

# Angular clustering properties of the DESI QSO target selection using DR9 Legacy Imaging Surveys

Edmond Chaussidon<sup>1</sup>,<sup>\*</sup> Christophe Yèche,<sup>1</sup> Nathalie Palanque-Delabrouille,<sup>1</sup> Arnaud de Mattia,<sup>1</sup> Adam D. Myers,<sup>2</sup> Mehdi Rezaie<sup>3</sup>, Ashley J. Ross,<sup>4</sup> Hee-Jong Seo<sup>3</sup>, David Brooks,<sup>5</sup> Enrique Gaztañaga,<sup>6</sup> Robert Kehoe,<sup>7</sup> Michael E. Levi,<sup>8</sup> Jeffrey A. Newman,<sup>9,10</sup> Gregory Tarlé<sup>11</sup> and Kai Zhang<sup>8</sup>

<sup>1</sup>IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France

<sup>2</sup>Department of Physics and Astronomy, University of Wyoming, 1000 E. University Ave., Laramie, WY 82071, USA

<sup>3</sup>Department of Physics and Astronomy, Ohio University, Athens, OH 45701, USA

<sup>4</sup>Center of Cosmology and AstroParticle Physics, The Ohio State University, Columbus, OH 43210, USA

<sup>5</sup>Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK

<sup>6</sup>Institute of Space Sciences (ICE, CSIC), Campus UAB, Carrer de Can Magrans sn, E-08193 Bellaterra, Barcelona, Spain

<sup>7</sup>Department of Physics, Southern Methodist University, 3215 Daniel Ave., Dallas, TX 75205, USA

<sup>8</sup>Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

<sup>9</sup>Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260, USA

<sup>10</sup>Pittsburgh Particle Physics, Astrophysics, and Cosmology Center (PITT PACC), University of Pittsburgh, Pittsburgh, PA 15260, USA

<sup>11</sup>Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA

Accepted 2021 November 5. Received 2021 November 4; in original form 2021 August 14

## ABSTRACT

The quasar target selection for the upcoming survey of the Dark Energy Spectroscopic Instrument (DESI) will be fixed for the next 5 yr. The aim of this work is to validate the quasar selection by studying the impact of imaging systematics as well as stellar and galactic contaminants, and to develop a procedure to mitigate them. Density fluctuations of quasar targets are found to be related to photometric properties such as seeing and depth of the Data Release 9 of the DESI Legacy Imaging Surveys. To model this complex relation, we explore machine learning algorithms (random forest and multilayer perceptron) as an alternative to the standard linear regression. Splitting the footprint of the Legacy Imaging Surveys into three regions according to photometric properties, we perform an independent analysis in each region, validating our method using extended Baryon Oscillation Spectroscopic Survey (eBOSS) EZ-mocks. The mitigation procedure is tested by comparing the angular correlation of the corrected target selection on each photometric region to the angular correlation function obtained using quasars from the Sloan Digital Sky Survey (SDSS) Data Release 16. With our procedure, we recover a similar level of correlation between DESI quasar targets and SDSS quasars in two-thirds of the total footprint and we show that the excess of correlation in the remaining area is due to a stellar contamination that should be removed with DESI spectroscopic data. We derive the Limber parameters in our three imaging regions and compare them to previous measurements from SDSS and the 2dF QSO Redshift Survey.

**Key words:** methods: data analysis – surveys – large-scale structure of Universe – cosmology: observations.

## 1 INTRODUCTION

The Dark Energy Spectroscopic Instrument (DESI)<sup>1</sup> will conduct a spectroscopic survey of a third of the sky (i.e. 14 000 deg<sup>2</sup>) over a period of 5 yr (DESI Collaboration 2016). At the end of the survey, DESI is expected to produce a catalogue of 35 million objects covering redshifts up to  $\sim 4$ . This will be the largest catalogue of spectroscopic sources ever observed, increasing by an order of magnitude the number of sources observed by the Sloan Digital Sky Survey<sup>2</sup> (SDSS; Ahumada et al. 2020).

