965b). This through is due to the presence of neutral hydrogen in the intergalactic medium (IGM) that absorbs photons in the Lyman-alpha transition, leaving a complete absorption feature (a trough) in the spectrum at wavelengths shorter than the Lyman-alpha line.

In the spectra of quasars with redshifts $z \gtrsim 6$, we begin to observe the Gunn-Peterson trough, indicating the presence of significant amounts of neutral hydrogen in the universe at that time. This tells us that the universe was transitioning from a neutral to an ionised state (the epoch of reionisation), shedding light on one of the most important phases in cosmic history.

The Gunn-Peterson can be described by its optical depth τ_{GP} which represents the optical depth of neutral hydrogen in the IGM to the Lyman-α transition (Gunn &Peterson 1965b), it is given by:

$$\tau_{\mathrm{GP}} = \frac{\pi e^2}{m_e c} \lambda_\alpha f_{12} n_H, \tag{0.12}$$

where n_H is the density of neutral hydrogen, and f_{12} is the oscillator strength of the Lyman-α transition, m_e the mass of an electron, approximately 9.109^{-31} kg, c the speed of light ($3{\times}10^8$ m/s) and λ_α the wavelength of the Lyman-α transition.

Quasar Luminosity and Black Hole Masses

High-redshift quasars are typically powered by supermassive black holes (SMBHs) with masses ranging from millions to billions of solar masses. Understanding the growth of these black holes over cosmic time is crucial for understanding the formation of large-scale structures in the universe.

The luminosity of a quasar is directly tied to the accretion of material onto the black hole, with higher luminosity implying a higher accretion rate. The *Eddington luminosity*

(Eddington 1926), which represents the maximum luminosity where the outward radiation pressure balances the inward gravitational pull, is often used to estimate the masses of black holes:

$$L_{\text{Edd}} = \frac{4\pi G M m_p c}{\sigma_T}, \tag{0.13}$$

where G is the gravitational constant, M is the black hole mass, m_p is the proton mass, c the speed of light (3×10^8 m/s), and σ_T is the Thomson scattering cross-section.

High-redshift quasars help us explore the formation and early growth of SMBHs in the first billion years of the universe, which remains one of the great mysteries in cosmology.

Opacity and Reionisation

Opacity is a fundamental concept in the study of high-redshift quasars, as it governs how light from these distant objects interacts with the IGM. The absorption of light by neutral hydrogen, as seen in the Gunn-Peterson trough and the Lyman-alpha forest, provides essential clues about the structure and evolution of the IGM, the epoch of reionisation, and the large-scale geometry of the universe.

During reionisation, the IGM was opaque to ultraviolet light due to the high density of neutral hydrogen. As reionisation progressed and more hydrogen became ionised, the opacity of the IGM decreased, allowing light from high-redshift quasars to travel more freely through space. This reduction in opacity can be traced through the evolution of the Lyman-alpha forest and the Gunn-Peterson trough.

Annexe B

```
workdir : /home/mezziati/Documents/OWL-Z_V0.1 # Parent path

#QSO SEDs
templates_QSO :   /data/model_data/QSO/templates.list #list
templates_QSO_dir :   /data/model_data/QSO/ # SEDs

#GAL SEDs
    templates_GAL :  /data/model_data/GLX/BC03/log_temp.txt #list
templates_GAL_dir :   /data/model_data/GLX/BC03/    # SEDs

# MLT SEDs
templates_MLT: /data/model_data/LT/MLT_star_templates/
templates_M_list: /data/model_data/LT/MLT_star_templates/Mstars.txt
templates_L_list: /data/model_data/LT/MLT_star_templates/Lstars.txt
templates_T_list: /data/model_data/LT/MLT_star_templates/Tstars.txt

# Absolute magnitudes with K-correction
MABS_QSO_list :    /results/colors/log_qso.txt
MABS_GAL_list :    /results/colors/log_gal.txt

# liste of the used spectral types
SptM:  /data/model_data/LT/spt_M.dat
SptL: /data/model_data/LT/spt_L.dat
SptT:  /data/model_data/LT/spt_T.dat

#local densities
M_dens:  /data/model_data/LT/spt_M.txt
L_dens: /data/model_data/LT/spt_L.txt
T_dens:  /data/model_data/LT/spt_T.txt

#Filters
filters_list :  /data/filters/euclid/list.txt
filters_dir :  /data/filters/euclid/
reference_filt : 3 # reference filter  index

#results paths
colors_result_path:  /results/colors/
probas_result_path:  /results/probas/

# candidates
candidates: /data/candidates_files/

# model parameters
LIM_mag : [25.5,24,24,24] # limiting mangitudes
```

```
redshift : # quasar redshift
  - 6
  - 10
  - 0.1

redshift_g : # galaxy redshift range
  - 1
  - 2.1
  - 0.1

magY: # model mangitude rnage
  - 18
  - 25
  - 0.1

# Miscellaneous parameters
light_speed : 3.0e18
cosmolgy : [70.0,0.3]
EBmV : 0.02
hR : 2250. # in parsecs
hZ : 1000 # in parsecs
Z_sol : 27. # in parsecs
r_sol : 8600. # in parsecs
SW : 1. # surface
ZP : -48.572 # magnitude AB zero point
const : 3631e-23 # hypothetical constant source for zero point
```

Astronomy & Astrophysics manuscript no. output
February 11, 2025
©ESO 2025

OWL-z: a Bayesian tool to select $z \gtrsim 7$ quasars

M. Ezziati[1]⋆, R. Pello[1], J.-G. Cuby[1,4], P. Pudlo[2], F.-X. Dupé[3],J.-C. Lambert[1], J.-C. Cuillandre[5], O. Ilbert[1], S. de la Torre[1], S. Arnouts[1], E. Jullo[1],D. Yang[6]

(Affiliations can be found after the references)

February 11, 2025

ABSTRACT

This paper presents OWL-z, a Bayesian code aiming at identifying $z \geq 7$ quasars in wide field optical and near-infrared surveys. By construction, the code can also be used to select objects that contaminate the high-z quasar population, i.e. brown dwarfs and early-type galaxies at intermediate redshifts. The code can be adapted for the selection of high-redshift galaxies, and although it has been tuned to the *Euclid* Wide Survey, it can be easily adapted to other photometric surveys. The code input data are the object's photometric data and its galactic longitude and latitude, and the code output data are the probabilities of the modelled populations of high-redshift quasars, brown dwarfs and early-type galaxies at intermediate redshift. As part of the validation, OWL-z could re-identify all spectroscopically confirmed quasars at *zgeq*7, demonstrating the code's versatility in applying to different photometric catalogues. The performance of OWL-z, based on a metric combining completeness and purity called F-measure, is analysed in the case of *Euclid* using simulated data in a wide range of redshifts ($7 \leq z \leq 12$) and H-band *Euclid* magnitudes ($18 \leq H_{\rm E} \leq 24.5$). The results show that OWL-z reaches full performance for bright sources ($H_{\rm E} \lessapprox 22$), somewhat independently of the redshift. We show that the probability threshold used to select promising quasar candidates can be adjusted after processing to fine-tune the F-measure value of candidates depending on their magnitude and redshift estimates. We show that for objects brighter than about two magnitudes above the survey detection limit, OWL-z provides a good classification that will facilitate the optimization of photometric and spectroscopic confirmation campaigns. In conclusion, OWL-z is a powerful public tool to help select high-redshift quasars, brown dwarfs or early-type galaxies at intermediate redshifts in *Euclid* or other wide-field surveys.

Key words. Methods: statistical; Surveys; Cosmology: observations; reionization; quasars: general

1. Introduction

High-redshift quasars are significant in illuminating the early Universe, shedding light on the genesis of initial galaxies and black holes. The luminosity of these cosmic beacons allows for unprecedented insights into the evolution and patchiness of neutral hydrogen and its re-ionization, particularly during the epoch of reionization (EoR) (see e.g. Fan et al. 2006; Becker et al. 2015). Beyond tracing the EoR, the identification of supermassive black holes (SMBHs) using quasars at high redshifts challenges conventional models, prompting investigations into alternative scenarios (Bennett et al. 2024). In particular, the identification and study of $z > 7$ quasars are challenging due to their ambiguous nature and the limited observations/identifications of such objects in the early Universe (see e.g. Bañados et al. 2018a). Also, the recent discoveries of young quasars with small Lyman-alpha near zones add complexity, opening avenues for refining our understanding of early cosmic phenomena. Spectroscopic observations with the James Webb Space Telescope (JWST) have uncovered AGN signatures in the spectra of the most distant galaxies known today (Maiolino et al. 2024), providing further insights on their environment and formation history (Scholtz et al. 2024).

Despite their high intrinsic luminosity, high-redshift quasars appear faint at $z > 7$, and their photometric selection on wide cosmological surveys is particularly challenging. Indeed, their red colours, often used to identify them with the appropriate colour-cuts or photometric redshifts, are subject to severe contamination by brown dwarfs and intermediate-redshift galaxies. With the arrival of massive cosmological surveys, such as *Euclid* (Laureijs et al. 2011; Euclid Collaboration et al. 2024), the Nancy Grace Roman Space Telescope (Akeson et al. 2019) and the Large Synoptic Survey Telescope (LSST) (Ivezić et al. 2019a), it becomes possible to identify significant samples of quasars at high redshifts, which will greatly improve our understanding of galaxy and SMBH formation mechanisms. However, for this, it is essential to build efficient and optimised methods and tools. This is precisely the purpose of this work.

The method introduced here, and its associated code OWL-z, is a Bayesian comparison model aiming at optimising the statistical analysis in such a way that the probability of detecting high-redshift quasars is enhanced with respect to the classical colour-cuts method. Our approach is directly inspired by the Bayesian method developed by Mortlock et al. 2012 and Euclid Collaboration: Barnett et al. 2019 for the search of quasars at high redshifts.

Our method is one of the methods based on Bayesian analysis, along with Euclid Collaboration: Barnett et al. 2019, but there are other probabilistic methods, such as the one presented in Nanni et al. 2022, which uses a probabilistic classification approach using density estimation in flux ratio space, specifically using the extreme deconvolution (XD) technique to accurately model the density distribution of high redshift quasar and contaminant populations (Bovy et al. 2011). There are also other efficient colour-cutting methods paired with radio detections, such as those used in Belladitta et al. 2020, Bañados et al. 2014 and Bañados et al. 2016a, which proved to be very efficient. Last but not least, there are more recent methods that use Random Forests, a supervised machine learning approach, to effectively select high redshift quasars using photometric data from surveys such as Pan-STARRS and WISE (Wenzl et al. 2021).

⋆ Corresponding author, e-mail: meriam.ezziati@lam.fr

OWL-z also offers a flexible, fast, and efficient tool for analysing large photometric data sets, looking for $z > 7$ quasars with high performance in terms of completeness and purity of the extracted samples. From a technical point of view, OWL-z is an open-source code developed specifically for the selection of high-redshift quasars in wide field near-infrared surveys such as the *Euclid* Wide Survey (hereafter EWS) (Euclid Collaboration: Scaramella et al. 2022), but it is adaptable to tackle various other surveys, as also shown in this work.

This paper is organized as follows. In Sect. 2, we present the probabilistic method used for the identification of high-z quasars, the primary objective being the Bayesian selection and classification of each source as either a high-z quasar, an intermediate-redshift contaminant galaxy, or a dwarf star. The detailed modelling performed for each one of these populations is also presented in this section. A brief technical description of OWL-z is presented in Sect. 3. Sect. 4 is devoted to the validation of OWL-z using two different approaches: the successful re-identification of known and spectroscopically-confirmed quasars at $z > 7$ selected from near-IR surveys and simulations of the EWS allowing us to quantify the performance of the code based on the measurement of the completeness and purity of the extracted samples. In Sect. 5, we discuss the influence of several important parameters of the method on the performance of OWL-z, such as the threshold used for the selection of high-z quasars, and we provide some guidelines to optimize their selection and follow-up photometric or spectroscopic campaigns. The summary of conclusions is presented in Sect. 6.

Throughout this paper, the following cosmological parameters have been adopted: $H_0 = 70$ km s^{-1} Mpc^{-1}, $\Omega_m = 0.3$ and $\Omega_\lambda = 0.7$. All magnitudes are given in the AB system (Oke & Gunn 1983).

2. Probabilistic selection of sources at high redshift

This section presents the probabilistic selection method for identifying high-redshift quasars in the EWS. The method can be adapted to other wide-field surveys and to the selection of other high-redshift astronomical sources. The objective of this method is to identify and select a distinct set of potentially high-redshift sources, referred to as candidates, by effectively distinguishing them from low-redshift sources designated as contaminants. Our approach is based on the Bayesian method developed by Mortlock et al. 2012 and Euclid Collaboration: Barnett et al. 2019 for the search of quasars at high redshift.

The primary spectral signature of objects with a redshift exceeding 7 ($z > 7$) is the quasi-absence of flux below the Lyman alpha line, which is a consequence of the combined effects of the Lyman forest and the Gunn-Peterson trough. It is thus possible to identify objects with a redshift of more than 7 by the lack of signal in the optical domain below $\approx 1\mu$m (0.97 μm). This allows us to restrict the modelling of high-redshift quasars and their contaminants to objects detected in near-infrared bands and pre-selected on the basis of the absence or near-absence of detection in optical bands. This work can be compared with others, such as Euclid Collaboration: Barnett et al. 2019 and Pipien et al. 2018b. For details, please see Sect. 5.5

We employ Bayesian selection methodology to ascertain the posterior probability that a specific observed astronomical source, initially identified as non-detected in optical bands, is classified as a high-redshift quasar or as a contaminant. This is achieved by considering the prior surface density of the quasars and contaminants. Contaminant populations of high-redshift galaxies and quasars in near-infrared and mid-infrared bands are well known (see e.g. Mortlock et al. 2012; Euclid Collaboration: Barnett et al. 2019) and consist of early-type galaxies at intermediate redshift ($1 \lesssim z \lesssim 2$) (see e.g. Euclid Collaboration: van Mierlo et al. 2022) and low mass dwarf stars of spectral types late-M, L or T (MLT hereafter) (Stern et al. 2007; Caballero et al. 2008; Wilkins et al. 2014; Hainline et al. 2024a). The difficulty in distinguishing high-redshift quasars from the above-mentioned contaminants is illustrated in Fig. 1, showing the comparison of near-infrared (NIR) and optical-NIR colours for the three populations considered in this work using *Euclid* filters (Euclid Collaboration et al. 2022). Analysis of this figure highlights the distinct colour characteristics of these populations in the $Y_\mathrm{E} - J_\mathrm{E}$ colour space, which is key to the quasar selection process. The top panel of the figure illustrates that at higher redshifts ($z > 8$), quasars are significantly redder than the contaminating populations, allowing for effective separation based on their $Y_\mathrm{E} - J_\mathrm{E}$ colours. However, the analysis also notes that at lower redshifts ($7 < z < 8$), the NIR broadband colours of quasars and contaminants overlap more, making it difficult to distinguish between them without deep complementary data, especially in the LSST z-band (lower panel).

2.1. Principles

In this section, we present the Bayesian formalism in a general context and then apply it in more detail to each of the populations. Bayesian model selection (Robert 2007) is a relevant method to establish the nature of a source in a survey given photometric data (see Mortlock et al. 2012, and the references therein). The posterior probability that a given source is of a type $t \in \mathcal{T}$, given the data $\mathbf{D}$, is driven by two terms: the prior probability $\mathbb{P}(t)$ of the source being of type t, and the integrated likelihood or the evidence of the data given the type of source, $\mathbb{P}(\mathbf{D}|t)$. The latter is the marginal likelihood of the data, which is the integral of the likelihood of the data given the parameters of the model, $\mathbb{P}(\mathbf{D}|\boldsymbol{\theta}_t, t)$, with respect to the prior distribution of the parameters, $\mathbb{P}(\boldsymbol{\theta}_t|t)$. The marginal likelihood is the key quantity in Bayesian model selection: Unlike methods based on best-fit parameters and type, it accounts for the diversity of photometric data of a given type. The posterior probability of a source being of type t is then given by Bayes' theorem:

$$\mathbb{P}(t|\mathbf{D}) = \frac{W(t|\mathbf{D})}{\sum_{t' \in \mathcal{T}} W(t'|\mathbf{D})}, \tag{1}$$

where $W(t|\mathbf{D}) = \mathbb{P}(t)\mathbb{P}(\mathbf{D}|t)$ is the weighted evidence of the model t given the data $\mathbf{D}$. The weighted evidence of type t is given by

$$\begin{aligned} W(t|\mathbf{D}) = \mathbb{P}(t)\mathbb{P}(\mathbf{D}|t) &= \mathbb{P}(t) \int \mathbb{P}(\boldsymbol{\theta}_t|t)\mathbb{P}(\mathbf{D}|\boldsymbol{\theta}_t, t)\mathrm{d}\boldsymbol{\theta}_t \\ &= \int \rho_t(\boldsymbol{\theta}_t)\mathbb{P}(\mathbf{D}|\boldsymbol{\theta}_t, t)\mathrm{d}\boldsymbol{\theta}_t, \end{aligned} \tag{2}$$

where $\boldsymbol{\theta}_t$ is the vector of parameters of t-type object and $\rho_t(\boldsymbol{\theta}_t)$ is the prior value of surface density of detected objects of type t and parameters $\boldsymbol{\theta}_t$ in the sky. A photometric dataset is a set $\widehat{\mathbf{F}} = (\widehat{F}_1, \ldots, \widehat{F}_N)$ of N measurements of the flux of a source in N photometric bands, and their standard error $\widehat{\boldsymbol{\sigma}} = (\widehat{\sigma}_1, \ldots, \widehat{\sigma}_N)$. The statistical model of each type of source is built on a set of spectral energy densities (SEDs) of n_t templates. Given the i-th SED of type t, we can compute expected fluxes in each band, $\mathbf{F}_{t,i}(\boldsymbol{\theta}_t) = (F_{t,i,1}(\boldsymbol{\theta}_t), \ldots, F_{t,i,N}(\boldsymbol{\theta}_t))$ at value $\boldsymbol{\theta}_t$ of the parameter. The likelihood of the data given $\boldsymbol{\theta}_t$ and t is set as the mixture of

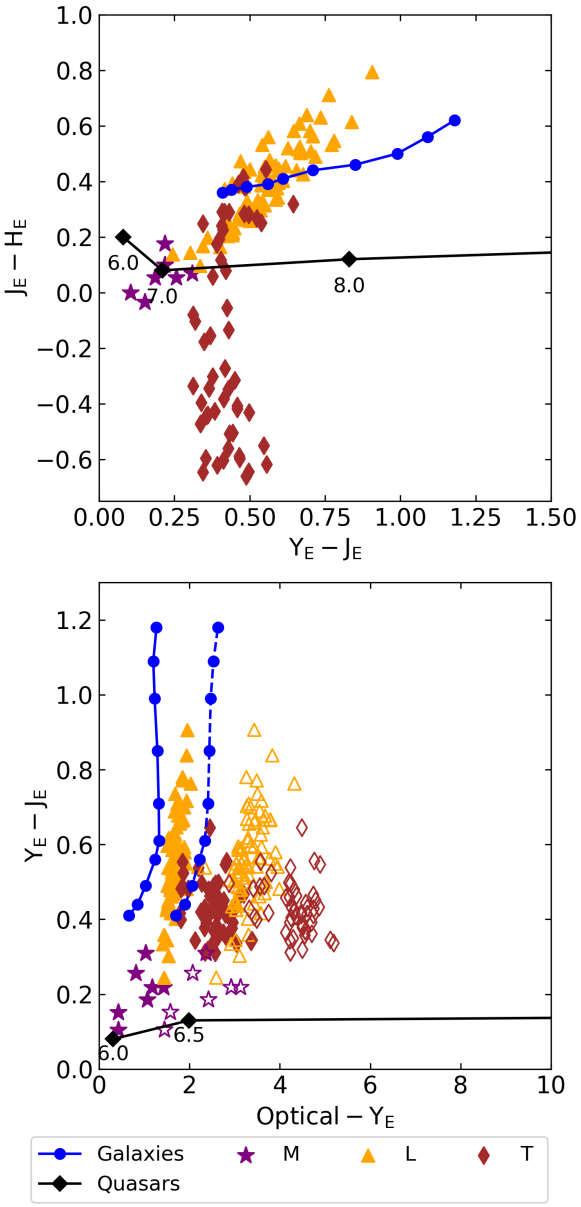


Fig. 1: NIR (top) and Optical-NIR (bottom) colours for the three classes of objects considered in this work, using *Euclid* filters. Black solid lines display quasars in the redshift domains captured by the filters, with redshifts indicated directly on the lines. Blue lines display galaxies at $(1 \leq z \leq 2)$, susceptible to contaminating the quasar samples. MLT: stars M, triangles L, diamonds. For the M, L and T, filled: $z - Y_E$ and hollow: $I_E - Y_E$

multivariate densities is given by

$$\mathbb{P}[\widehat{\mathbf{F}}|\boldsymbol{\theta}_t, t] \sim \frac{1}{n_t}\sum_{i=1}^{n_t} p\left(\widehat{\mathbf{F}}\middle|\mathbf{F}_{t,i}(\boldsymbol{\theta}_t), \widehat{\boldsymbol{\sigma}}\right), \tag{3}$$

centred at $\mathbf{F}_{t,i}(\boldsymbol{\theta}_t)$ and with dispersion given by $\widehat{\boldsymbol{\sigma}}$.