The DESI survey will provide the most detailed 3D map of the matter distribution in the Universe as traced by four different target classes: galaxies from a Bright Galaxy Survey (BGS) in the redshift range  $0.05 < z < 0.4$ ; luminous red galaxies (LRGs) in the range  $0.4 < z < 1.1$ ; emission line galaxies (ELGs) in the range  $0.6 < z < 1.6$ ; and quasars or quasi-stellar objects (QSOs) at  $z > 0.8$ . The DESI collaboration will use these data sets to perform analyses of baryon acoustic oscillations and redshift-space distortions to constrain the expansion history of the Universe and the growth of structure. Similar studies were conducted on the final SDSS data set (Data Release 16 – DR16) by Gil-Marín et al. (2020), Wang et al. (2020), Bautista et al. (2021) (using LRGs), Tamone et al. (2020), Raichoor et al. (2021), De Mattia et al. (2021) (using ELGs), Neveux et al. (2020), Hou et al. (2021) (using QSOs), and du Mas des Bourboux et al. (2020) (using the Ly $\alpha$  forest). DESI should also enable precise measurements of

\* E-mail: edmond.chaussidon@cea.fr

<sup>1</sup><https://www.desi.lbl.gov/>

<sup>2</sup><https://www.sdss.org/dr16/>

the sum of the neutrino masses, as well as investigations of modified gravity and theories of inflation.

A promising approach to probe inflation is through the tiny imprint left on the matter power spectrum by inflation-induced primordial non-Gaussianity. As noted for instance in Ross et al. (2013), Castorina et al. (2019), and Mueller et al. (2021), this measurement is known to be limited by systematic effects on large scales, most of which are due to imaging systematics imprinted on the density of spectroscopic targets. To prepare for upcoming clustering analyses with DESI, it is therefore crucial to study and be able to mitigate the imaging systematics that impact target selection.

The aim of this paper is to mitigate systematics for the DESI QSO target selection. We compute the angular clustering properties of the QSO targets in order to validate the selection method and to provide a control sample of QSO targets stripped of residual biases from imaging systematics or selection criteria. This is particularly important as the QSO target selection is known to be strongly contaminated by stars in addition to being impacted by imaging systematics. This work also serves as a crucial input to the selection of the QSO targets for DESI, which will soon be finalized for the 5-yr duration of the survey, in order to avoid strong imprints into the spectroscopic QSO sample due to imaging systematics that occur during the target selection. Finally, it is also a proof of concept. Since the stellar contamination of the target selection will be removed after the spectroscopic survey, the same method could be used to mitigate systematics in the DESI spectroscopic QSO sample.

Different strategies have been developed to deal with imaging systematic effects and improve the reliability of clustering studies. In this work, we follow the approach that was used for SDSS studies (Myers et al. 2006; Ross et al. 2011, 2017, 2020; Ho et al. 2012; Raichoor et al. 2021) and for Dark Energy Survey (DES Collaboration 2021) studies (Leistedt et al. 2016; Elvin-Poole et al. 2018). This method models the variation of target density as a linear function of imaging features (see e.g. Myers et al. 2015; Prakash et al. 2016) in order to remove fluctuations caused by imaging systematics. A correction weight is then computed and applied to the data. Since this method smooths density fluctuations, one needs to check to what level the mitigation procedure affects the cosmological signal. Another less common approach is based on mode projection (Rybicki & Press 1992; Tegmark et al. 1998; Leistedt et al. 2013; Elsner, Leistedt & Peiris 2016; Kalus et al. 2019): modes (in Fourier space) or pixels (in configuration space) are assigned increased variance where the systematics map exhibits large values, such that the covariance matrix has larger values in the presence of systematics. This is a robust method, which, however, only mitigates systematics using also a linear combination of the imaging maps. It cannot model the non-linear effects that are now observed, as illustrated in Ho et al. (2012). The correction-weights and mode-projection strategies can be combined in a common framework as explained in Weaverdyck & Huterer (2021).

To circumvent the assumption that only known features can completely explain all imaging systematics, one can also use a forward-modelling approach (Suchyta et al. 2016; Burleigh 2018; Kong et al. 2020). Such an approach accounts for source detection and target selection processes in an end-to-end fashion, by injecting fake galaxies into raw images, running source detection on the images, and applying target selection algorithms to the resulting sources.

Some studies of how imaging systematics affect DESI target selection have already been undertaken. For instance, Kitanidis et al. (2020) gave a first overview of imaging systematics for different DESI target classes selected from Data Release 7 (DR7) of the DESI Legacy Imaging Surveys (see e.g. Dey et al. 2019), and found that the

DESI QSO target sample suffered from strong contaminating effects. In another study of DR7 of the Legacy Imaging Surveys, Rezaie et al. (2020) analysed how imaging systematics affect the extended Baryon Oscillation Spectroscopic Survey (eBOSS)-like ELG selection, and used an artificial neural network to mitigate non-linear effects.

The aim of this work is to analyse the final QSO target selection that will be used for the next 5 yr by DESI. The selection is based on the Data Release 9 (DR9) of the Legacy Imaging Surveys, which is greatly improved compared to prior data releases. For instance, DR9 covers the full footprint of the Legacy Imaging Surveys (about 20 000 deg<sup>2</sup>), incorporates improved background fitting, and uses the latest definition of bright star and large-galaxy masks. To mitigate systematic effects on the QSO selection, we will explore machine learning approaches based on random forests (RFs) and neural networks (NNs), which we will compare to each other and to a traditional linear treatment. We will test our mitigation process by measuring the angular properties of the QSO target selection, before and after applying the weights derived from these techniques. The density fluctuation of the QSO target selection is non-linear with respect to imaging properties due to significant stellar contamination. We account for this contamination by including the Milky Way and Sagittarius Stream stellar distributions as additional inputs to our mitigation procedure.

This paper is organized as follows. In Section 2, we outline the DESI DR9 Legacy Imaging Surveys and introduce the strategy used QSO target selection. The observational features we study and the method used to mitigate systematics are detailed in Section 3. In Section 4, we show the results of our systematic mitigation and discuss how density fluctuations in the QSO target sample are related to observational features. In Section 5, we analyse the angular clustering properties of the DESI QSO target sample, illustrating the efficiency of our methods. Finally, our conclusions are presented in Section 6.

All magnitudes in this paper will be quoted on the AB system, including magnitudes from the Wide-field Infrared Survey Explorer (WISE) that are often given on the Vega system. In addition, except when mentioned otherwise, all computations with HEALPIX pixels are done with  $N_{\text{side}} = 256$  (a pixel area of  $\sim 0.05$  deg<sup>2</sup>) and all maps are plotted in a Mollweide projection with a HEALPIX resolution of  $N_{\text{side}} = 64$ .

## 2 QUASAR TARGETS IN LEGACY IMAGING SURVEYS

In this section, we introduce the Legacy Imaging Surveys that are used to perform the DESI QSO target selection. We then present the QSO target selection method itself, and finally discuss areas of the DESI footprint that have particularly high target density.

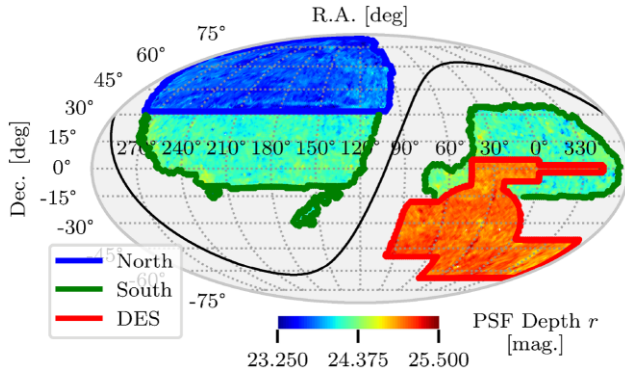
### 2.1 Data Release 9 of Legacy Imaging Surveys

To select targets for the DESI spectroscopic survey, the Legacy Imaging Surveys<sup>3</sup> programme was conducted over more than 14 000 deg<sup>2</sup> of sky from the Northern hemisphere, in three optical bands:  $g$  (470 nm),  $r$  (623 nm), and  $z$  (913 nm). The optical surveys were complemented by two infrared bands from the custom ‘unWISE’ coadds (e.g. Meisner, Lang & Schlegel 2017) of the all-sky data of the WISE satellite (Wright et al. 2010), namely: W1 (3.4  $\mu\text{m}$ ) and W2 (4.6  $\mu\text{m}$ ). A full description of the Legacy Imaging Surveys is

<sup>3</sup><https://www.legacysurvey.org/dr9/>

**Table 1.** Median values of the PSF depth and PSF size (cf. Section 3.1.1) for the three imaging surveys that together constitute the DR9 Legacy Imaging Surveys. DECaLS is split according to the DES region since the quality of the photometry inside this region is significantly better.