The above mixture sets a uniform prior distribution on the SED templates modelling each population. If the photometric bands do not overlap, we can neglect the correlation between the fluxes in different bands. Thus, we rely on a product of Gaussian distributions truncated to be non-negative, given by

$$p(\widehat{\mathbf{F}}|\mathbf{F}_{t,i}(\boldsymbol{\theta}_t), \widehat{\boldsymbol{\sigma}}) = \prod_{j=1}^{N} \mathcal{L}_j, \tag{4}$$

where each truncated Gaussian distribution is given by

$$\mathcal{L}_j = \frac{\exp\left(-\frac{1}{2}\left(\frac{\widehat{F}_j - F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)^2\right)\mathbb{I}_{\widehat{F}_j \geq 0}}{\widehat{\sigma}_j\sqrt{2\pi}\left(1 - \mathcal{N}\left(-\frac{F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)\right)}, \tag{5}$$

where $\mathbb{I}_{\widehat{F}_j \geq 0}$ is the indicator function that the flux is non-negative and $\mathcal{N}$ is the cumulative distribution function of the standard normal distribution given by

$$\mathcal{N}(z) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{z}{\sqrt{2}}\right).$$

Note that the χ^2-criterion that is used to produce the best-fit parameters in the frequentist approach appears in the exponent of the likelihood given in Eq. (5).

In a general context, faint sources are measured with a low signal-to-noise ratio (S/N) in one or more bands. Forced photometry is a practice that is applied in many surveys, including *Euclid*, using detected images from which aperture fluxes are measured for all undetected sources. The likelihood is then calculated using Eq. (5) even when the forced photometry flux has a negative or null value. However, when this is impossible i.e., there is no estimation of the flux in a certain band, we re-write the likelihood with an unknown measured flux. This allows us to calculate the probability that a source is observed with a measured flux below the stated detection limit. This probability can be expressed as follows:

$$\mathcal{L}_j = \frac{\mathcal{N}\left(\frac{\widehat{F}_{lim} - F_{t,i,j}(\boldsymbol{\theta}_t)}{\sqrt{\widehat{\sigma}_j}}\right)}{1 - \mathcal{N}\left(-\frac{F_{t,i,j}(\boldsymbol{\theta}_t)}{\widehat{\sigma}_j}\right)}. \tag{6}$$

In OWL-z, Eq. (2) is calculated in two separate steps: initially to build a model and subsequently during comparison of the model with the sources' photometry.

For each population, OWL-z calculates the colours in all bands to build a model by generating maps of the values of colour for each population for each parameter value. The SED and the desired ranges of the parameters are used. Subsequently, the colours are calculated through the integration of magnitudes, as outlined by Hogg et al. 2002, across all SEDs employed. The model comparison step is located within the central portion of the code, where OWL-z examines each candidate for which comprehensive photometry is available (including associated errors), and contrasted it with the models. For each parameter set, it compares the photometry to all model fluxes. The model fluxes are calculated using the colours and model reference magnitude that we choose to be the H band in this work. However, any reference magnitude could be selected, provided that a detection is made in this band. It should be noted that OWL-z does not calculate the probability from the best value of the weighted evidence but rather from the integration of all values. Subsequently, the probability for each population is calculated using

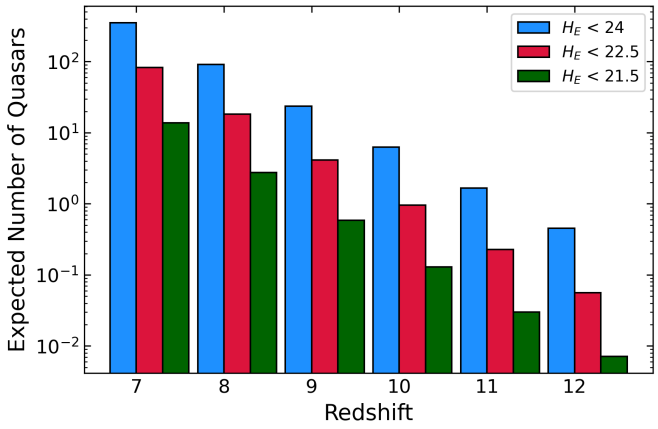


Fig. 2: The expected number of quasars per redshift bin in the 15,000 deg^2 EWS given the high-redshift quasar Luminosity Function adopted in this paper. In blue are shown the numbers of quasars expected up to $H_E < 24$, in crimson with $H_E < 22.5$ and in green with $H_E < 21.5$

Bayes probability in Eq. (1). Although OWL-z is not a best-fit extraction code, it incorporates snippets of code that facilitate the recovery of parameters enabling the calculation of the maximum a posteriori, thus providing the best-fit parameters for a given candidate. Moreover, several interesting output parameters are yielded by the code, in addition to the classification result, such as redshift, magnitude or templates. Note that the "best-fit" output of a Bayesian code is the "maximum posterior" (MAP), calculated by maximising the posterior, that is max(w_q), where $w_q = (w_{q_1}, w_{q_2}, ..., w_{q_{n_i}})$ are given by

$$w_{q_i} = \rho_{q_i}(H_{\rm mod}, z) \times P(\mathrm{D}, H_\mathrm{E}|H_\mathrm{E}, q_i). \quad (7)$$

In the following sections, we use the terms best-fit or MAP, but we always refer to the same Bayesian definition above.

2.2. The high-redshift quasars population

The statistical model of the quasar (QSO) population is based on a collection $n_{\rm QSO}$ of SEDs (also referred to as spectral templates or templates) and a set of parameters $\boldsymbol{\theta}_{\rm QSO}$ describing the quasars' properties. The spectral templates used in this paper are smoothed versions of the quasar composites presented in Bañados et al. 2016b and parametric SED models that accurately reproduce the observed optical and NIR colours of luminous Type 1 quasars over a wide range of redshifts and luminosities from Temple et al. 2021. The parameters are the redshift z and the apparent magnitude in a reference band, which we choose to be the *Euclid* H band, denoted H_E. This allows us to compute the expected fluxes in each band, $\mathbf{F}_{{\rm QSO},i}(\boldsymbol{\theta}_{\rm QSO})$ at the $\boldsymbol{\theta}_{\rm QSO} = (H_\mathrm{E}, z)$ value of the parameter set given the i-th SED model. The QSO likelihood of the data given $\boldsymbol{\theta}_{\rm QSO}$ is defined as the mixture of multivariate densities in Eq. (3). For the purposes of our paper, we let the redshift vary in the range $7 \leq z \leq 12$. We aim for a maximum redshift of 12 as the *Euclid* bands will cover up to the H_E band, and the Lyman-alpha emission line stays in the J_E band until $z \approx 12$.

To characterize the prior distribution of high-redshift quasars, we adopt the double power law parametric Luminosity Function (LF) from Willott et al. (2010) and adjust the parameters fit to Matsuoka et al. 2023:

$$\Phi(M_{1450}, z) = \frac{10^{k(z-7)}\Phi^*}{10^{0.4(\alpha+1)(M_{1450}-M^*_{1450})} + 10^{0.4(\beta+1)(M_{1450}-M^*_{1450})}}, \quad (8)$$

where M_{1450} is the quasar rest-frame absolute magnitude at $\lambda = 1450$Å and $\Phi(M_{1450}, z)$ is the number of quasars per magnitude bin, and unit volume at redshift z. The parameters are the normalisation Φ^*, the break magnitude M^*_{1450}, the bright end slope β and the faint end slope α, their values can be found in Table 1. Using the apparent magnitude in the H_E band, we can write the surface density of quasars in the sky in mag^{-1}×deg^{-2}× dz^{-1} units as:

$$\rho_{\rm QSO}(H_E, z) = \frac{1}{4\pi} \times \frac{dV_c}{dz} \times \Phi\left[H_E - \mu - K_{corr}(z), z\right], \quad (9)$$

where

- μ represents the distance modulus, defined as a function of the luminosity distance D_L by

$$\mu = \frac{5\log_{10}(D_L/10\ \mathrm{pc})}{\mathrm{pc}}, \quad (10)$$

- $K_{\rm corr}(z)$ denotes the K-correction, which converts the absolute magnitude at rest-frame $\lambda = 1450$Å into the observed H-band magnitude, and
- $\frac{1}{4\pi} \times \frac{dV_c}{dz}$ is the co-moving volume element per steradian and per redshift interval dz.

Fig. 2 illustrates the number of quasars per redshift bin and up to three different values of H_E across the entire 15,000 deg^2 of the EWS, as derived from Eq. (9). This demonstrates that *Euclid* will detect tens of quasars brighter than $H_\mathrm{E} \approx 22.5$ between redshifts 7 and 8, and potentially a few up to z ≈ 9. Up to this H_E magnitude, the selection with OWL-z will be robust while becoming increasingly more hazardous beyond it.

Table 1: Quasar Luminosity function parameters

α	β	k	M^*_{1450}	$\phi^* 10^{-8}$ (Mpc^{-3}mag^{-1})
−1.2	−3.34	−0.47	−24.38	0.475

2.3. The MLT stars population

For the purpose of our Bayesian analysis, we need to model the spatial distribution of MLT dwarfs, their luminosity, and spectral energy distributions from the optical to the near-infrared. For spectral energy distributions, we use the stellar dwarf Luyten Half-Second (LHS) library (Bakos et al. 2002) for M3-M5 brown dwarfs and the brown dwarf spectral library SpeX prism (Burgasser 2014; Burgasser & Splat Development Team 2017) for other spectral types. The use of spectral data enables magnitudes and colours to be correctly determined and modelled in the *Euclid* filters or any other filters as long as they overlap the spectral range of the spectra. Moreover, there are between 5 and ≈ 50 spectra available for each spectral type. Our models also include WISE photometric data for the W1 and W2 bands, anchored to the average JHK photometry of brown dwarfs as tabulated in Best et al. 2017. This allows us to include in our model the WISE photometry that cannot be calculated from the MLT

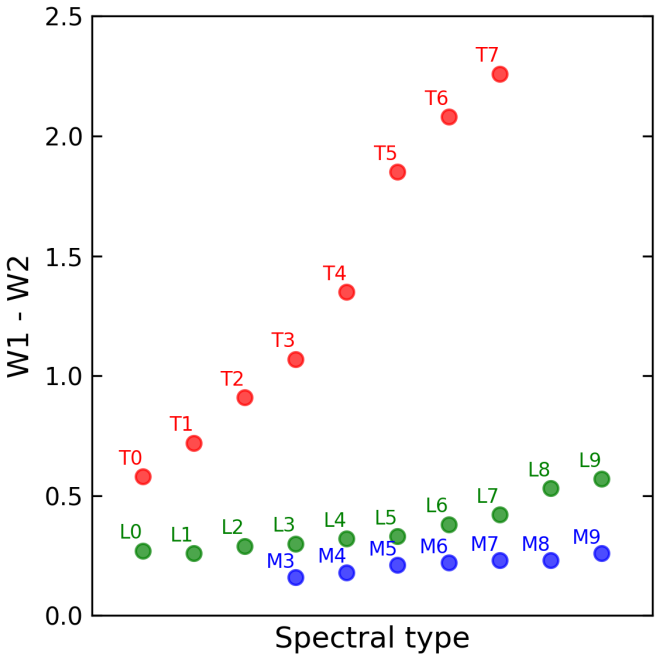


Fig. 3: WISE $W1 - W2$ colours of MLT dwarfs from Best et al. 2017 shown in blue, green, and red, respectively, with the values indicating their spectral type.

spectra that do not extend beyond the K band. Fig. 3 shows the $W1 - W2$ colours for all MLT types as indicated in Best et al. 2017, showing that T dwarfs, especially late types, have very strong $W1 - W2$ colours that can easily be discriminated from the colours of high-redshift quasars.

To estimate the probability of brown dwarf contamination when searching for high redshift objects in wide-field imaging datasets, the spatial distribution of stellar dwarfs in the Galaxy is usually modelled and restricted to the thin disc of the Milky Way (Caballero et al. 2008; Euclid Collaboration: Barnett et al. 2019; Pipien et al. 2018b). Detailed modelling of the thin disc is beyond the scope of this paper, and we therefore follow Caballero et al. 2008 by assuming a similar thin-disc scale height for LT dwarfs to its value determined from earlier stellar dwarf types (GKM) and a simplified exponential vertical and horizontal distribution. For detailed theoretical modelling of thin disc parameters as a function of stellar age and metallicity, calibrated against Gaia and APOGEE data, see, for example, Sysoliatina, K. & Just, A. 2022.

At the magnitudes corresponding to the depth of the EWS ($\sim$ 24), the distance to T dwarfs varies from $\sim$ 150 pc for the coldest to $\sim$ 450 pc for the hottest. With a scale height of the thin disc of the order of 300 pc (Gaia Collaboration et al. 2023; Vieira et al. 2022), we infer that the sole thin disc approximation will reasonably well represent the population of T dwarfs detectable in the EWS. However, the same will not be true for M dwarfs. An M6 dwarf at an apparent magnitude $\approx$ 24 is at a distance of $\sim$ 4000 pc, extending well beyond the thin disc into the thick disc or the halo. As a matter of fact, the *James Webb* Space Telescope has revealed in extragalactic fields a significant number of extremely faint and low-temperature brown dwarfs extending well into the thick disc and possibly into the halo, at distances of up to 2 kpc (Hainline et al. 2024b,a; Burgasser et al. 2024). An analysis of the thick disc contamination is therefore warranted.

For this analysis, we follow a similar approach to that used in Ryan & Reid 2016 and compare the distribution of stellar dwarfs in thin and thick discs as a function of magnitude. We adopt a value of 330 pc for the scale height of the thin disc (Caballero et al. 2008), 800 pc for the scale height of the thick disc (Vieira et al. 2023), 2250 pc (Caballero et al. 2008) for the value of the scale length of both discs and a normalisation factor between the thick and thin discs local densities of 10%. Using Eq. (12) of Pipien et al. 2018b, we compare brown dwarf densities for different spectral types in the thin and thick discs for two lines of sight: $b = 90°$ and $(l, b) = (90°, 30°)$ (the lowest Galactic latitude of the EWS is $b = 23°$). For local volume densities in the Galactic plane and absolute *Euclid* magnitudes by spectral type, we refer to Table 2 of Euclid Collaboration: Barnett et al. 2019 and the references therein (Dupuy & Liu 2012; Skrzypek et al. 2016; Bochanski et al. 2010). Late-type M dwarfs can potentially contaminate searches for galaxies or quasars at high redshift in some surveys, whereas, as we will see later, only L and T-type brown dwarfs contaminate the search for $z \geq 7$ with *Euclid*. Consequently, for the purposes of this paper, we restrict our analysis to the brown dwarf population within the M dwarf population, that is, dwarfs of a type later than approximately M6.

The analysis (see Fig. 4) indicates that up to *Euclid* magnitude $H_{\rm E} \sim 24$, the thin disc dominates the number of L and T dwarfs at low and high Galactic latitudes. For M6 to M9 dwarfs, the thick disc is the main contributor above H $\sim$ 22.5 magnitudes at high Galactic latitudes, whereas at low Galactic latitudes, the thin disc dominates up to $H_{\rm E} \sim 24.5$ and beyond. For the purposes of our paper, Galactic latitude is, therefore, the dominant parameter affecting the population of contaminating stars. We explore the effect of Galactic latitude on the performance of our code in Sect. 4.

This simple analysis assumes that stellar dwarf populations are the same in both thin and thick discs. This oversimplification ignores the differences in age and metallicity between the constituents of the thin and thick discs and, therefore, the significant differences that may exist between the stellar dwarf populations within them. Early M-type dwarfs (M0 to M5) have masses above the hydrogen-burning minimum mass, and their luminosity evolves little on time scales up to 10 billion years, whereas brown dwarfs below this limit cool on much faster time scales (Reid 2013). As a result, late M-type dwarfs evolve into later spectral types, and their luminosities drop rapidly over timescales of a Gyr or less. The thick disc, which is older than the thin disc, should, therefore, be depleted of late-type brown dwarfs (M6 to M9 and L), and T dwarfs should be significantly less luminous in the thick disc than in the thin disc (see also the discussion in Caballero et al. 2008). This suggests that the stellar densities reported in Fig. 4 are likely to be seriously overestimated in the thick disc at a given magnitude for spectral types later than M5.

As a conclusion to this analysis, we restrict the modelling of the Galaxy's brown dwarf population to the thin disc and analyse the impact of stellar density through its dependence on Galactic latitude. In the coming years, *Euclid* and *James Webb* Space Telescope data will enable a more detailed understanding of the relative populations of brown dwarfs in the thin and thick discs.