	PSF depth (mag)			PSF size (arcsec)		
	<i>g</i>	<i>r</i>	<i>z</i>	<i>g</i>	<i>r</i>	<i>z</i>
DECaLS (non-DES)	24.7	24.2	23.3	1.51	1.38	1.31
DES	25.2	25.0	23.8	1.42	1.24	1.14
BASS	24.2	23.7		1.89	1.67	
MzLS			23.3			1.24



**Figure 1.** Distribution of the PSF depth  $r$  (cf. Section 3.1.1) in the DR9 Legacy Imaging Surveys footprint. The  $r$  band is used to define the magnitude limit for DESI QSO target selection. The solid black line shows the Galactic plane. Three different imaging footprints are highlighted. The blue region is the combination of BASS and MzLS (designated *North* in this paper). The red region is the DES part of DECaLS (designated *DES*). The green region, which excludes the red and the blue regions, is the non-DES part of DECaLS (designated *South*).

available in Dey et al. (2019). The optical bands were collected via three independent programmes.

(i) The Beijing–Arizona Sky Survey (BASS; Zou et al. 2019) observed  $\sim 5100 \text{ deg}^2$  of the North Galactic Cap (NGC) in  $g$  and  $r$  using the 2.3-m Bok Telescope. The area surveyed corresponded to approximately  $\text{Dec.} > 32^{\circ}37'$ .

(ii) The Mayall  $z$ -band Legacy Survey (MzLS; Silva et al. 2016) provided  $z$ -band observations over the same footprint as BASS using the 4-m Mayall Telescope. Because the median value of the point spread function (PSF) size is significantly better than in the BASS data, the MzLS data are critical for deblending sources and deriving source morphology.

(iii) The Dark Energy Camera Legacy Survey (DECaLS) was performed with the Dark Energy Camera (DECam; Flaugher et al. 2015) on the 4-m Blanco Telescope. DECaLS observed the bulk of the Legacy Imaging Surveys footprint in  $g$ ,  $r$ , and  $z$ . DECam was initially built to conduct the Dark Energy Survey (DES) and DECaLS expanded the DES area using publicly available DECam time. However, DES imaging is significantly deeper than standard DECaLS observations as it is covered by more exposures (more than four in each band).

The median values of the PSF depth and PSF size that quantify the quality of the photometry, as explained in Section 3.1.1, are given in Table 1 for each program. Fig. 1 shows the PSF depth  $r$  in the Legacy Imaging Surveys and highlights three distinct regions.

(i) In blue, the combination of BASS and MzLS covering the northern part ( $\sim 5100 \text{ deg}^2$ ) of the DESI footprint (designated *North* hereafter).

(ii) In red, the DES part of DECaLS covering  $\sim 4600 \text{ deg}^2$  (designated *DES*).

(iii) In green, the non-DES part of DECaLS covering  $\sim 9900 \text{ deg}^2$  (designated *South*).

The region around the Large Magellanic Cloud (RA, Dec. in the ranges  $[52^{\circ}, 100^{\circ}]$  and  $[-70^{\circ}, -50^{\circ}]$ , respectively) is excluded from our study as it is heavily contaminated by stars.

## 2.2 QSO target selection

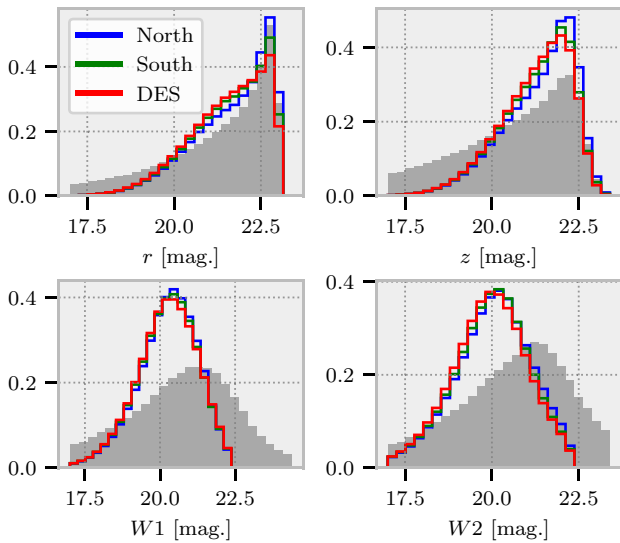
The QSO selection used in this study is an updated version of the preliminary selection described in Yèche et al. (2020). It follows the same general principles but is now based upon Legacy Surveys DR9 instead of DR8. We briefly summarize below the key properties of the selection.