We now return to the formulation of our Bayesian model for the description of the stellar dwarf population. The parameters of this population are magnitude (or heliocentric distance), and spectral type $\{H_{\rm E}, spt\}$. We use the *Euclid* $H_{\rm E}$ band as the reference band for calculating magnitudes and the spectral types from M6 to T9. Using the notation described in Sect. 2, we write the weighted evidence of this population as follows:

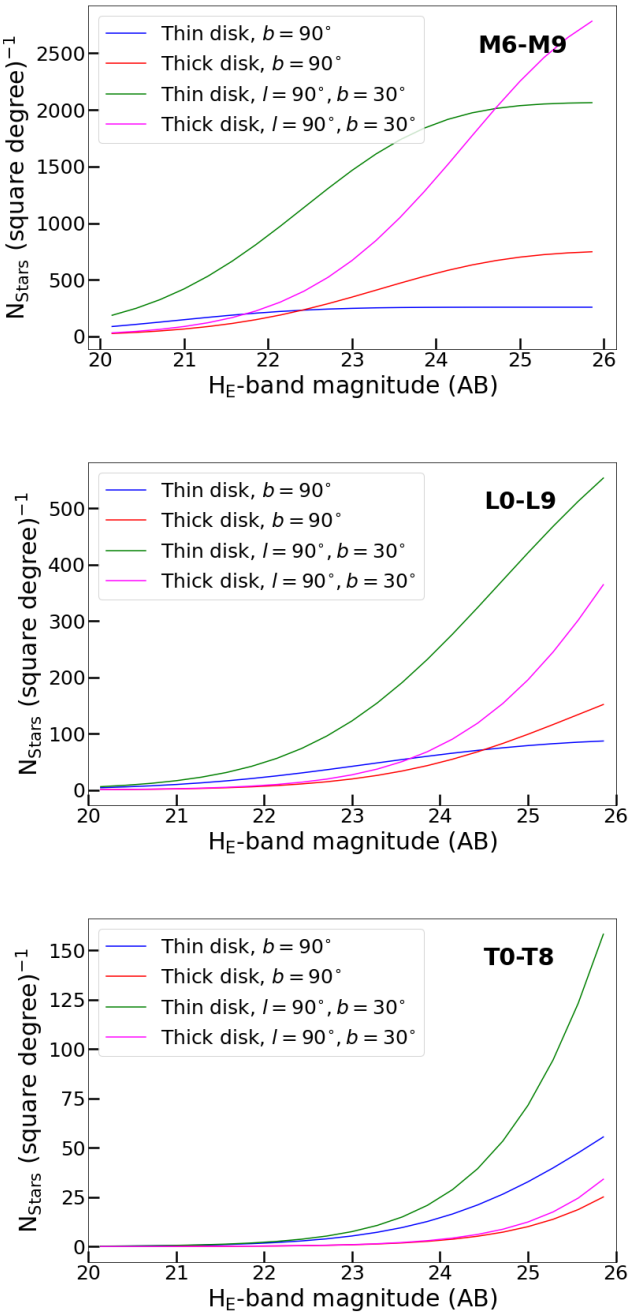


Fig. 4: Brown dwarf number counts as a function of H-band magnitude for the thin and thick discs in two different field locations (b = 90° and (l,b)=(90°,30°)). Top: M6 to M9 spectral types, middle: L0 to L9 and bottom: T0 to T8.

$$W_s = \sum_{M6}^{T9} \int_{-\infty}^{+\infty} \rho_s^{(l,b)}(H_{mod}, spt)\mathbb{P}(D|H_{mod}, spt)\mathrm{d}H_{mod}, \quad (11)$$

where $\rho_s(H_{mod}, spt)$ is the surface density of the contaminating brown dwarf population of *Euclid* magnitude H_{mod} and spectral type spt, at a given Galactic longitude l and latitude b. We model the spatial distribution across the Galaxy in mag^{-1} × deg^{-2} × spt^{-1} units as a function of the heliocentric distance to the star d,

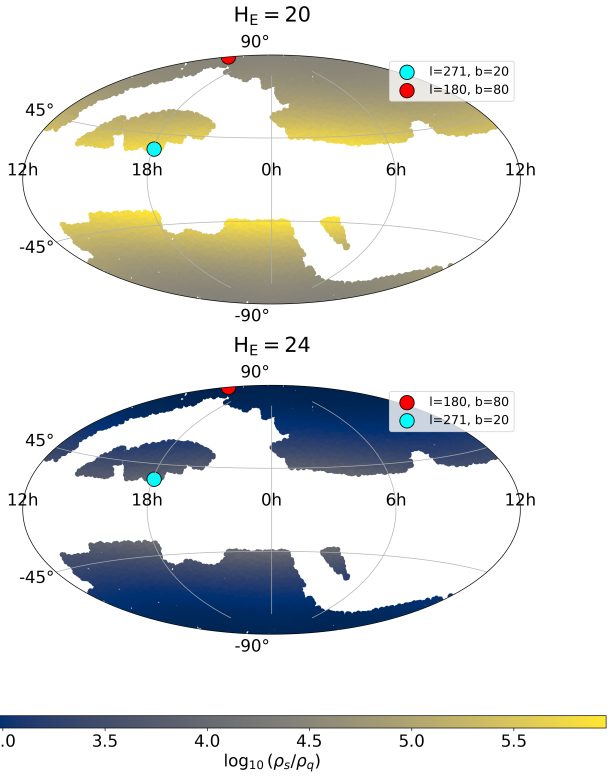


Fig. 5: Ratio of the surface density of brown dwarfs of spectral type M6 to T8 to the surface density of $z > 7$ quasars over the *Euclid* footprint shown in Galactic coordinates and galactic projection. Top: for magnitude H_E= 20, bottom for magnitude H_E= 24.

assumed to be far smaller than the solar galactocentric distance $R_\odot$ (Caballero et al. 2008):

$$\rho_s^{(l,b)}(H_{mod}, spt) \approx \rho_{0,spt} \exp\left(\frac{\mp Z_\odot}{h_Z}\right) \mathcal{R}(d_{H_{mod}}, l, b), \quad (12)$$

where,

$$\mathcal{R}(d_{H_{mod}}, l, b) = \exp\left[-d_{H_{mod}}\left(-\frac{\cos(b)\cos(l)}{h_R} \pm \frac{\sin(b)}{h_Z}\right)\right] \quad (13)$$

where $\rho_{0,spt} \exp\left(\frac{\mp Z_\odot}{h_Z}\right)$ is the local volume density of brown dwarfs of spectral type spt, $Z_\odot$ the height of the Sun above the Galactic plane (assumed to be 27 pc), h_Z and h_R the scale height and length of the thin disc as mentioned above, and the sign convention indicates whether the source is above or below the Galactic plane (see Caballero et al. 2008). Finally, H_{mod} and $d_{H_{mod}}$ are related through:

$$H_{mod} - M_{H_{mod}} = 5\log(d_{H_{mod}}) - 4 + 3.09\ E(B-V), \quad (14)$$

where H_{mod} and $M_{H_{mod}}$ are the apparent and absolute magnitudes in the reference band H_E and $E(B-V)$ the interstellar reddening.

Using the modelling of the high-z quasar population (Sect. 2.2) and the MLT population described in this section and the MLT population described in this section (M6 to T9), we derive the ratio of the brown dwarf to high-z QSO density on the *Euclid* footprint. This is shown in Fig. 5 for two different apparent magnitudes.

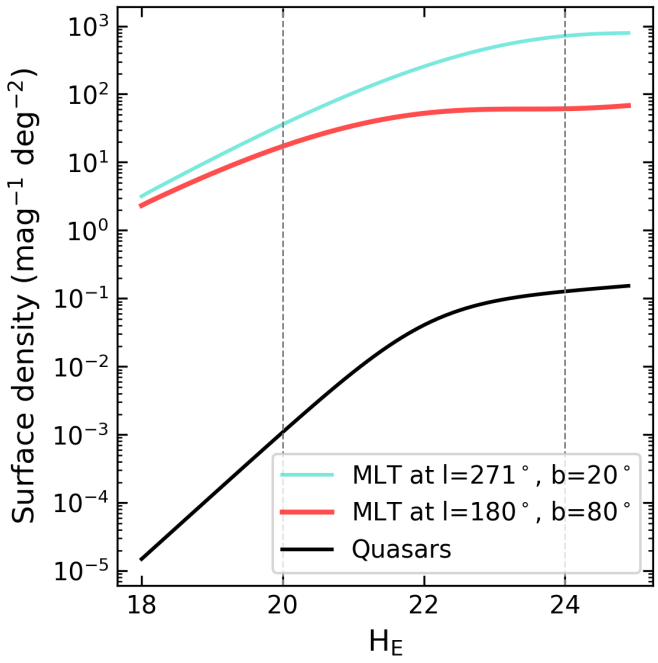


Fig. 6: Surface number densities of MLT stars compared to $z > 7$ quasars, as a function of apparent magnitude, in two different locations of the *Euclid* footprint.

Fig. 6 shows the densities of quasars and M6 to T9 brown dwarfs for two positions with different Galactic coordinates in the *Euclid* footprint.

2.4. The contaminant galaxy model

There are two different types of galaxies susceptible to contaminating the quasar samples because they display red colours that mimic those of high-z quasars, in particular at low SNR: early-type galaxies at intermediate redshift ($1 < z < 2$) on the one hand, and dusty star-burst galaxies on the other hand. The former exhibit an important 4000Å break because they are dominated by an old stellar population, whereas the presence of dust reddens the latter. To assess the extent of this contamination by intermediate- z galaxies, we have carried out two different tests. The first one consists of precisely determining the contaminant population by using straightforward simulations. The second one is to quantify the contamination in the real Universe based on the COSMOS2022 field (Weaver et al. 2022) by computing the Luminosity Function (LF) of the contaminant population. The LF will be used later to model the abundance of contaminant galaxies.

To precisely identify the contaminant population, we run OWL-z on a simulated catalogue containing 100,000 galaxies uniformly sorted in redshift (between $0 \leq z \leq 6$), spectrophotometic templates, and extinction values (between $0 \leq A_V \leq 3$ magnitudes), using the Calzetti extinction law (Calzetti et al. 2000). The simulated catalogue was generated using `make_catalogue`, a tool from the HyperZ tool (Bolzonella et al. 2000). The template library includes two evolutionary synthesis models: a delta burst -Single Stellar Population (SSP)- from the Bruzual & Charlot code (Bruzual & Charlot 2003), with Chabrier IMF (Chabrier 2003) and solar metallicity, and 10 Starbursts99 templates (Leitherer et al. 1999), including emission lines, for single bursts and constant star formation rate models, each one spanning five metallicities (Z =0.04, 0.02, 0.008, 0.004 and 0.001) and 37 ages for the stellar population (between 0 and 1 Gyr). Apparent magnitudes and associated errors have been computed using these templates, sorted to uniformly sample the range of magnitudes where galaxies are expected to be the dominant population in the EWS, in the reference H_E-band, that is $22 \leq H_E \leq 25$. Photometric error bars and noise are scaled to apparent magnitudes, assuming a Gaussian distribution, according to the expected SNR in the different bands used in the EWS. The probability threshold for the selection as a quasar is set to $P_q = 0.1$ for this analysis, following Mortlock et al. 2012 and Euclid Collaboration: Barnett et al. 2019.

The first result of this experiment is that young starbursts, represented by the Starburst99 models, only contaminate the sample at low SNR values for extremely high values of $A_V \geq 1.5$, spanning a relatively broad domain in redshift ($1 \leq z \leq 5$). It is important to acknowledge the significant influence of SNR values in the reference magnitude on our selection process. Indeed, a low SNR in the reference filter H induces a noisy determination of the $J - H$ colour, which in turn leads to a higher degree of confusion in the probability calculation and a greater likelihood of misidentifying high-redshift quasars. This is a general remark affecting all contaminant galaxies but is particularly important for dusty starbursts. The main contaminants among this population are galaxies at $z \gtrsim 4$ with $A_V \geq 1.5$, and galaxies at $1 \leq z \leq 2$ with $A_V \geq 3$. In the two cases, such extremely reddened galaxies are relatively rare in the real Universe at these redshifts.

Our analysis shows that the main contamination comes from early-type galaxies, represented by SSP models, with ages above 1 Gyr, at intermediate redshifts, $1 < z < 2$. Indeed, the spectral energy distribution of early-type galaxies, such as elliptical and lenticular galaxies, is characterised by an old stellar population, well represented by an initial burst of star formation followed by a rapid decline (see e.g. Ali et al. 2024). Contrary to dusty starbursts, these galaxies are relatively abundant at $1 < z < 2$. For this reason, in the following, we consider that the main contamination comes from early-type galaxies. In this regard, our results are consistent with the assumptions of previous works (e.g. Euclid Collaboration: Barnett et al. 2019). The abundance of these galaxies has been well studied at $z \sim 1.5$, for instance, by Zucca et al. 2006, who studied the evolution of the LF for different filters and spectrophotometric types of galaxies. We can use these previous findings to guide our modelling in the sensitive magnitude domain, as shown below.

After identifying the nature of the contaminant galaxies, we determine the LF of this population. For this need, we use the Euclidized COSMOS2022 catalogue (Weaver et al. 2022), named E-COSMOS hereafter. This catalogue is described in more detail in Sect. 4.3.3. Note that for galaxies detected in the J_E-band at intermediate redshift, the relevant filter for determining the LF is the rest-frame B band.

To represent the LF, we use the classical parametrisation provided by the Schechter function (Schechter 1978) in terms of magnitude given by:

$$n(M) = (0.4 \ln 10)\, \phi^* \, [10^{0.4(M^*-M)}]^{\alpha+1} \exp[-10^{0.4(M^*-M)}]. \quad (15)$$

Here, $n(M)\, dM$ represents the number of galaxies per comoving Mpc^3 with magnitudes between M and $M + dM$. The parameters are as follows:

– ϕ^*: The normalisation factor, representing the overall volume density of galaxies in Mpc$^{-3}\times$ mag^{-1}.

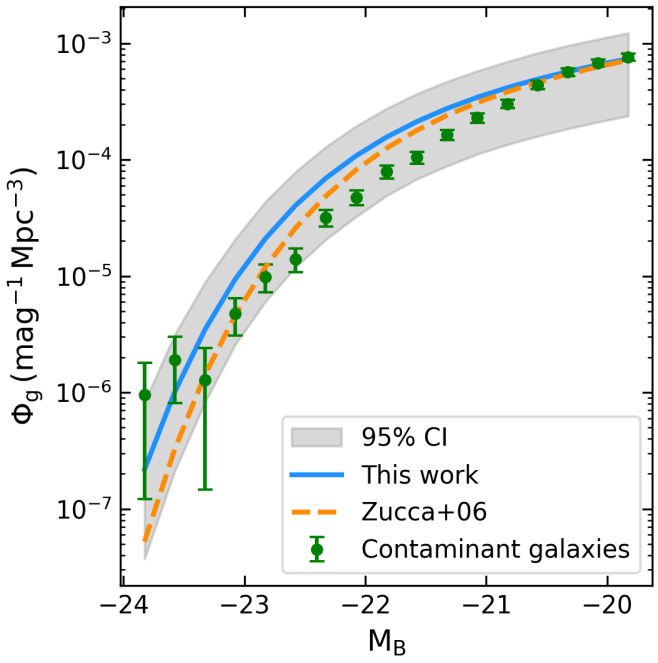


Fig. 7: Luminosity function fit in the B-band (blue) in addition to its 95% confidence interval (grey) for the population of early-type galaxies at $1 < z < 2$ in the E-COSMOS catalogue data (green), compared to the early-type galaxies (dashed orange) luminosity function fit by Zucca et al. 2006.

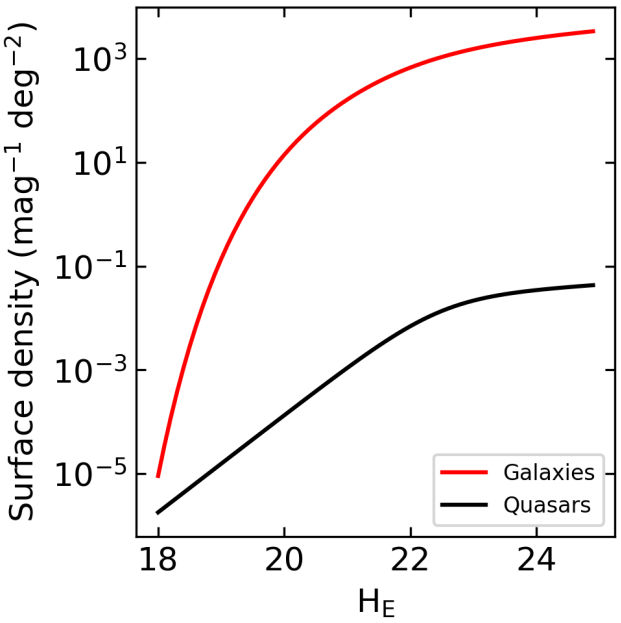


Fig. 8: The surface number density of early-type galaxies integrated in the redshift interval $1 < z < 2$ compared to the quasars surface number density integrated over the redshift interval $7 < z < 12$ quasars, as a function of apparent magnitude.

- M^*: The characteristic magnitude represents the cut-off between the (bright) luminosity regime dominated by the exponential function and the (faint) regime dominated by the power-law.
- α: The faint-end slope of the luminosity function, describing the distribution of the faintest galaxies.

For the sake of consistency, before selecting the target population of contaminant galaxies in E-COSMOS, we run OWL-z on the E-COSMOS catalogue, with only one contamination source: MLT stars, and the same threshold as before for the selection of high-z quasars. We check the spectrophotometric type and redshift of the 16 galaxies that have been selected as quasars. They are all early-type galaxies at intermediate redshift ($z \sim 1.5$), as expected from previous simulations.

In order to model the LF for this population of contaminants, we extract from the E-COSMOS catalogue the subset of early-type galaxies within the sensitive redshift domain $1 < z < 2$. To compute absolute magnitudes, we model the spectral energy distribution with a template well suited to represent this population, namely a short, exponentially decaying model with characteristic star-formation time $\tau = 0.1$ Gyr, age $= 10$ Gyr and solar metallicity ($Z = 0.02$). It is worth mentioning that the precise choice of this model does not change the results. The LF points and their associated error bars are then computed; error bars in the LF include Poisson noise and field-to-field variance. Field-to-field variance is derived using the Trenti & Stiavelli method and calculator (Trenti & Stiavelli 2008). We then fit the data points with the Schechter function in Eq. (15) using a χ^2 minimization.

Towards the faint end, we applied a magnitude cut to $M_B < -19.7$ because beyond this limit, our sample suffers from incompleteness. The parameters were chosen to enable easy comparison with the work of Zucca et al. 2006. This motivated us to fix the same value for the parameter α. For comparison, a value of $\alpha = 1.0$ gives a similar result to the function adopted by Euclid Collaboration: Barnett et al. 2019 to represent the same population. Note that the precise choice of α is irrelevant here because, on the one hand, only the faintest part of the LF is dominated by the power law and, on the other hand, the population of contaminant galaxies lies preferentially in the bright end.

Table 2: Comparison of the Schechter function parameters obtained in this work with the results by Zucca et al. 2006 for early-type galaxies at redshift $1 < z < 1.2$. The value of the slope has been fixed to $\alpha = -1.23$ in our fit.

	α	M^*	ϕ^* $\times 10^{-4}$ Mpc^{-3}
Zucca et al. (2006)	−1.23	$-21.5^{+0.48}_{-0.57}$	$3.0^{+1.4}_{-1.2}$
This work	−1.23	$-21.7^{+0.17}_{-0.17}$	$2.9^{+0.6}_{-0.6}$

Fig. 7 shows that our LF points are in good agreement with Zucca et al. 2006. Our fit parameters are shown in Table 2. The LF obtained above is then used to model the prior distribution of early-type galaxies $\rho_{g_i}(H_{\rm mod}, z)$ given in mag$^{-1}\times$ deg$^{-2}\times$ dz^{-1} units by:

$$\rho_g(H, z_g) = \frac{1}{4\pi} \times \frac{dV_c}{dz_g} \times \Phi_g \left[H_{\rm mod} - \mu - K_{\rm corr}(z_g), z_g\right], \qquad (16)$$

where z_g is the redshift range of the contaminant galaxies $1 < z < 2$ and Φ_g is the luminosity function fit obtained in this work.

For a given set of parameters $\{z_g, H_{\rm mod}\}$ and an early-type galaxy model g_i, the weighted evidence can be quantified as:

$$W_{g_i} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \rho_{g_i}(H_{\rm mod}, z_g) P(\text{bands}, H_{\rm obs} | H_{\rm obs}, g_i) dH_{\rm mod} dz_g. \qquad (17)$$

Fig. 8 displays a comparison between the surface number density of early-type galaxies in the $1 < z < 2$ interval and $z > 7$ quasars as a function of apparent magnitude. The population of contaminant galaxies clearly dominates the number counts.

3. Technical description of OWL-z

The OWL-z code calculates the probability of each preselected candidate belonging to one of the modelled populations, namely high-redshift quasars, early-type galaxies at intermediate redshifts, and MLT dwarfs. The code is self-contained and written in the Python programming language and its various packages, such as NumPy, SciPy, and Astropy. It is capable of functioning without any dependencies on external software.

3.1. Inputs

The OWL-z code relies on the computation of colours from the SEDs for all three populations in addition to the K-corrections for quasars and galaxies. To do so, the model requires the filter transmission curve data for each band, as well as a designated reference magnitude. Additional parameters are the magnitude range and the redshift range to explore, for each population. In the case of brown dwarfs, for each spectral type, the range of absolute magnitudes and the local densities should be provided. Templates for each population should be provided, and the code distribution already includes a large number of them for the three populations considered. It is possible to add as many templates as required, provided that they are in the correct format.

3.2. Outputs

The program provides a list of probabilities of belonging to each class for each source. The user then interprets the data and makes a decision. Furthermore, it provides a list of useful output parameters, including the optimal redshift for galaxies and quasars and the optimal spectral type for the best-fit stellar model. It also provides the SEDs of the galaxies and quasars that provide the maximum posterior condition.

3.3. Efficiency

OWL-z has several efficiency indicators:

- OWL-z requires the user to fill a configuration file which is straightforward to modify, requiring only the input of the user to enable its operation.
- The code is self-sustained, whereby the models and K-corrections are calculated concurrently with the code and the input parameters.
- The code is flexible regarding the numbers and nature of bands used, provided that they do not overlap.
- The code is written with high-performance computing methods in mind and is partly written in Cython for the purpose of enhancing numerical speed.
- In addition to providing a probability value, the model also provides additional output parameters for each source, thus facilitating a more accurate interpretation of the results.

3.4. Limitations

A natural limitation of the code is that the filters correspond to the spectral range of the templates used, for brown dwarfs on the one hand and for galaxies and quasars in the redshift ranges to be explored for each of them.

In the released version of the code corresponding to the one described in this paper, the brown dwarf templates are observed spectra and cover up to 2500 nm, thus imposing that the filters that can be provided are limited to the K-band. The galaxy and quasar templates have a sufficient spectral range for the redshift ranges explored in this paper and do not impose additional limits.

As will be seen later in this paper, in Sect. 4.2, the colours in the WISE filters have been used to further model the SED beyond the K-band, but this is exclusively possible with the tabulated data provided in Best et al. 2017 and used in the code and cannot be generalised to other datasets extending beyond K-band. Extending the filters that can be used beyond the K-band would require brown dwarf templates with extended spectral ranges that are not readily available.

4. Validation and Performance

4.1. Methodology

This section is dedicated to the validation of OWL-z using two different approaches. The first one consists of re-identifying all the $z > 7$ spectroscopically confirmed quasars available in the literature. To achieve this, OWL-z was configured on a case-by-case basis with the photometric data from which the individual quasars were identified. The second approach relies on simulating different flavours of the EWS catalogues, allowing us to quantitatively evaluate the performance expected for OWL-z based on the measurement of the completeness, a measurement of how many (if any) quasars we misidentify with this process and purity, a measurement enables us to ascertain whether we are still contaminated among the objects identified as quasars by OWL-z. A global performance metric is also introduced, combining both completeness and purity. Performance estimators are important for optimising the selection of high-redshift quasar candidates with OWL-z, minimizing the demand for spectroscopic follow-up and maximizing the confirmation rate with a reasonable observational effort.

4.2. Re-identifying known $z > 7$ quasars

We analyse the performance of our code in recovering known and spectroscopically-confirmed quasars at $z > 7$ and selected from near-IR surveys. The versatility of OWL-z allows us to apply it to the different optical and NIR data sets used in the discovery of these quasars. The objective here is to determine if OWL-z is able to identify them as quasars and to compare their spectroscopic redshift with the redshift obtained by the code.

Table 3 lists the quasars used for this analysis, the photometric bands used for their discovery, their measured spectroscopic redshifts, and their reference magnitude.

The photometric data encompass data from DECaLS (Dey et al. 2019), UKIDSS (Warren et al. 2007), and WISE (Wright et al. 2010). Given the shallow depth and non-constraining photometry of WISE bands $W3$ and $W4$, we only use photometric data from bands W1 and W2 in the following.

The pre-selection methods used to select these quasars were mainly based on colour selection and the identification of strong Lyman breaks ($z - J > 4$) around $\sim 1\mu$m. An additional criterion was used for the five cases out of eight where WISE photometry was available, consisting of a $W1 - W2 < 0.7$ colour selection enabling the rejection contamination by T-dwarfs (see Fig. 3).

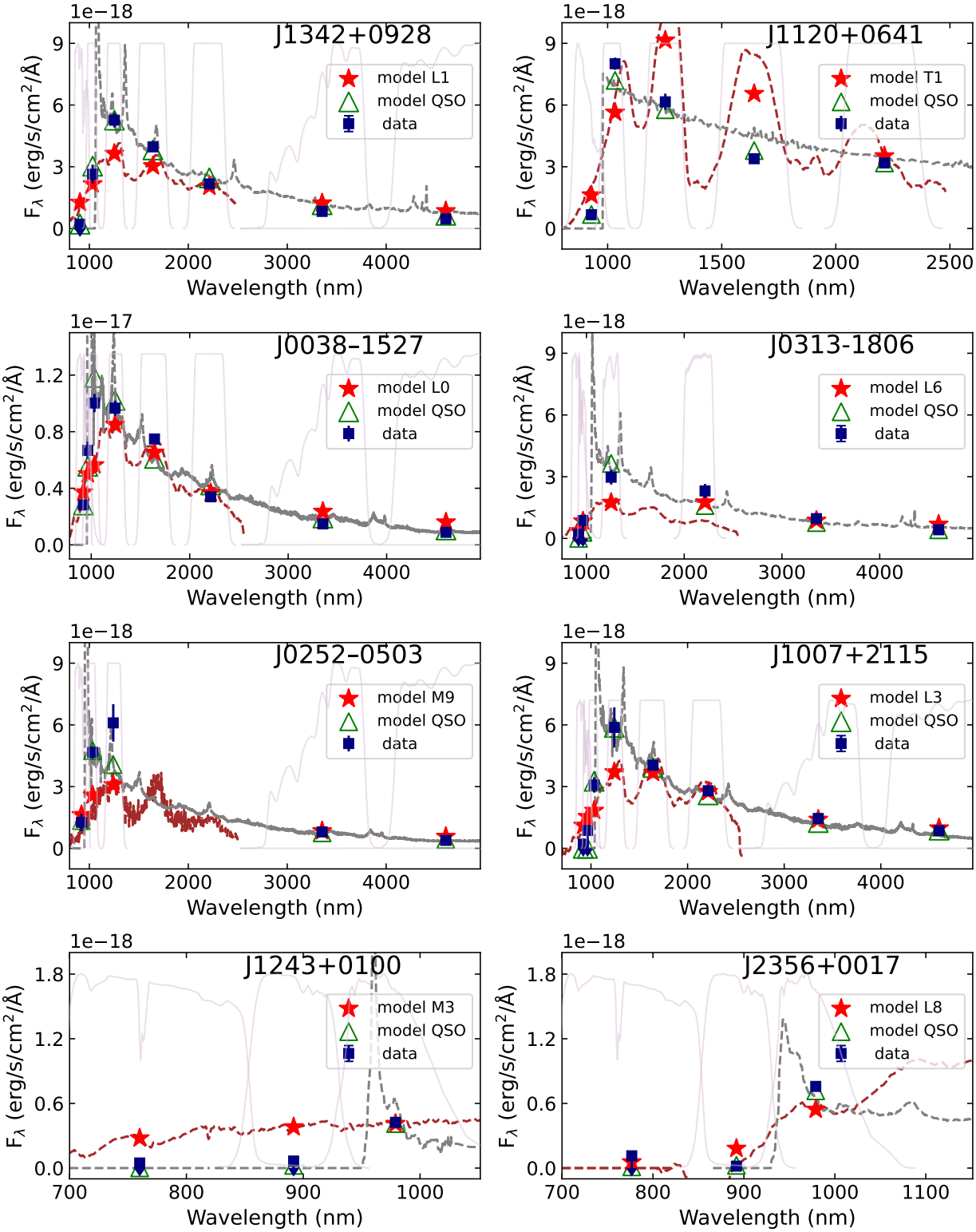


Fig. 9: Photometric data for the eight spectroscopically confirmed quasars at $z > 7$ and the OWL-z best-fit solutions for quasars and MLT dwarfs. The original photometric data points are represented by blue squares. The best-fit quasar model is in dashed grey, the best-fit model for an MLT dwarf is in dashed brown, and the filter transmission curves are shown in solid light purple. Model photometry for best-fit solutions is represented by green triangles for quasars and red stars for MLT dwarfs.

The results of OWL-z are also presented in Table 3, in particular, the output redshifts and P_q values returned by the code. The primary potential contaminants expected for the QSOs in this study are L brown dwarfs. OWL-z yields a probability $P_q > 0.99$ for all of the aforementioned QSOs, showing the high performance of the code in re-identifying the whole sample. Fig. 9 shows the photometric data, including error bars, for all quasars

Table 3: Known spectroscopically confirmed quasars at $z > 7$. The photometric data used for their discovery and spectroscopic redshift (z_{spec}) are indicated. The columns z_{out} and P_q refer to the redshift and probability of being a quasar calculated by OWL-z.

QSO name	Photometry	z_{spec}	z_{out}	P_q	mag	Reference
J1342+0928	z_{DECam}, Y,J,H,K, $W1$,$W2$	7.54	7.8	1	J 20.75±0.11	Bañados et al. 2018b
J1120+0641	z_{DECam}, Y,J,H,K,$W1$,$W2$	7.08	7.05	1	J 20.13±0.07	Mortlock et al. 2011
J0313-1806	z_{PS1},y_{PS1} ,J,K	7.64	7.7	1	J 20.92 ±0.13	Wang et al. 2021
J0038–1527	z_{DECam},y_{PS1}, Y,J,H,K, $W1$,$W2$	7.02	6.95	1	J 19.66±0.07	Wang et al. 2018
J1007+2115	z_{DECam},y_{PS1}, Y,J,H,K, $W1$,$W2$	7.51	7.65	1	J 20.20±0.18	Yang et al. 2020
J0252–0503	z_{DECam}, Y, J, $W1$, $W2$	7.02	6.9	1	J 20.16±0.07	Yang et al. 2019
J1243+0100	g_{HSC}, r_{HSC}, i_{HSC}, z_{HSC}, y_{HSC}	7.07	9.9	1	y_{HSC} 23.57±0.08	Matsuoka et al. 2019b
J2356+0017	i_{HSC}, z_{HSC}, y_{HSC}	7.01	6.75	1	y_{HSC} 22.94±0.05	Matsuoka et al. 2019a

and the best-fit quasar and MLT dwarf solutions returned by the code.

Some results deserve a specific comment. OWL-z has provided an excellent result for all quasars, including those detected in a single band with non-detection constraints in the others (J2356+0017 and J1243+0100) because the photometric bands in HSC are very deep, and the Lyα break much easier to detect. These quasars have broad Lyα emission and are low-luminosity quasars, matching well the properties of the quasar population model (Sect. 2.2), unlike the SEDs of the MLT dwarfs (Sect. 2.3). As expected, the best-fit MLT solutions for quasars with WISE photometry are late M or L dwarfs, given the aforementioned colour selection criterion ($W1 - W2 < 0.7$) that excludes T dwarfs from photometric selection (Fig. 3). Conversely, best-fit solutions by T-dwarfs are more likely when WISE data are not available, as is the case for J1120+0641. WISE data proved crucial for J0252-0503, without which there would have been only one detection.

4.3. Expected performance on EWS simulated data

The performance of OWL-z on the EWS is estimated in terms of completeness and purity in order to ascertain the sensitivity to confusion and contamination by MLT dwarfs and early-type galaxies. To this end, a series of catalogues is simulated in the following sections using *Euclid* filters (optical I_E and NIR $Y_E J_E H_E$) and sensitivities detailed in Table 4.

Table 4: Used bands characteristics. The *Euclid* bands characteristics are a duplicate of Table 7 in Euclid Collaboration: Scaramella et al. 2022. The SNR statistics and depths (magnitudes) in the region of interest. All the statistics have been computed for each channel: optical I_E and LSST z-band and NIR Y_E, J_E, H_E, and WISE channels W1 and W2. The median depth here is evaluated for 5σ point-like source

band	λ_{eff}	median SNR	median depth
I_E	6869.74	15.9	26.2
z	8678.90	5	25.9
Y_E	10688.60	6.5	24.3
J_E	13422.62	7.8	24.5
H_E	17410.63	7.2	24.4
W1	33526.00	5	19.8
W2	46028.00	5	18.9

4.3.1. Performance: Completeness

To estimate the completeness of the OWL-z selection method, we create a mock catalogue of 500,000 quasars. In order to gain a deeper understanding of the impact of using only Euclid data and of replacing or supplementing the existing data with other bands, three different scenarios have been developed. The first scenario involves using only Euclid data. In the second scenario, the optical band I_E is replaced by the optical z-band of LSST. In the third scenario, in addition to the Euclid optical and NIR data, the photometry is extended to the WISE coverage in the W1 and W2 bands (see Table5). The sensitivities of these additional bands can also be found in Table 4 The completeness is calculated on the EWS DR6 (Euclid Collaboration: Scaramella et al. 2022) footprint, which covers 15,000 deg^2. Quasars are randomly selected with a flat distribution in redshift in the range $7 \leq z \leq 12$, a flat distribution in absolute magnitude M_{1450} in the range $-29 \leq M_{1450} \leq -22$, a flat distribution over all the SEDs in our library, and finally, a flat spatial distribution within the 15,000 deg^2 of the EWS footprint. For each object in the catalogue, we calculate apparent magnitudes in the *Euclid* bands, in the LSST z band, and in the WISE W_1 and W_2 bands. For photometric errors, we adopt those corresponding to the DR1 release of the LSST data (Ivezić et al. 2019b) and those for the WISE data; we use the W1 and W2 bands and their corresponding errors from (`wise2.ipac.caltech.edu`). For *Euclid* photometric errors, we use the signal-to-noise ratio maps across the EWS footprint as described in Euclid Collaboration: Scaramella et al. 2022; all of this information can be found in Table 4.

We define the completeness C of the samples selected by OWL-z as follows:

$$C = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{18}$$

where TP (True Positive) is the number of quasars that have been successfully classified as high-redshift quasars, and FN (False Negatives) is the number of quasars that have been incorrectly classified as contaminants. For the sake of consistency with other work, we define a successful classification when $P_q > 0.1$, and we will examine in Sect. 5 how a different definition affects the results.

For all the objects in our mock quasar catalogue, we calculate the probability P_q, P_s and P_g that the quasar is identified as a quasar, star and galaxy, respectively, as defined in Eq. (1). Additionally, the SEDs for each category that allow for a maximum posterior are retrieved.

Results are presented as a function of redshift and H_E magnitude in Fig. 10, for a selection from *Euclid*'s $I_E Y_E J_E H_E$ data. The colours correspond to different increments of the completeness value. We also evaluate how completeness varies when the

Table 5: Scenarios explored in the simulations for the completeness estimation.

Survey	NIR bands	optical bands
Euclid	$Y_E J_E H_E$	I_E
Euclid +LSST	$Y_E J_E H_E$	z
Euclid +WISE	$Y_E J_E H_E$W1 W2	I_E

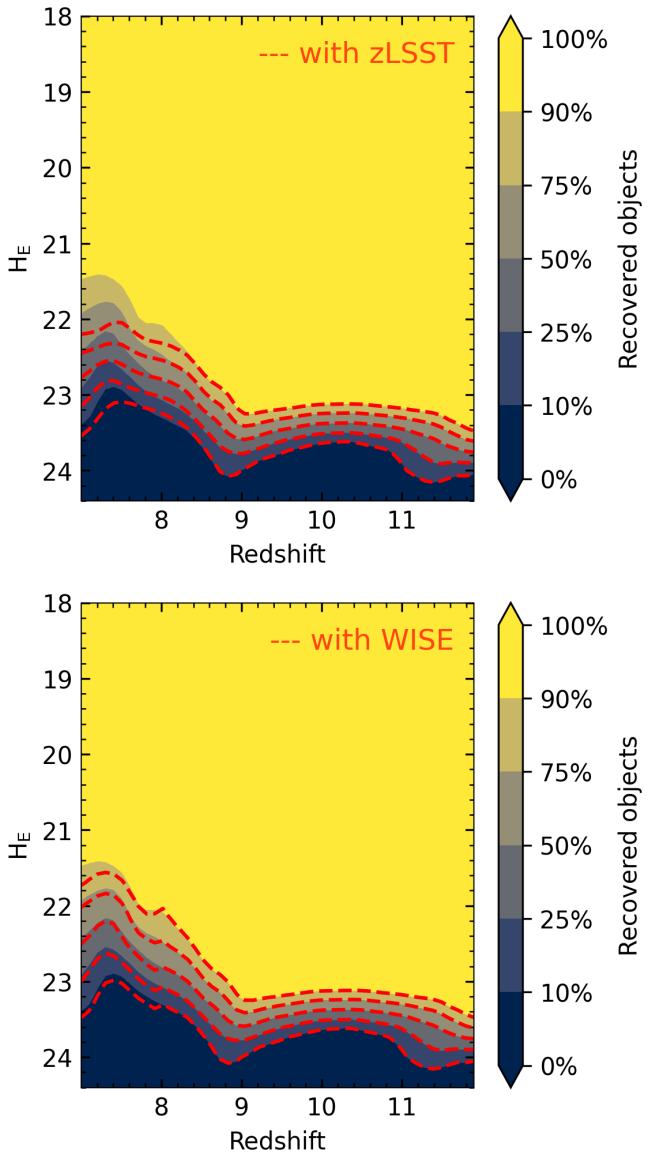


Fig. 10: The selection functions of high redshift quasars determined using OWL-z for *Euclid* $I_E Y_E J_E H_E$ data. The selection function designs the level of completeness per redshift bin; the completeness of a bin is calculated by counting the percentage of quasars in the bin with probability is$\mathbb{P} > 0.1$. Several cases are shown: *(top panel)* In red, the optical band O of *Euclid* is replaced by the z-band from LSST. *(bottom panel) Euclid* $I_E Y_E J_E H_E$ data and in red *Euclid* $I_E Y_E J_E H_E$ in addition to WISE (W1 and W2 bands).

LSST z band (top-hand figure) and the WISE $W1$ and $W2$ bands (bottom-hand figure) are added to the photometric dataset. This is represented by the dotted red contour lines corresponding to the completeness increments.

To achieve maximum completeness, quasar magnitudes need to be 2 to 3 magnitudes brighter than the EWS detection limit, depending on the redshift. The closer to the detection limit, and therefore the lower the SNR, the greater the confusion between quasars and brown dwarfs or galaxies at intermediate redshifts, preventing reliable identification.

Between redshift 7 and 8, completeness is lower than at higher redshift, up to $H_E \approx 21$, due to greater confusion with late-type dwarfs. Due to the strong increase in late-type dwarf flux in the reddest part of the optical band and the wide width of the *Euclid* I_E band, any late-type dwarf will appear significantly fainter in the I_E band than in any z-band image of comparable depth (see Fig. 1). As a result, z-band data from LSST or from the UNIONS survey (`https://www.skysurvey.cc/`) provide significantly improved discrimination against L- and T- dwarfs (dashed red curves in the left-hand plot of Fig. 10).

Fig. 10 also shows a loss in completeness around redshift 10 and between magnitudes 22 and 23 of H_E, compared with its value at lower redshift values. We attribute this loss in completeness mainly to an increase in confusion with early-type galaxies as their colours are very similar (see Fig. 11).

Finally, the analysis of completeness, including WISE data (bottom panel in Fig. 10) shows no significant improvement, in contrast to the improvement mentioned in Sect. 4.2 for the re-identification of known quasars at high redshift. Only a modest improvement in the decrease of the confusion with late-type dwarfs is perceptible for the brightest quasars at the lowest redshift. This is due to the fact that the depth of the WISE data (AB magnitude ~ 19 at 5σ in the $W1$ and $W2$ bands) does not match that of the *Euclid* data, making them unconstrained and allowing no significant improvements over the *Euclid* data alone, either above or below the WISE detection limit.

4.3.2. Redshift estimation

Following Sect. 4.1, the redshift estimated by OWL-z corresponds to the maximum posterior value of the parameter z, that is, the maximum posterior $\int W_q(z_{out}, H_E) dH_E$. We refer to it as "photometric redshift" or z_{out} hereafter.

The accuracy of photometric redshift has been estimated by comparing the true redshift injected in the simulation z_{true} to z_{out}. Fig. 12 displays this comparison for the redshift interval explored in the simulations presented in Sect. 4.3.1. To evaluate the accuracy of the photometric redshift, the sample has been limited to objects brighter than $H_E < 24$, identified as quasars using the same criterion as in Sect. 4.3.1, that is $P_q > 0.1$.

As shown in Fig. 12, an excellent correlation between z_{true} and z_{out} is found, irrespective of the photometric dataset. For the $I_E Y_E J_E H_E$ filter set, the Pearson correlation coefficient is found to be 0.97, whereas the Spearman correlation coefficient is 0.98, indicating a strong linear correlation. The same results are found for the z$Y_E J_E H_E$ dataset. The classical quality estimators used for photometric redshift yield the following results, with

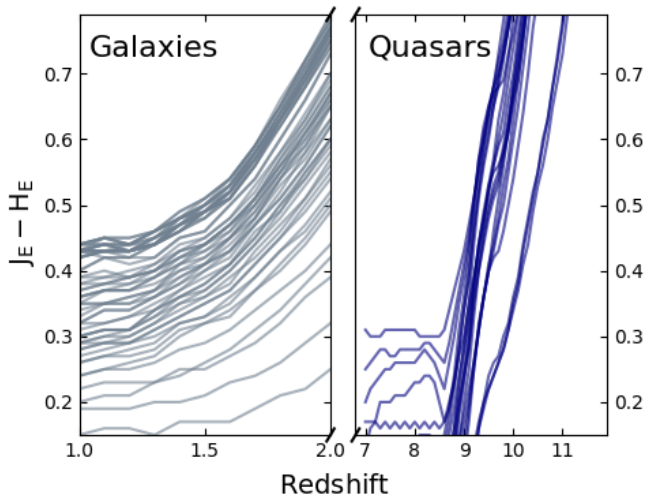


Fig. 11: Tracks of J_E-H_E colours as a function of redshift of Early-Type galaxies (left panel) and high redshift quasars (right panel). For the early-type galaxies, the lower in colour the track is the younger the galaxy.

$\Delta z = z_{out} - z_{true}$: $\sigma(\Delta z/(1+z)) = 0.030$, the median of $(\Delta z/(1+z))$ = -0.007, and the normalized median absolute deviation, which is less sensitive to outliers: $\sigma_{z,MAD} = 1.48 \times$ median $(|\Delta z|/(1+z))$ = 0.021. Regarding the fraction of outliers, defined in a conservative way as sources with $\Delta z > 0.1(1 + z_{true})$, it is found to be only 2.45% (2.23%) for the $I_E Y_E J_E H_E$(z$Y_E J_E H_E$) datasets. Only 13.7% (13.3%) of the sample exhibit $|\Delta z| > 0.05(1 + z_{true})$ for the $I_E Y_E J_E H_E$(z$Y_E J_E H_E$) datasets.

These results demonstrate the quality and the reliability of the photometric redshift obtained by OWL-z for sources identified as high-z quasars. This is of utmost importance, given the use expected for the output redshift in the selection of samples for spectroscopic follow-up. It is also important in the determination of the purity, as shown in Sect. 4.3.3.

4.3.3. Purity

The purity of the sample expresses the expected reliability of OWL-z in extracting true quasars from photometric observations of real fields, which is applied to the EWS in this particular case. Here, we use the *Euclid* bands $I_E Y_E J_E H_E$, the photometric depths and the variable SNR maps of *Euclid*.

We define the purity as:

$$P = \frac{1 + \mathrm{TP}}{1 + \mathrm{TP} + \mathrm{FP}}, \tag{19}$$

that is, the number of true positive identifications (TP) divided by the total number of quasar identifications (TP + FP), where FP stands for false positives. With the above definition, a purity of 100% can either mean that no object(quasar or contaminant) has been identified as a quasar or all the identified objects are quasars. To estimate the purity, we need to simulate as accurately as possible the content of the input photometric catalogues, including the contaminant populations of stars and galaxies in a realistic way. Bright high-z quasars are rare, according to the current LF. For this reason, we force the inclusion of such sources in the simulated catalogues, as explained below, to determine how difficult it will be to identify and study these sources if they ever exist. In order to achieve this, the simulations are conducted on a surface that is 10 times larger than that of the EWS. Subsequently, scaling back to the original surface (1000 deg^2) is performed, whereby only those sources that the LF predicts are accounted for. Note that the definition of the purity as given above accounts for the possible detection of these unique, bright and rare quasars.

The COSMOS2020 field (Weaver et al. 2022) represents an ideal base for this exercise, given the extended wavelength coverage and the exceptionally good quality of the photometric redshifts that can be used as spectroscopic redshifts for our needs. All sources in this field have been fitted using LePHARE (Arnouts & Ilbert 2011) and a library of reference templates, and this provided us with a best-fit template and a redshift. We use these results to compute the *Euclid* Optical and NIR photometry via a process we call "Euclidization" (additional details will be provided in *Euclid* preparation, EC & Garnett et al.). This process has also allowed us to simulate the SNR and associated photometric errors in each filter. This new *Euclid*-like COSMOS catalogue will be called hereafter E-COSMOS.

The original COSMOS catalogue covers 2 deg^2. To cover a statistically significant field of view, multiple realisations of the parent E-COSMOS catalogue are needed. This is done by randomly sorting the coordinates of the field centre within the EWS footprint in such a way that a full (not overlapping) large catalogue is created. This catalogue covers 1000 deg^2 separated into six different areas. The rationale behind this choice is that it represents the smallest surface area over which we can account for a sufficient number of quasars. Furthermore, this surface has been divided across the region of interest in order to accommodate the various scenarios in galactic latitude/longitude, as well as to account for the inevitable differences in resolution, however slight they may be. Fig. 13 and Table 6 display the location of these areas around the *Euclid* Footprint. It is worth noting that the location of the field is expected to have a direct impact on the performance, given the different distribution of contaminant stars on the one hand and the different SNR achieved across the EWS due to zodiacal light variations on the other hand (Euclid Collaboration: Scaramella et al. 2022). These effects are discussed below. The E-COSMOS catalogue has been cleansed of all stellar objects and is situated at a considerable distance from the Galactic plane. To properly account for the presence of contaminant stars, we randomly inject additional MLT dwarfs into the catalogues, following the distribution presented in Sect. 2.3, reproducing the number densities expected at the galactic coordinates where the simulated field is located. Since E-COSMOS does not contain any spectroscopically identified high-redshift quasar, this population has been randomly injected in the simulated fields, following the prescriptions presented in Sect. 4.3.1 to properly account for the SNR in the EWS. High-z quasars are randomly sorted according to their LF. Obviously, the number of bright $z > 7$ quasars is expected to be very small in the EWS field. In particular, the current LF does not predict any quasar brighter than $M_{1450} < -23$ in the entire field. For these rare bright objects, if they exist, we force the measurement of the purity as explained above. To better capture this population, 10 realisations of the entire *Euclid* footprint are performed, allowing us to retrieve some quasars towards the brightest luminosities and up to redshifts $z > 10$. However, in order to facilitate consistent comparison of the statistical performance estimators C and P, it is necessary to normalise all populations to the same surface area (15,000 deg^2).

OWL-z has been run on these simulated catalogues. The preselection criteria and $P_q > 0.1$ threshold applied are identical

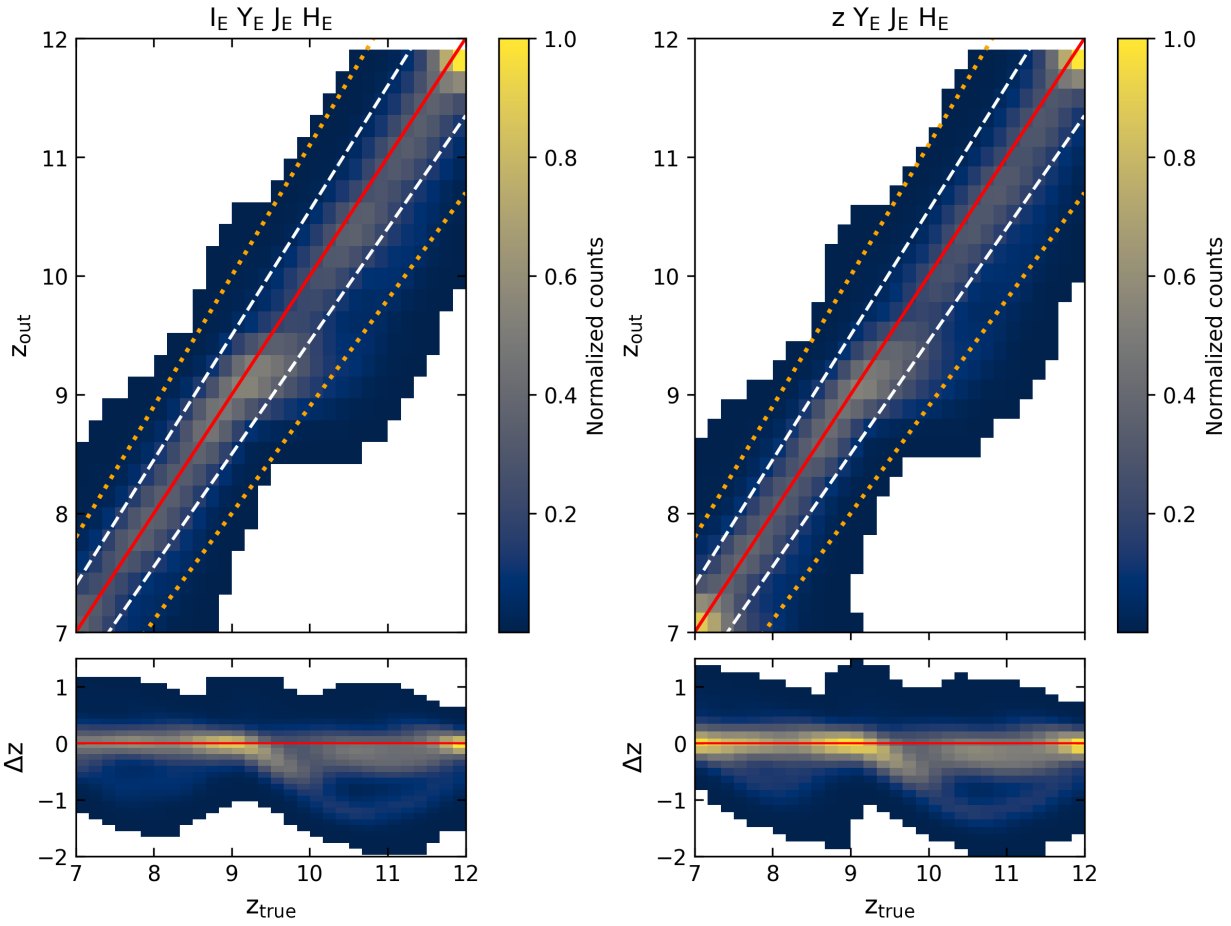


Fig. 12: Top panel shows the comparison between the true redshift and the redshift estimated by OWL-z with both *Euclid* $I_{\rm E}Y_{\rm E}J_{\rm E}H_{\rm E}$ dataset (left) and z$Y_{\rm E}J_{\rm E}H_{\rm E}$ dataset (right). Red solid lines display the 1:1 bisector; the thresholds corresponding to $|\Delta z| > 0.1(1+z_{true})$ and $|\Delta z| > 0.05(1+z_{true})$ are also displayed to guide the eye as orange dotted and white dashed lines, respectively. The bottom panel shows the residual $\Delta z = z_{out} - z_{true}$ as a function of redshift, with the red line corresponding to $\Delta z = 0$

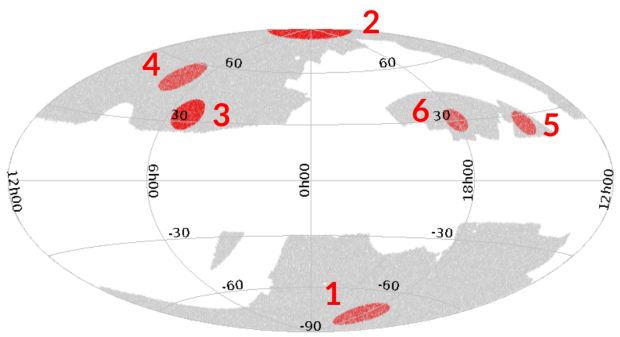


Fig. 13: (*red*) The location of the regions of focus used in the computation of the purity how in Galactic coordinates and galactic projection. Details on these surfaces can be found in Table 6. (*grey*) the EWS DR6 footprint.

to those used for the completeness in Sect. 4.3.1. We study the variation of the purity as a function of redshift (z_{out}) and apparent magnitude in the reference filter. The results presented in Fig. 14 provide a summary of the performance of OWL-z in terms of completeness and purity. In each bin, the figure displays the values of C and P. Additionally, it illustrates another performance estimator, the F-measure, which will be defined in Sect. 4.3.4. As shown in this figure, the code is expected to achieve a high purity for the brightest sources, irrespective of the redshift. The influence of the precise choice for the threshold in P_q is discussed in Sect. 5.1.

4.3.4. Global performance

In order to evaluate the global performance of OWL-z by combining both the completeness and purity measurements, we introduce the F-measure as follows:

$$F = 2\frac{CP}{C+P}, \tag{20}$$

where P is the purity and C is the completeness, according to the definitions in Sect. 4.3.1 and Sect. 4.3.3 respectively. The F-measure is a statistical measurement employed to assess the performance of a classifier (Baeza-Yates & Ribeiro-Neto 2011).

The classification performance—specifically, the code's ability to identify high-z quasars—is evaluated using the F-measure, which indicates the quality of the classification. A higher F-measure indicates a superior classification performance. We use F-measure hereafter as a performance metric.

The results summarising the performance of OWL-z in terms of F-measure are also provided in Fig. 14. As shown in this figure, high values of F-measure are obtained for bright ($H_E < 21$) sources at all redshifts between $7 < z < 11$, indicating that the performance in the spectroscopic confirmation of bright rare quasars is expected to be high. On the contrary, the F-measure drops significantly towards fainter magnitudes and high redshifts, following the drop in the purity value. Note that satisfying results can still be achieved for $H_E < 23$ and $z < 8$. Also, the performance depends on the choice of P_q threshold, as discussed in Sect. 5.1.

5. Discussion

In this section, we study the influence of different parameters on the performance of OWL-z, such as the threshold used to select quasars as a function of redshift and magnitude. We also give some guidelines to optimize the selection of high-z quasars and their spectroscopic follow-up. The discussion in this section is based on the performance expected on the EWS.

5.1. Influence of the selection threshold

In the previous section, we consider an object as a quasar candidate when the posterior probability $\mathbb{P}$ is above the threshold $\zeta = 0.1$. This parameter clearly impacts the performance because lower values will improve the completeness while inducing worse purity values and vice-versa. There is a trade-off to find to optimize the performance. Here, we discuss the influence of ζ on F-measure.

To this end, we use the data discussed in Sect. 4.3.3, change the value of ζ to select quasar candidates and recalculate the F-measure in each bin. The results are presented in Fig. 14 for $\zeta = 0.1$ and in Fig. 15 for $\zeta = 0.9$.

Two main trends are observed for sources respectively brighter/fainter than $H_E = 22$.

Increasing the threshold improves purity, albeit at the expense of completeness for faint sources ($H_E > 22$). For bright sources ($H_E < 22$), the completeness is almost unaffected by increasing the classification threshold, thanks to the ability of the code to compute high P_q values for bright quasars.

For faint sources, the performance of OWL-z drops rapidly towards the faintest magnitudes above $H_E = 22$, irrespective of the redshift bin and ζ due to low SNR values, which contribute to a rapid decrease in purity, even when completeness is still good. The analysis of the results shows that early-type galaxies are the main contributors to contamination at these faint magnitudes ($H_E > 22$). Conversely, at bright magnitudes ($H_E < 22$ and brighter), the SNR is high and enables adequate classification between stars, galaxies and quasars with good completeness, and higher values of ζ improve the purity and the values of the F-measure increase.

In Fig. 14, the $8 < z < 10$ redshift bins for bright candidates ($19 < H_E < 22$) are clearly affected by low purity values. The main sources of contamination in this area are early-type galaxies at redshifts $1 < z < 2$ due to the confusion between the 4000 Å break at intermediate redshift and the Lyman-α break at high redshift, as explained in Sect. 2.4. There is also a minor contribution from contamination by late L-type dwarfs. Increasing the classification threshold has the immediate effect of reducing this contamination and, therefore, improving the purity. As completeness remains high, this results in a significant improvement in $F-$measure values.

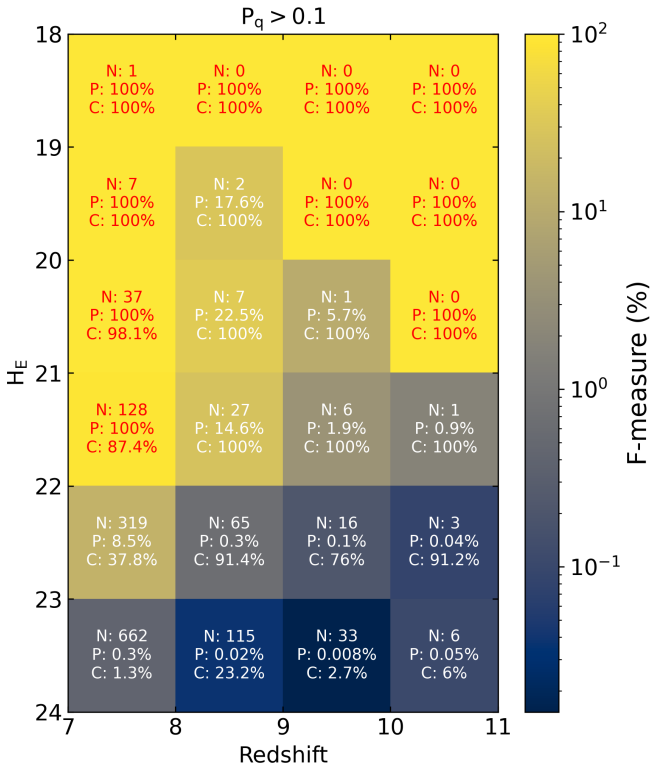


Fig. 14: Classification performance of OWL-z determined using the threshold $\zeta = 0.1$, in each magnitude and redshift bin, N is the number of quasars in the bin, P is the purity, and C is the completeness. The colour code indicates the performance measurement F-measure in each bin

This analysis suggests that an adaptive threshold can be used to optimise the $F-$ measure values in the magnitude/redshift parameter space. This is the subject of the next section.

5.2. Optimising the identification of $z > 7$ quasars

We now investigate if the value of the F-measure can be further increased by adjusting the selection threshold ζ in each redshift and magnitude bin. To this end, we use the data discussed in Sect. 4.2, change the value of ζ in increments of 0.1 and select the value that maximises F in each bin. The result is summarised in Fig. 16, showing the value of ζ that maximises F in each bin of magnitude H_E and redshift z_{out}. The results show that, for bright magnitudes ($H_E < 21$), a low threshold value $\zeta = 0.1$ is sufficient to maximise F-measure. This is due to the robustness of the selection for high SNR values irrespective of the redshift, as shown previously. On the contrary, for faint magnitudes ($H_E > 21$), the value of ζ needs to be increased to compensate for the decline in purity, requiring ζ values as high as 0.9 towards the faintest magnitudes.

In conclusion, adopting a variable threshold enables the optimization of the F-measure and, consequently, the effectiveness of photometric and spectroscopic follow-up campaigns based on OWL-z selected candidates. Finding the best threshold value ζ for a quasar candidate with magnitude H_E and a redshift z_{out} from any other survey data can be done by following the same prescription presented in this article in the case of EWS.

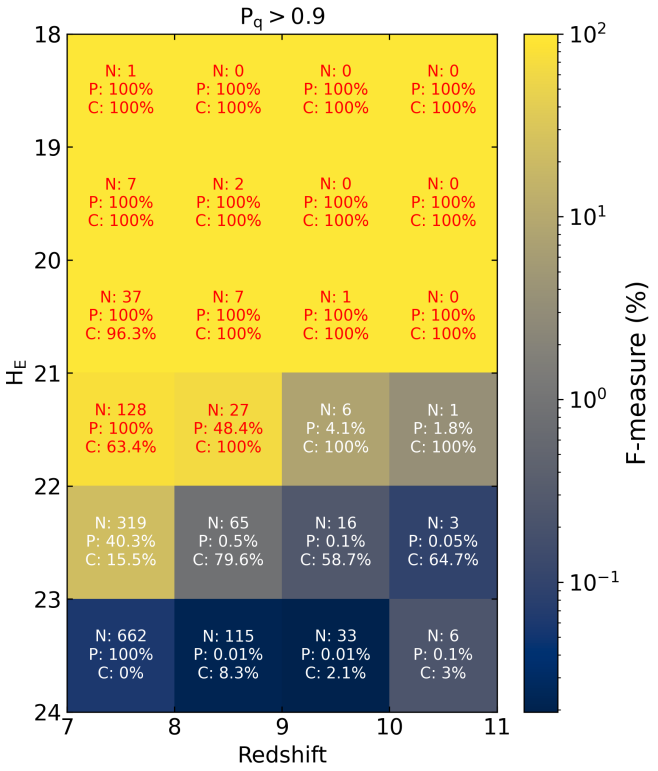


Fig. 15: Classification performance of OWL-z determined using the threshold $\zeta = 0.9$, in each bin we report N: the number of quasars injected in the bin, P: the purity and C: the completeness. The colour code indicates the performance measurement F-measure in each bin.

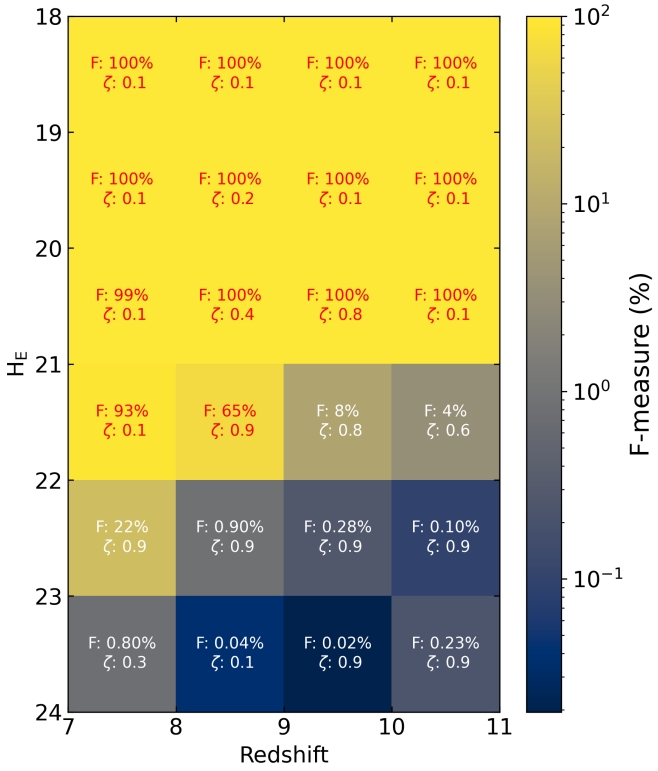


Fig. 16: Classification performance of OWL-z on the magnitude H_E and redshift z_{out} plane. The threshold value of ζ that maximises F-measure is given for each bin.

5.3. Influence of the position in the EWS footprint

In this section, we evaluate the effect of different locations in the EWS footprint. As described in Euclid Collaboration: Scaramella et al. 2022, different zodiacal light levels across the footprint modify the SNR for a given magnitude in each filter, and different sky positions and Galactic coordinates introduce varying stellar densities, as described in Sect. 2.3. To this end, we select six different regions in the *Euclid* footprint, representative of different zodiacal light levels and Galactic coordinates. These regions are shown in Fig. 13, and Table 6 lists their Galactic coordinates and SNR values at reference magnitudes in all Euclid filters.

According to the results presented in Sect. 4, the confusion by MLT dwarfs is only sensible in the redshift interval $z = [7, 8]$, where there is also contamination by early-type galaxies. Therefore, we focus the analysis on this redshift bin because it is more susceptible to being affected by the precise distribution assumed for MLT stars. In addition, this redshift bin is also the most populated by quasars, meaning that the results are expected to be statistically more significant than in the other bins. For each one of these regions, we compute simulated catalogues following the prescriptions in Sect. 4.3.3, and we run OWL-z to derive the value of F-measure.

The results are shown in Fig. 17 and display minor variations in the values of F-measure. Regions 5 and 6 exhibit a slight decline in F-measure within the magnitude bin $[21, 22]$ compared to the others. This can be attributed either to the fact that these regions possess the lowest Galactic latitudes ($b = 26$ and $b = 30$) or to the most unfavourable SNR values. According to the results presented in Sect. 4, the Comparison with region 3, which is also at a low Galactic latitude but has higher SNR values, suggests that SNR variation is the dominant factor in variations in F-measure. In other words, brown dwarf contamination does not appear to depend significantly on Galactic latitude within the *Euclid* footprint, which is consistent with figure 4 and with the fact that up to magnitude ≈ 22 contamination is dominated by late-type L and T dwarfs at distances small compared to the scale height of the Galactic disc.

5.4. Influence of thick disc MLT

We aim to evaluate the robustness of our selection process when adding contamination by thick disc MLTs that are not included in the model. We evaluate the impact on the F-measure of adding a contribution of thick-disk brown dwarfs to the input catalogue, following the prescriptions described in section 4.3, and ignoring differences in brown dwarf populations between the thin and thick discs. To do so, we employ the density function outlined in Sect. 2 to generate a population of MLT dwarfs from the thick disc, in addition to the previously added MLT dwarfs from the thin disc, and we analyse the impact on F-measure on the six regions introduced previously. The results are presented in Fig. 18, comparing two scenarios: one where only thin disc MLT dwarfs are injected in the input catalogue (shown by yellow lines) and another where thick disc MLT dwarfs are added (shown by blue lines). The trends on these graphs can be qualitatively explained by the relative density of stars between the thin and thick discs as a function of magnitude. At low magnitudes, the thin disk dominates at all Galactic latitudes, and there are no differences

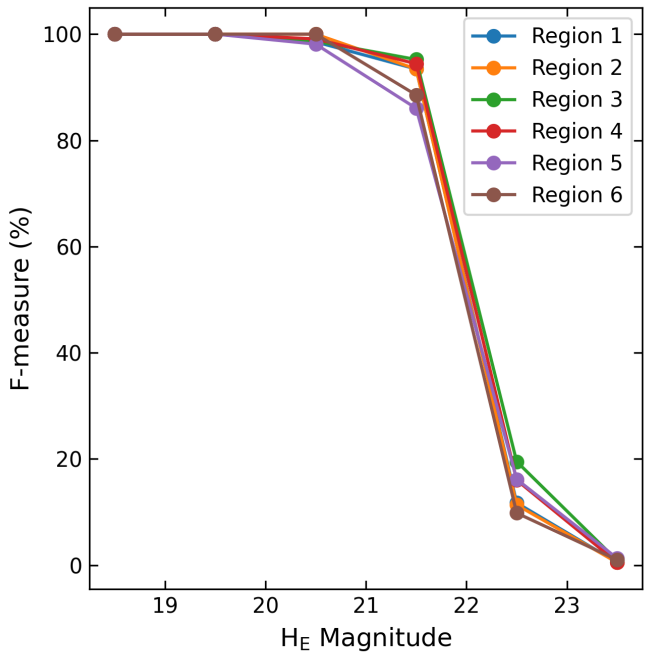


Fig. 17: F-measure evolution as a function of the $H_{\rm E}$ magnitude in each of the regions shown in Table 6, computed in the $7 < z < 8$ interval.

between the two scenarios. At magnitudes above 21, however, and at high Galactic latitudes only (see figure 4), the thick disk population begins to contribute to the counts and introduce a difference between the two scenarios (regions 1, 2, 4 and to a lesser extent region 3, which is at a relatively low Galactic latitude but closest to the Galactic centre in longitude). Overall, this analysis serves as a sanity check in confirming that the precise modelling of the thick disc is unlikely to significantly affect the results for the EWS, but would be warranted for deeper surveys.

5.5. Comparison with previous works

The results obtained in this article can be compared to previous similar works, at least qualitatively. A proper comparison will need to run the various codes on the same simulated catalogues, which is currently impossible because not all codes are publicly available. The Bayesian classification approach is similar to that adopted by Pipien et al. 2018b for the analysis of the CFHQSIR survey (Pipien et al. 2018a) searching for quasars at $6.5 < z < 7.5$. Our results in the lowest redshift bin are comparable to those of Pipien et al. 2018b. Note that Pipien et al. 2018b do not include the population of contaminant galaxies, a choice that has no impact on their results because the dominant contamination in this bin comes mainly from LT stars, as shown in Sect. 4.

Our work also compares to and complements the study conducted by Euclid Collaboration: Barnett et al. 2019 on the EWS using a similar Bayesian classification method and the same contaminant populations of stars and galaxies. With respect Euclid Collaboration: Barnett et al. 2019, the work with OWL-z extends the study beyond $z > 9$, and introduces different performance metrics in addition to the completeness, namely purity and F-measure (combining completeness and purity). Our definition of purity has also been adapted to account for rare quasars beyond $z > 9$ that could be found in the EWS if they ever exist. We also

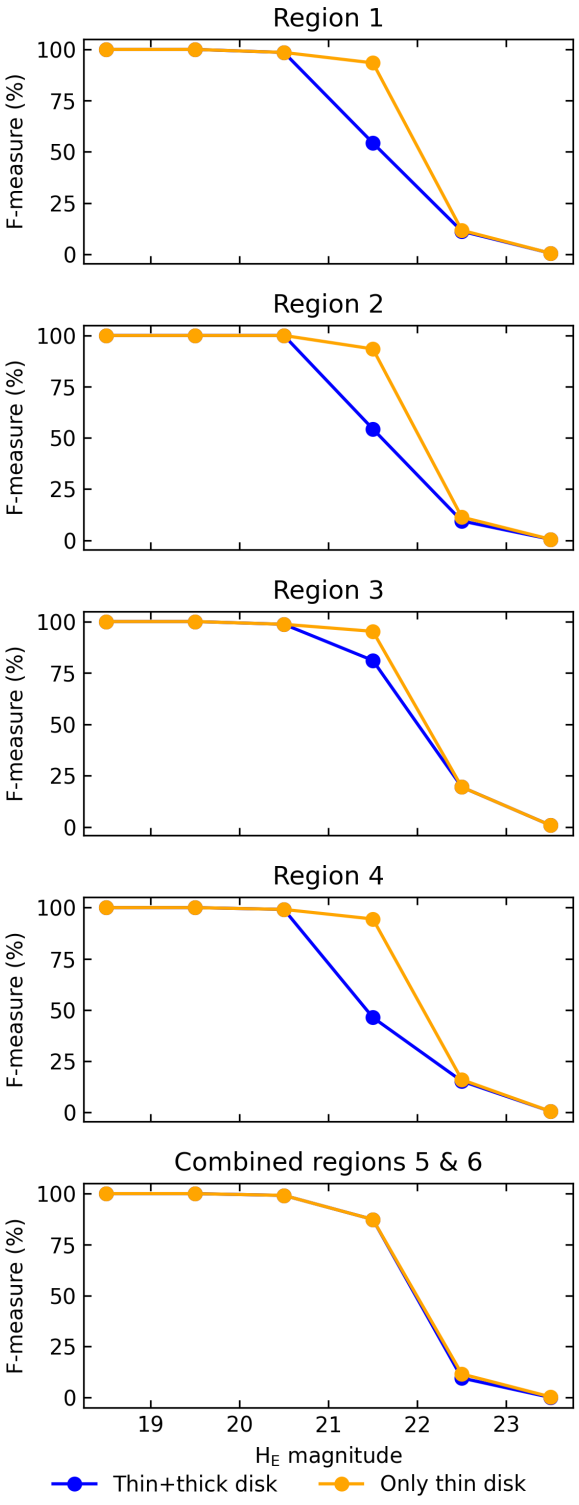


Fig. 18: F-measure evolution as a function of the $H_{\rm E}$ magnitude in each of the regions shown in Table 6, computed in the $7 < z < 8$ interval. Two cases are shown for the contribution of MLT stars. In yellow is presented the evolution when only the thin disc contribution is accounted for, both in the simulation and the Bayesian modelling. In blue, the same result is shown when the contribution of the thick disc is added to the simulations.

use updated prescriptions for the quasar LF and the contaminant populations. It is worth noting that we have used here the EWS

Table 6: Properties of areas and catalogues simulated across the EWS footprint at reference magnitudes of 24 for $Y_E J_E H_E$ and 24.5 for I_E. The SNR values are averaged over each surface.

Region	Surface (deg^2)	l(deg)	b(deg)	SNR$_{I_E}$	SNR$_{Y_E}$	SNR$_{J_E}$	SNR$_{H_E}$
1	200	270	−75	17.5	7.0	8.45	7.9
2	200	180	85	16.5	6.6	7.8	7.4
3	200	78	35	17.8	7.0	8.5	7.7
4	200	107	55	18.6	7.4	8.8	8.22
5	100	227	26	13.8	5.6	6.6	6.1
6	100	270	30	14.3	5.9	7.1	6.5

SNR maps, making our predictions closer to the real observations. The performance of OWL-z in terms of completeness in the common redshift range with Euclid Collaboration: Barnett et al. 2019 is found to be better than in this previous work. The reason for this improvement is mainly the fact that the global SNR is expected to be better than assumed by Euclid Collaboration: Barnett et al. 2019. On the contrary, the effect of the detailed 2D modelling of the SNR in the EWS has little impact on the performance of OWL-z and similar codes, possibly because the depth is not dramatically different across the survey, as discussed in Sect. 5.3.

Possibly the main advantage of OWL-z with respect to other existing codes is its versatility in adapting to different photometric surveys, as shown in Sect. 4.2. The code will be publicly available and open to access.

6. Conclusions

The conclusions of this work can be summarised as follows:

- This work is intended to be the reference article presenting OWL-z, a public Bayesian code aiming at identifying $z > 7$ quasars in massive photometric surveys. OWL-z is based on the classification of sources as high-z quasars or contaminants, the latter being MLT stars or early-type galaxies at intermediate redshifts. The populations of contaminants have been carefully modelled for the needs of the Bayesian model selection. Although the code has been tuned to be used on the EWS, it can be easily adapted to different photometric surveys, as shown in this work.
- OWL-z has been able to re-identify all spectroscopically-confirmed quasars at $z > 7$. This exercise has also demonstrated the versatility of OWL-z regarding its application to different photometric catalogues.
- The performance of OWL-z based on F-measure, a metric combining completeness and purity, has been estimated using simulations for the EWS in a wide range of redshifts ($7 < z < 12$) and reference magnitudes ($18 < H_E < 24.5$). The results show that OWL-z reaches full performance for bright sources ($H_E < 22$) irrespective of the redshift, meaning that the performance in the spectroscopic confirmation of bright rare quasars selected by OWL-z is expected to be high. The performance drops significantly towards fainter magnitudes and high redshifts.
- The threshold value for the selection of quasars can be applied in the post-treatment phase, meaning that this value can be tuned to optimize the selection of $z > 7$ quasars. For instance, a small/large value can be used for the brightest/faintest sources ($H_E < 21$ and > 21, respectively) to optimize the performance. Given the steep change of the optimum threshold on the magnitude versus redshift plane, a relatively crude cut on this plane could be adopted to optimize the spectroscopic follow-up.
- A uniform performance is obtained with OWL-z across the EWS footprint, irrespective of the Galactic coordinates and relative SNR values.
- The results of the code remain robust despite inevitable uncertainties in the detailed modelling of the contaminating stellar populations, in particular, the thin and thick discs.

Acknowledgements. This work has been carried out within the framework of the Euclid Consortium. We would like to thank the members of the Euclid Science Working Group Primeval Universe, particularly the Quasar Work Package, for their feedback and comments during the preparation of this article. Special thanks are extended to Eduardo Bañados and Daniel Mortlock for their invaluable input and cooperation. This research has also benefited from the use of quasar templates provided by Paul Hewett. Additionally, we acknowledge the contributions of Yōsuke Matsuoka, Feige Wang, and Jinyi Yang, who kindly provided spectra for quasars at $z > 7$. This research has made use of the Spanish Virtual Observatory project (https://svo.cab.inta-csic.es) funded by MCIN/AEI/10.13039/501100011033/ through grant PID2020-112949GB-I00, in particular for the SpeX Prism Library containing spectral templates for brown dwarfs. We also acknowledge the use of the SVO Filter Profile Service "Carlos Rodrigo", funded by MCIN/AEI/10.13039/501100011033/ through grant PID2020-112949GB-I00 This research has been possible thanks to the computing facilities operated by CeSAM data centre at LAM, Marseille, France.

References

Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv e-prints, arXiv:1902.05569
Ali, S. S., De Propris, R., Chung, C., et al. 2024, ApJ, 966, 50
Arnouts, S. & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate, Astrophysics Source Code Library, record ascl:1108.009
Bañados, E., Venemans, B. P., Decarli, R., et al. 2016a, ApJS, 227, 11
Bañados, E., Venemans, B. P., Decarli, R., et al. 2016b, ApJS, 227, 11
Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018a, Nature, 553, 473
Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018b, Nature, 553, 473
Bañados, E., Venemans, B. P., Morganson, E., et al. 2014, AJ, 148, 14
Baeza-Yates, R. & Ribeiro-Neto, B. 2011, Modern Information Retrieval the Concepts and Technology Behind Search
Bakos, G. Á., Sahu, K. C., & Németh, P. 2002, ApJS, 141, 187
Becker, G. D., Bolton, J. S., Madau, P., et al. 2015, MNRAS, 447, 3402
Belladitta, S., Moretti, A., Caccianiga, A., et al. 2020, A&A, 635, L7
Bennett, J. S., Sijacki, D., Costa, T., Laporte, N., & Witten, C. 2024, MNRAS, 527, 1033
Best, W. M. J., Liu, M. C., Dupuy, T. J., & Magnier, E. A. 2017, ApJ, 843, L4
Bochanski, J. J., Hawley, S. L., Covey, K. R., et al. 2010, AJ, 139, 2679
Bolzonella, M., Miralles, J. M., & Pelló, R. 2000, A&A, 363, 476
Bovy, J., Hogg, D. W., & Roweis, S. T. 2011, Annals of Applied Statistics, 5, 1657
Bruzual, G. & Charlot, S. 2003, MNRAS, 344, 1000
Burgasser, A. J. 2014, in Astronomical Society of India Conference Series, Vol. 11, Astronomical Society of India Conference Series, 7–16
Burgasser, A. J., Bezanson, R., Labbe, I., et al. 2024, The Astrophysical Journal, 962, 177
Burgasser, A. J. & Splat Development Team. 2017, in Astronomical Society of India Conference Series, Vol. 14, Astronomical Society of India Conference Series, 7–12
Caballero, J. A., Burgasser, A. J., & Klement, R. 2008, A&A, 488, 181
Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, ApJ, 533, 682
Chabrier, G. 2003, PASP, 115, 763
Dey, A., Schlegel, D. J., Lang, D., et al. 2019, AJ, 157, 168
Dupuy, T. J. & Liu, M. C. 2012, ApJS, 201, 19

Euclid Collaboration, Mellier, Y., Abdurro'uf, et al. 2024, arXiv e-prints, arXiv:2405.13491
Euclid Collaboration, Schirmer, M., Jahnke, K., et al. 2022, A&A, 662, A92
Euclid Collaboration: Barnett, R., Warren, S. J., Mortlock, D. J., et al. 2019, A&A, 631, A85
Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, A&A, 662, A112
Euclid Collaboration: van Mierlo, S. E., Caputi, K. I., Ashby, M., et al. 2022, A&A, 666, A200
Fan, X., Strauss, M. A., Becker, R. H., et al. 2006, AJ, 132, 117
Gaia Collaboration, Recio-Blanco, A., Kordopatis, G., et al. 2023, A&A, 674, A38
Hainline, K. N., D'Eugenio, F., Sun, F., et al. 2024a, arXiv e-prints, arXiv:2407.08781
Hainline, K. N., Helton, J. M., Johnson, B. D., et al. 2024b, ApJ, 964, 66
Hogg, D. W., Baldry, I. K., Blanton, M. R., & Eisenstein, D. J. 2002, arXiv e-prints, astro
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019a, ApJ, 873, 111
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019b, ApJ, 873, 111
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193
Leitherer, C., Schaerer, D., Goldader, J. D., et al. 1999, ApJS, 123, 3
Maiolino, R., Scholtz, J., Witstok, J., et al. 2024, Nature, 627, 59
Matsuoka, Y., Iwasawa, K., Onoue, M., et al. 2019a, ApJ, 883, 183
Matsuoka, Y., Onoue, M., Iwasawa, K., et al. 2023, ApJ, 949, L42
Matsuoka, Y., Onoue, M., Kashikawa, N., et al. 2019b, ApJ, 872, L2
Mortlock, D. J., Patel, M., Warren, S. J., et al. 2012, MNRAS, 419, 390
Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Nature, 474, 616
Nanni, R., Hennawi, J. F., Wang, F., et al. 2022, MNRAS, 515, 3224
Oke, J. B. & Gunn, J. E. 1983, ApJ, 266, 713
Pipien, S., Basa, S., Cuby, J. G., et al. 2018a, A&A, 616, A55
Pipien, S., Cuby, J. G., Basa, S., et al. 2018b, A&A, 617, A127
Reid, I. N. 2013, Brown Dwarfs, ed. T. D. Oswalt & M. A. Barstow (Dordrecht: Springer Netherlands), 337–395
Robert, C. P. 2007, The Bayesian choice: from decision-theoretic foundations to computational implementation, 2nd edn. (Springer)
Ryan, R. E., J. & Reid, I. N. 2016, AJ, 151, 92
Schechter, P. 1978, The luminosity function for galaxies and the clustering of galaxies
Scholtz, J., Witten, C., Laporte, N., et al. 2024, A&A, 687, A283
Skrzypek, N., Warren, S. J., & Faherty, J. K. 2016, A&A, 589, A49
Stern, D., Kirkpatrick, J. D., Allen, L. E., et al. 2007, The Astrophysical Journal, 663, 677
Sysoliatina, K. & Just, A. 2022, A&A, 666, A130
Temple, M. J., Hewett, P. C., & Banerji, M. 2021, MNRAS, 508, 737
Trenti, M. & Stiavelli, M. 2008, ApJ, 676, 767
Vieira, K., Carraro, G., Korchagin, V., et al. 2022, The Astrophysical Journal, 932, 28
Vieira, K., Korchagin, V., Carraro, G., & Lutsenko, A. 2023, Galaxies, 11, 77
Wang, F., Yang, J., Fan, X., et al. 2021, ApJ, 907, L1
Wang, F., Yang, J., Fan, X., et al. 2018, ApJ, 869, L9
Warren, S. J., Cross, N. J. G., Dye, S., et al. 2007, arXiv e-prints, astro
Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, The Astrophysical Journal Supplement Series, 258, 11
Wenzl, L., Schindler, J.-T., Fan, X., et al. 2021, AJ, 162, 72
Wilkins, S. M., Stanway, E. R., & Bremer, M. N. 2014, Monthly Notices of the Royal Astronomical Society, 439, 1038
Willott, C. J., Delorme, P., Reylé, C., et al. 2010, AJ, 139, 906
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868
Yang, J., Wang, F., Fan, X., et al. 2020, ApJ, 897, L14
Yang, J., Wang, F., Fan, X., et al. 2019, AJ, 157, 236
Zucca, E., Ilbert, O., Bardelli, S., et al. 2006, A&A, 455, 879

[1] Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France
[2] Aix Marseille Univ, CNRS, I2M, Marseille, France
[3] Aix Marseille Univ, CNRS, LIS, Marseille, France
[4] Canada-France-Hawai'i Telescope, Waimea, Hawai'i
[5] Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191, Gif-sur-Yvette, France
[6]Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA Leiden, The Netherlands

Bibliography

Sivia, D. & Skilling, J. 2006, Data Analysis: A Bayesian Tutorial

Abdo, A. A., Ackermann, M., Ajello, M., et al. 2010, Astrophysical Journal, 722, 520

Akeson, R., Armus, L., Bachelet, E., et al. 2019, arXiv e-prints, arXiv:1902.05569

Albrecht, A., Bernstein, G., Cahn, R., et al. 2006, arXiv e-prints, astro

Ali, S. S., De Propris, R., Chung, C., et al. 2024, Astrophysical Journal, 966, 50

Amendola, L., Appleby, S., Avgoustidis, A., et al. 2018, Living Reviews in Relativity, 21, 2

Antonucci, R. 1993, Annual Review of Astron and Astrophys, 31, 473

Arnouts, S. & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate, Astrophysics Source Code Library, record ascl:1108.009

Assef, R. J., Stern, D., Kochanek, C. S., et al. 2013, Astrophysical Journal, 772, 26

Atek, H., Gavazzi, R., Weaver, J. R., et al. 2024, arXiv e-prints, arXiv:2405.13504

Bañados, E., Venemans, B. P., Decarli, R., et al. 2016, Astrophysical Journal, Supplement, 227, 11

Bañados, E., Venemans, B. P., Mazzucchelli, C., et al. 2018, Nature, 553, 473

Baeza-Yates, R. & Ribeiro-Neto, B. 2011, Modern Information Retrieval the Concepts and Technology Behind Search

Bakos, G. Á., Sahu, K. C., & Németh, P. 2002, Astrophysical Journal, Supplement, 141, 187

Barbary, K. 2016, Journal of Open Source Software, 1, 58

Barkana, R. & Loeb, A. 2001a, Physics Reports, 349, 125

Barkana, R. & Loeb, A. 2001b, Physics Reports, 349, 125

Battaglieri, M., Belloni, A., Chou, A., et al. 2017, arXiv e-prints, arXiv:1707.04591

Bayes, T. & Price, R. 1763, Philosophical Transactions of the Royal Society, 53, 370, communicated by Mr. Price in a letter to John Canton, A.M., F.R.S.

Belladitta, S., Moretti, A., Caccianiga, A., et al. 2020, Astronomy and Astrophysics, 635, L7

Bertin, E. & Arnouts, S. 1996, Astronomy and Astrophysics, Supplement, 117, 393

Bertin, E., Mellier, Y., Radovich, M., et al. 2002, in Astronomical Society of the Pacific Conference Series, Vol. 281, Astronomical Data Analysis Software and Systems XI, ed. D. A. Bohlender, D. Durand, & T. H. Handley, 228

Best, W. M. J., Liu, M. C., Dupuy, T. J., & Magnier, E. A. 2017, Astrophysical Journal, Letters, 843, L4

Betoule, M., Kessler, R., Guy, J., et al. 2014, Astronomy and Astrophysics, 568, A22

Blandford, R. D. & Königl, A. 1979, Astrophysical Journal, 232, 34

Blandford, R. D. & Rees, M. J. 1978, Physica Scripta, 17, 265

Bochanski, J. J., Hawley, S. L., Covey, K. R., et al. 2010, Astronomical Journal, 139, 2679

Bouwens, R. J., Illingworth, G. D., Oesch, P. A., et al. 2015, Astrophysical Journal, 811, 140

Bovy, J., Hennawi, J. F., Hogg, D. W., et al. 2011a, Astrophysical Journal, 729, 141

Bovy, J., Hogg, D. W., & Roweis, S. T. 2011b, Annals of Applied Statistics, 5, 1657

Bowler, R. A. A., Dunlop, J. S., McLure, R. J., & McLeod, D. J. 2017, Monthly Notices of the RAS, 466, 3612

Boyle, B. J., Shanks, T., Croom, S. M., et al. 2000, Monthly Notices of the RAS, 317, 1014

Boyle, B. J., Shanks, T., & Peterson, B. A. 1988, Monthly Notices of the RAS, 235, 935

Brandt, W. N. & Hasinger, G. 2005, Annual Review of Astron and Astrophys, 43, 827

Brotherton, M. S. 2014, Nature, 513, 181

Bruton, S., Scarlata, C., Haardt, F., et al. 2023, Astrophysical Journal, 953, 29

Bruzual, G. & Charlot, S. 2003, Monthly Notices of the RAS, 344, 1000

Burgasser, A. J. 2014, in Astronomical Society of India Conference Series, Vol. 11, Astronomical Society of India Conference Series, 7–16

Burgasser, A. J., Bezanson, R., Labbe, I., et al. 2024, The Astrophysical Journal, 962, 177

Burgasser, A. J. & Splat Development Team. 2017, in Astronomical Society of India Conference Series, Vol. 14, Astronomical Society of India Conference Series, 7–12

Caballero, J. A., Burgasser, A. J., & Klement, R. 2008, Astronomy and Astrophysics, 488, 181

Calzetti, D., Armus, L., Bohlin, R. C., et al. 2000, Astrophysical Journal, 533, 682

Colless, M., Dalton, G., Maddox, S., et al. 2001, Monthly Notices of the RAS, 328, 1039

Collin-Souffrin, S. 1979, Annales de Physique, 4, 177

Collin-Souffrin, S., Dumnont, S., Joly, M., & Pequignot, D. 1986, Astronomy and Astrophysics, 166, 27

Collinson, J. S., Ward, M. J., Landt, H., et al. 2016, Monthly Notices of the Royal Astronomical Society, 465, 358

Cuillandre, J. C., Bertin, E., Bolzonella, M., et al. 2024, arXiv e-prints, arXiv:2405.13496

Czerny, B. & Elvis, M. 1987, Astrophysical Journal, 321, 305

Davé, R., Anglés-Alcázar, D., Narayanan, D., et al. 2019, Monthly Notices of the RAS, 486, 2827

Dayal, P., Ferrara, A., & Dunlop, J. S. 2013, Monthly Notices of the RAS, 430, 2891

Delvecchio, I., Daddi, E., Aird, J., et al. 2020, Astrophysical Journal, 892, 17

Dey, A., Schlegel, D. J., Lang, D., et al. 2019, Astronomical Journal, 157, 168

Duncan, K., Conselice, C. J., Mortlock, A., et al. 2014, Monthly Notices of the RAS, 444, 2960

Dupuy, T. J. & Liu, M. C. 2012, Astrophysical Journal, Supplement, 201, 19

Eddington, A. S. 1926, The Internal Constitution of the Stars

Efstathiou, G., Ellis, R. S., & Peterson, B. A. 1988, Monthly Notices of the RAS, 232, 431

Eisenstein, D. J., Zehavi, I., Hogg, D. W., et al. 2005, Astrophysical Journal, 633, 560

Emerson, J. P., Sutherland, W. J., McPherson, A. M., et al. 2004, The Messenger, 117, 27

Euclid Collaboration, Mellier, Y., Abdurro'uf, et al. 2024a, arXiv e-prints, arXiv:2405.13491

Euclid Collaboration, Schirmer, M., Jahnke, K., et al. 2022, Astronomy and Astrophysics, 662, A92

Euclid Collaboration, Zalesky, L., McPartland, C. J. R., et al. 2024b, arXiv e-prints, arXiv:2408.05296

Euclid Collaboration: Barnett, R., Warren, S. J., Mortlock, D. J., et al. 2019, Astronomy and Astrophysics, 631, A85

Euclid Collaboration: Scaramella, R., Amiaux, J., Mellier, Y., et al. 2022, Astronomy and Astrophysics, 662, A112

Euclid Collaboration: van Mierlo, S. E., Caputi, K. I., Ashby, M., et al. 2022, Astronomy and Astrophysics, 666, A200

Falomo, R., Pesce, J. E., & Treves, A. 1993, Astrophysical Journal, Letters, 411, L63

Fan, X., Carilli, C. L., & Keating, B. 2006a, Annual Review of Astron and Astrophys, 44, 415

Fan, X., Carilli, C. L., & Keating, B. 2006b, Annual Review of Astron and Astrophys, 44, 415

Fan, X., Narayanan, V. K., Lupton, R. H., et al. 2001a, Astronomical Journal, 122, 2833

Fan, X., Narayanan, V. K., Lupton, R. H., et al. 2001b, Astronomical Journal, 122, 2833

Fan, X., Strauss, M. A., Schneider, D. P., et al. 2001c, Astronomical Journal, 121, 54

Fan, X., Strauss, M. A., Schneider, D. P., et al. 2001d, Astronomical Journal, 121, 54

Fanaroff, B. L. & Riley, J. M. 1974, Monthly Notices of the RAS, 167, 31P

Fanidakis, N., Baugh, C. M., Benson, A. J., et al. 2012, Monthly Notices of the RAS, 419, 2797

Fechner, C. & Reimers, D. 2007, Astronomy and Astrophysics, 463, 69

Feigelson, E. D., Martin, A. L., McNeill, C. J., Broos, P. S., & Garmire, G. P. 2009, Astronomical Journal, 138, 227

Ferrarese, L. & Ford, H. 2005a, Space Science Reviews, 116, 523

Ferrarese, L. & Ford, H. 2005b, Space Science Reviews, 116, 523

Ferrarese, L. & Merritt, D. 2000, Astrophysical Journal, Letters, 539, L9

Finkelstein, S. L., Ryan, Jr., R. E., Papovich, C., et al. 2015, Astrophysical Journal, 810, 71

Fisher, R. A. 1922, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222, 309

Fitzpatrick, E. L. & Massa, D. 2007, Astrophysical Journal, 663, 320

Frank, S., Mathur, S., Pieri, M., & York, D. G. 2010, Astronomical Journal, 140, 817

Frye, B. & Broadhurst, T. 1998, Astrophysical Journal, Letters, 499, L115

Fukugita, M., Hogan, C. J., & Peebles, P. J. E. 1998, The Astrophysical Journal, 503, 518

Furlanetto, S. R., McQuinn, M., & Hernquist, L. 2006, Monthly Notices of the RAS, 365, 115

Gaia Collaboration, Recio-Blanco, A., Kordopatis, G., et al. 2023a, Astronomy and Astrophysics, 674, A38

Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2023b, Astronomy and Astrophysics, 674, A1

Gaur, H., Gu, M., Ramya, S., & Guo, H. 2019, Astronomy and Astrophysics, 631, A46

Gauss, C. F. & Stewart, G. W. 1995, Theory of the Combination of Observations Least Subject to Errors, Part One, Part Two, Supplement (Society for Industrial and Applied Mathematics)

Gebhardt, K., Bender, R., Bower, G., et al. 2000, Astrophysical Journal, Letters, 539, L13

Giommi, P., Padovani, P., Polenta, G., et al. 2012, Monthly Notices of the RAS, 420, 2899

Gnedin, N. Y. & Madau, P. 2022, Living Reviews in Computational Astrophysics, 8, 3

Goto, T. 2006, Monthly Notices of the RAS, 371, 769

Goulding, A. D. et al. 2023, Astrophys. J. Lett., 955, L24

Grazian, A., Boutsia, K., Giallongo, E., et al. 2023, Astrophysical Journal, 955, 60

Grazian, A., Giallongo, E., Boutsia, K., et al. 2024, Astrophysical Journal, 974, 84

Grazian, A., Giallongo, E., Boutsia, K., et al. 2018, Astronomy and Astrophysics, 613, A44

Greene, J. E., Labbe, I., Goulding, A. D., et al. 2024, Astrophysical Journal, 964, 39

Greig, B., Mesinger, A., Haiman, Z., & Simcoe, R. A. 2017, Monthly Notices of the RAS, 466, 4239

Gunn, J. E. & Peterson, B. A. 1965a, Astrophysical Journal, 142, 1633

Gunn, J. E. & Peterson, B. A. 1965b, Astrophysical Journal, 142, 1633

Guth, A. H. & Steinhardt, P. J. 1984, Scientific American, 250, 116

Haiman, Z., Abel, T., & Rees, M. J. 2000, Astrophysical Journal, 534, 11

Hainline, K. N., D'Eugenio, F., Sun, F., et al. 2024a, arXiv e-prints, arXiv:2407.08781

Hainline, K. N., Helton, J. M., Johnson, B. D., et al. 2024b, Astrophysical Journal, 964, 66

Häring, N. & Rix, H.-W. 2004, Astrophysical Journal, Letters, 604, L89

Hasinger, G. 2008, Astronomy and Astrophysics, 490, 905

Hasinger, G., Miyaji, T., & Schmidt, M. 2005, Astronomy and Astrophysics, 441, 417

Hewett, P. C., Foltz, C. B., & Chaffee, F. H. 1995, Astronomical Journal, 109, 1498

Hewett, P. C., Warren, S. J., Leggett, S. K., & Hodgkin, S. T. 2006, Monthly Notices of the RAS, 367, 454

Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, Astrophysical Journal, Supplement, 208, 19

Hoag, A., Bradač, M., Huang, K., et al. 2019, Astrophysical Journal, 878, 12

Hogg, D. W., Baldry, I. K., Blanton, M. R., & Eisenstein, D. J. 2002, arXiv e-prints, astro

Hubble, E. 1929, Proceedings of the National Academy of Science, 15, 168

Hughes, A. C. N. 2023, Statistical Methods in the Era of Large Astronomical Surveys

Hui, L. & Gnedin, N. Y. 1997, Monthly Notices of the RAS, 292, 27

Hummels, C. B., Bryan, G. L., Smith, B. D., & Turk, M. J. 2013, Monthly Notices of the RAS, 430, 1548

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019a, Astrophysical Journal, 873, 111

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019b, Astrophysical Journal, 873, 111

Jarvis, M. J., Bonfield, D. G., Bruce, V. A., et al. 2013, Monthly Notices of the RAS, 428, 1281

Jaynes, E. T. & Bretthorst, G. L. 2003, Probability Theory

Jones, M. L., Hickox, R. C., Black, C. S., et al. 2016, Astrophysical Journal, 826, 12

Kaiser, N., Aussel, H., Burke, B. E., et al. 2002, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4836, Survey and Other Telescope Technologies and Discoveries, ed. J. A. Tyson & S. Wolff, 154–164

Kakiichi, K., Ellis, R. S., Laporte, N., et al. 2018, Monthly Notices of the RAS, 479, 43

Kannan, R., Garaldi, E., Smith, A., et al. 2022, Monthly Notices of the RAS, 511, 4005

Katgert, P., Mazure, A., Perea, J., et al. 1996, Astronomy and Astrophysics, 310, 8

Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, Monthly Notices of the RAS, 341, 33

Kellermann, K. I., Sramek, R., Schmidt, M., Shaffer, D. B., & Green, R. 1989, Astronomical Journal, 98, 1195

Koekemoer, A. M., Faber, S. M., Ferguson, H. C., et al. 2011, Astrophysical Journal, Supplement, 197, 36

Kormendy, J. & Ho, L. C. 2013, Annual Review of Astron and Astrophys, 51, 511

Kormendy, J. & Richstone, D. 1995, Annual Review of Astron and Astrophys, 33, 581

Larson, R. L., Finkelstein, S. L., Kocevski, D. D., et al. 2023, Astrophysical Journal, Letters, 953, L29

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv e-prints, arXiv:1110.3193

Le Fèvre, O., Vettolani, G., Garilli, B., et al. 2005, Astronomy and Astrophysics, 439, 845

Leitherer, C., Schaerer, D., Goldader, J. D., et al. 1999, Astrophysical Journal, Supplement, 123, 3

Li, W., Inayoshi, K., Onoue, M., et al. 2024, Astrophysical Journal, 969, 69

Li, Y.-R., Wang, J.-M., & Ho, L. C. 2012, Proceedings of the International Astronomical Union, 8, 259–260

Linde, A. D. 1982, Physics Letters B, 114, 431

Lupton, R., Blanton, M. R., Fekete, G., et al. 2004, Publications of the ASP, 116, 133

Madau, P., Giallongo, E., Grazian, A., & Haardt, F. 2024, Astrophysical Journal, 971, 75

Madau, P. & Haardt, F. 2015, Astrophysical Journal, Letters, 813, L8

Magorrian, J., Tremaine, S., Richstone, D., et al. 1998, Astronomical Journal, 115, 2285

Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024a, Astronomy and Astrophysics, 691, A145

Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024b, Astronomy and Astrophysics, 691, A145

Maiolino, R., Scholtz, J., Curtis-Lake, E., et al. 2024c, Astronomy and Astrophysics, 691, A145

Maiolino, R., Scholtz, J., Witstok, J., et al. 2024d, Nature, 627, 59

Maiolino, R., Scholtz, J., Witstok, J., et al. 2024e, Nature, 627, 59

Maiolino, R., Übler, H., Perna, M., et al. 2024f, Astronomy and Astrophysics, 687, A67

Malkan, M. A. & Sargent, W. L. W. 1982, Astrophysical Journal, 254, 22

Marconi, A. & Hunt, L. K. 2003, Astrophysical Journal, Letters, 589, L21

Margon, B. 1999, Philosophical Transactions of the Royal Society of London Series A, 357, 93

Matsuoka, Y., Iwasawa, K., Onoue, M., et al. 2019a, Astrophysical Journal, 883, 183

Matsuoka, Y., Onoue, M., Iwasawa, K., et al. 2023, Astrophysical Journal, Letters, 949, L42

Matsuoka, Y., Onoue, M., Kashikawa, N., et al. 2019b, Astrophysical Journal, Letters, 872, L2

Matthee, J., Naidu, R. P., Brammer, G., et al. 2024, Astrophysical Journal, 963, 129

McCracken, H. J., Milvang-Jensen, B., Dunlop, J., et al. 2012, Astronomy and Astrophysics, 544, A156

McDonald, P., Miralda-Escudé, J., Rauch, M., et al. 2000, Astrophysical Journal, 543, 1

McGreer, I. D., Mesinger, A., & D'Odorico, V. 2015, Monthly Notices of the RAS, 447, 499

McLure, R. J. & Dunlop, J. S. 2002, Monthly Notices of the RAS, 331, 795

Meiksin, A. A. 2009, Rev. Mod. Phys., 81, 1943

Miralda-Escudé, J., Haehnelt, M., & Rees, M. J. 2000, Astrophysical Journal, 530, 1

Mortlock, D. J., Patel, M., Warren, S. J., et al. 2012, Monthly Notices of the RAS, 419, 390

Mortlock, D. J., Warren, S. J., Venemans, B. P., et al. 2011, Nature, 474, 616

Nakane, M., Ouchi, M., Nakajima, K., et al. 2024, Astrophysical Journal, 967, 28

Nanni, R., Hennawi, J. F., Wang, F., et al. 2022, Monthly Notices of the RAS, 515, 3224

Netzer, H. 2015, Annual Review of Astron and Astrophys, 53, 365

Ni, Y., Di Matteo, T., Bird, S., et al. 2022, Monthly Notices of the Royal Astronomical Society, 513, 670

Oesch, P. A., Brammer, G., van Dokkum, P. G., et al. 2016, Astrophysical Journal, 819, 129

Ouchi, M., Shimasaku, K., Furusawa, H., et al. 2010, Astrophysical Journal, 723, 869

Padovani, P., Alexander, D. M., Assef, R. J., et al. 2017, Astronomy and Astrophysics Reviews, 25, 2

Padovani, P., Resconi, E., Giommi, P., Arsioli, B., & Chang, Y. L. 2016, Monthly Notices of the RAS, 457, 3582

Pallottini, A., Ferrara, A., Decataldo, D., et al. 2019, Monthly Notices of the RAS, 487, 1689

Peebles, P. J. E. 1993, Principles of Physical Cosmology

Pelló, R., Kneib, J. P., Le Borgne, J. F., et al. 1999, Astronomy and Astrophysics, 346, 359

Pello, R., Le Borgne, J.-F., Soucail, G., Mellier, Y., & Sanahuja, B. 1991, Astrophysical Journal, 366, 405

Perlmutter, S., Aldering, G., Goldhaber, G., et al. 1999, Astrophysical Journal, 517, 565

Pipien, S., Basa, S., Cuby, J. G., et al. 2018a, Astronomy and Astrophysics, 616, A55

Pipien, S., Cuby, J. G., Basa, S., et al. 2018b, Astronomy and Astrophysics, 617, A127

Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, Astronomy and Astrophysics, 641, A6

Puchwein, E., Haardt, F., Haehnelt, M. G., & Madau, P. 2019, Monthly Notices of the RAS, 485, 47

Rauch, M. 1998, Annual Review of Astron and Astrophys, 36, 267

Reed, S. L., McMahon, R. G., Martini, P., et al. 2017, Monthly Notices of the Royal Astronomical Society, 468, 4702

Reid, I. N. 2013, Brown Dwarfs, ed. T. D. Oswalt & M. A. Barstow (Dordrecht: Springer Netherlands), 337–395

Rieke, M. J., Kelly, D., & Horner, S. 2005, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 5904, Cryogenic Optical Systems and Instruments XI, ed. J. B. Heaney & L. G. Burriesci, 1–8

Riess, A. G., Filippenko, A. V., Challis, P., et al. 1998, Astronomical Journal, 116, 1009

Robert, C. P. 2007, The Bayesian choice: from decision-theoretic foundations to computational implementation, 2nd edn. (Springer)

Robertson, B. E., Ellis, R. S., Dunlop, J. S., McLure, R. J., & Stark, D. P. 2010, Nature, 468, 49

Robertson, B. E., Ellis, R. S., Furlanetto, S. R., & Dunlop, J. S. 2015, Astrophysical Journal, Letters, 802, L19

Rodighiero, G., Brusa, M., Daddi, E., et al. 2015, Astrophysical Journal, Letters, 800, L10

Rosdahl, J., Katz, H., Blaizot, J., et al. 2018, Monthly Notices of the RAS, 479, 994

Ryan, R. E., J. & Reid, I. N. 2016, Astronomical Journal, 151, 92

Sanders, D. B., Phinney, E. S., Neugebauer, G., Soifer, B. T., & Matthews, K. 1989, Astrophysical Journal, 347, 29

Sanders, D. B., Soifer, B. T., Elias, J. H., et al. 1988, Astrophysical Journal, 325, 74

Schaye, J., Crain, R. A., Bower, R. G., et al. 2015, Monthly Notices of the RAS, 446, 521

Schaye, J., Dalla Vecchia, C., Booth, C. M., et al. 2010, Monthly Notices of the RAS, 402, 1536

Schaye, J., Theuns, T., Rauch, M., Efstathiou, G., & Sargent, W. L. W. 2000, Monthly Notices of the RAS, 318, 817

Schechter, P. 1978, The luminosity function for galaxies and the clustering of galaxies

Schenker, M. A., Ellis, R. S., Konidaris, N. P., & Stark, D. P. 2014, Astrophysical Journal, 795, 20

Scheuer, P. A. G. 1965, Nature, 207, 963

Schlafly, E. F. & Finkbeiner, D. P. 2011, Astrophysical Journal, 737, 103

Schmidt, M. 1963, Nature, 197, 1040

Schmidt, M. & Green, R. F. 1983, Astrophysical Journal, 269, 352

Scholtz, J., Witten, C., Laporte, N., et al. 2024, Astronomy and Astrophysics, 687, A283

Schroeder, J., Mesinger, A., & Haiman, Z. 2013, Monthly Notices of the RAS, 428, 3058

Seyfert, C. K. 1943, Astrophysical Journal, 97, 28

Shankar, F., Weinberg, D. H., & Miralda-Escudé, J. 2013, Monthly Notices of the RAS, 428, 421

Shklovskii, I. S. 1965, Soviet Astronomy, 8, 638

Shulevski, A., Morganti, R., Barthel, P. D., et al. 2015, Astronomy and Astrophysics, 583, A89

Simpson, C. M., Pakmor, R., Marinacci, F., et al. 2016, Astrophysical Journal, Letters, 827, L29

Skrzypek, N., Warren, S. J., & Faherty, J. K. 2016, Astronomy and Astrophysics, 589, A49

Sohn, J., Fabricant, D. G., Geller, M. J., Hwang, H. S., & Diaferio, A. 2020, Astrophysical Journal, 902, 17

Stern, D., Kirkpatrick, J. D., Allen, L. E., et al. 2007, The Astrophysical Journal, 663, 677

Trenti, M. & Stiavelli, M. 2008, Astrophysical Journal, 676, 767

Ulrich, M.-H., Maraschi, L., & Urry, C. M. 1997, Annual Review of Astron and Astrophys, 35, 445

Urry, C. M. & Padovani, P. 1995, Publications of the ASP, 107, 803

Vieira, K., Carraro, G., Korchagin, V., et al. 2022, The Astrophysical Journal, 932, 28

Vieira, K., Korchagin, V., Carraro, G., & Lutsenko, A. 2023, Galaxies, 11, 77

Volonteri, M., Habouzit, M., & Colpi, M. 2021, Nature Reviews Physics, 3, 732

Volonteri, M., Pfister, H., Beckmann, R. S., et al. 2020, Monthly Notices of the RAS, 498, 2219

Volonteri, M., Trebitsch, M., Dubois, Y., et al. 2024, arXiv e-prints, arXiv:2408.12854

Wang, F., Yang, J., Fan, X., et al. 2021, Astrophysical Journal, Letters, 907, L1

Wang, F., Yang, J., Fan, X., et al. 2019, The Astrophysical Journal, 884, 30

Wang, F., Yang, J., Fan, X., et al. 2018, Astrophysical Journal, Letters, 869, L9

Warren, S. J., Cross, N. J. G., Dye, S., et al. 2007, arXiv e-prints, astro

Warren, S. J., Hewett, P. C., & Osmer, P. S. 1994, Astrophysical Journal, 421, 412

Weaver, J. R., Kauffmann, O. B., Ilbert, O., et al. 2022, The Astrophysical Journal Supplement Series, 258, 11

Weaver, J. R., Taamoli, S., McPartland, C. J. R., et al. 2024, arXiv e-prints, arXiv:2405.13505

Wenzl, L., Schindler, J.-T., Fan, X., et al. 2021, Astronomical Journal, 162, 72

Weymann, R. J., Carswell, R. F., & Smith, M. G. 1981, Annual Review of Astron and Astrophys, 19, 41

Whitler, L. R., Mason, C. A., Ren, K., et al. 2020, Monthly Notices of the RAS, 495, 3602

Wilkins, S. M., Stanway, E. R., & Bremer, M. N. 2014, Monthly Notices of the Royal Astronomical Society, 439, 1038

Willott, C. J., Delorme, P., Reylé, C., et al. 2010, Astronomical Journal, 139, 906

Wise, J. H. 2019, arXiv e-prints, arXiv:1907.06653

Wolf, C., Wisotzki, L., Borch, A., et al. 2003, Astronomy and Astrophysics, 408, 499

Worseck, G., Prochaska, J. X., Hennawi, J. F., & McQuinn, M. 2016, Astrophysical Journal, 825, 144

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010a, Astronomical Journal, 140, 1868

Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010b, Astronomical Journal, 140, 1868

Yang, J., Wang, F., Fan, X., et al. 2020, Astrophysical Journal, Letters, 897, L14

Yang, J., Wang, F., Fan, X., et al. 2019, Astronomical Journal, 157, 236

Zucca, E., Ilbert, O., Bardelli, S., et al. 2006, Astronomy and Astrophysics, 455, 879

Acknowledgements

I would first like to warmly thank the reviewers in the jury of my PhD defense, Giulia Rodighiero and Hervé Dole. Their careful reading, insightful comments, and constructive suggestions were essential in improving this work.

I would also like to thank the examiners, Marta Volonteri, Julyan Arbel, and Hakim Atek. Marta's publications opened my eyes to new perspectives and deepened my love for active galactic nuclei. Julyan took the time to read my manuscript and genuinely got excited about my project. Hakim provided guidance throughout my thesis follow-up committees, offering invaluable advice and encouraging words along the way.

I am very grateful to my thesis supervisors, Roser Pelló and Jean-Gabriel Cuby, Pierre Pudlo and François-Xavier Dupé, the supervision was sometimes complex, but it ultimately led to the completion of a fruitful thesis. Their guidance, patience, and scientific rigor were essential to this work.

I want to express my sincere gratitude to my colleagues from the GECO team at LAM. Olivier, Sylvain, Patrice, Raphael, Stéphane, your encouragement, stimulating discussions, and your guidance has been invaluable throughout my PhD. Thank you especially for helping me understand, often from first principles, important astronomical concepts and methods. Thank you all for your kindness, patience, and support not only helped me progress scientifically but also made my experience at LAM truly joyful.

Finally, I want to thank all the members of the Laboratoire d'Astrophysique de Marseille for providing such a stimulating environment. Their encouragement, good humour, and daily scientific discussions made these years of my PhD memorable.

Je tiens à adresser un remerciement tout particulier à Françoise Maxant, qui est sans doute l'une des personnes les plus formidables et attentionnées que j'aie jamais rencontrées. Sa gentillesse, sa bienveillance et son soutien ont illuminé mes journées pendant ma thèse (et après). Même un simple " bonjour " échangé avec elle suffisait à me faire sentir bien. Ses mots encourageants, son sourire et sa disponibilité ont été un énorme soutien moral tout au long de ce parcours. Merci du fond du cœur Françoise.

Je souhaite également remercier Jean-Charles Lambert, à qui je dois non seulement le bon fonctionnement de mon code (fruit de mon travail de thèse) mais aussi, en grande partie, mon équilibre mental tout au long de ce parcours. Ses longues discussions, ses conseils professionnels et son accompagnement attentif ont été précieux.

Je tiens également à remercier Laurent Jorda, avec qui j'ai pu construire une belle amitié professionnelle fondée sur notre amour commun pour la diffusion scientifique. Nous avons travaillé ensemble lors de la Fête de la Science et à d'autres occasions, transmettant notre passion au grand public. Ces expériences ont été très enrichissantes et, grâce à lui, j'ai pu saisir de nombreuses opportunités pour partager mes connaissances scientifiques. Sa présence et son soutien m'ont été extrêmement précieux pour prendre du recul par rapport à ma thèse et m'ouvrir à d'autres perspectives.

Je souhaite également remercier toute l'administration, et en particulier Solange, Miguel et Christine. Leur disponibilité, leur gentillesse et leurs conseils ont été d'une grande aide tout au long de ma thèse.

To my fellow phd students and others:

First and foremost, I want to thank Diana and Gayathri. My PhD would have never been possible without them. They listened to me, supported me, and tolerated me in my lowest moments; always with patience, kindness, and humour. They saw me when I was overcaffeinated, scattered, anxious, or exhausted, and somehow still managed to love me through all of it. Even when distance was imposed on us after they left for their postdocs, our friendship remained strong; we kept in touch, sharing everything from frustrations to laughter, and they never stopped being a grounding presence in my life. The bond we built goes far beyond work: they are truly my chosen family.

Another memeber of my acquired family: Camille, she helped me make sense of everything when nothing seemed to make sense at all. She completes me in a way that's hard to describe. I hope she knows, through these words, how much her friendship means to me and how deeply she shaped this journey.

Martin deserves a very special mention: my office mate during the hardest part of my PhD and a fellow Euclid member. We travelled together for work, shared long discussions, and supported each other through the ups and downs of research life. He was the sound, stable presence I always needed ,calm, reliable, and endlessly kind. Beyond being a wonderful officemate, he was also my running partner, helping me stay motivated when things felt overwhelming. His companionship brought balance to my days, and I am deeply grateful for that.

To Ilias, Sia, Paolo, Alexis, Mathilde and all the other PhD students and postdocs: thank you for all the wonderful social moments we shared. From dinners and barbecues to Eurovision evenings, poker nights, metal concerts (James ;)) and game nights, you made life outside of research vibrant and full of laughter. The hikes we did together, especially the long and hard ones! were unforgettable adventures that bonded us in the best way possible. And truly, to all the PhD students: every outing, every trip, every shared moment made this experience richer and more human.

To all of you: thank you for being there, for making me laugh, for helping me hold on, and for reminding me that even in the most demanding times, friendship and connection are what truly carry us forward. I end this with a famous Gandalf quote: "All we have to decide is what to do with the time that is given us", so "Allons-y, Alonso!" (10th).