QSOs are galaxies where a central supermassive black hole accretes large amounts of relativistic gas. The accreting region usually outshines the host galaxy by a large factor while being too small to be resolved, making these distant objects appear to us as point sources. The QSO selection therefore focuses solely on non-extended sources (i.e. sources with point source morphology) in the DR9 catalogue to avoid an almost 10-fold contamination by galaxies. The QSO selection then uses colours to distinguish the QSOs from other point sources (i.e. typically stars). We restrict the selection to magnitudes in the range  $16.5 < r < 23.0$  to reduce the stellar contamination at the bright end and limit to sources with reliable colours at the faint end. In addition, cuts on the *WISE* bands ( $W1 < 22.3$  and  $W2 < 22.3$ ) are used to further limit contamination by stars. We only consider sources that are not near bright stars, galaxies or globular clusters, i.e. excluding objects with MASKBITS 1, 12, or 13 set in the Legacy Surveys catalogues.<sup>4</sup>

In the absence of  $u$  band, we cannot apply a selection that relies upon the ultraviolet (UV) excess of quasars (e.g. Myers et al. 2015). In DESI, the main discrimination between quasars and stars is therefore obtained from the fact that quasars are about 2 mag brighter in the near-infrared at all redshifts than stars of comparable magnitudes and colours in the optical. The selection method uses a RF classification taking as input the  $r$ -band magnitude and all colours from two-band combinations of  $grzW1W2$ . The selection then identifies QSO candidates as sources that have a probability exceeding an  $r$ -dependent threshold. To take into account the differences in the photometry, the probability thresholds are tuned independently over the three imaging footprints in Fig. 1 to produce a median density of 300 targets per  $\text{deg}^2$  in weakly stars contaminated regions. For the North (respectively South), the median is computed in box of RA, Dec.  $\in [120^{\circ}, 240^{\circ}] \times [32^{\circ}3', 40^{\circ}]$  (respectively  $[120^{\circ}, 240^{\circ}] \times [24^{\circ}, 32^{\circ}2']$ ), and over all of the DES footprint. Since DES is less contaminated by stars than the other regions, the target density is on average smaller than 300 targets per  $\text{deg}^2$ . The median in DES is 285 targets per  $\text{deg}^2$ . The heat map of the resulting selection is illustrated in Fig. 3.

Fig. 2 shows the magnitude distribution of the QSO targets in the  $r$ ,  $z$ ,  $W1$ , and  $W2$  bands and demonstrates that these distributions are similar for each of the imaging footprints highlighted in Fig. 1. The three magnitude limits imposed on the selection are clearly visible on

<sup>4</sup><https://www.legacysurvey.org/dr9/bitmasks/>



**Figure 2.** Magnitude distribution of the QSO targets in the  $r$ ,  $z$ ,  $W1$ , and  $W2$  bands. These distributions are shown for the three independent imaging footprints in Fig. 1. Blue is for the North, green for the South, and red for DES. Each grey histogram depicts the magnitude distribution for the parent sample of sources (PSF sources with  $r < 23$ ) from which QSO targets are selected.

the corresponding histograms. Note that the  $r < 23.0$  limit also affects the  $z$ -band distribution, producing a sharp drop-off at the faint end for objects in the top right-hand panel. The two Gaussian distributions for  $W1$  and  $W2$  (bottom panels) demonstrate that the selection of a QSO target is not limited by the depth of the optical imaging (DECaLS or BASS/MzLS), but is sensitive to the determination of the fluxes in the *WISE* imaging.

To help contextualize which bands guide the DESI QSO target selection, it is worth noting that the colours that carry the largest weight in the selection are first  $z - W2$  and  $z - W1$  and then  $g - r$ ,  $W1 - W2$ , and  $g - z$ .

### 2.3 Overdensity and contamination

Fig. 3 exhibits several regions with higher density of QSO targets than average.

(i) The overdensity near the Galactic plane: the stellar density is higher near the Galactic plane (the black line in Fig. 3), which increases the stellar contamination in the QSO target selection. This effect is not obvious in the region bounded by  $RA, Dec. \in [120^\circ, 140^\circ] \times [-10^\circ, 15^\circ]$ , because the lower  $W1/W2$  PSF depth counters the excess of targets caused by the higher stellar density.

(ii) The overdensity in the Sagittarius Stream (see also Section 3.1.2): the stellar population of the Sagittarius Stream (which tracks the blue line in Fig. 3) is different from the Galactic stellar population. Most of the stars in the stream are bluer than Galactic stars and tend to have similar colours to the bulk of QSOs. Since there is little available data at magnitudes as faint as the stream, these stars cannot be efficiently rejected and end up contaminating our selected sample of QSO candidates.

(iii) The overdensity in the North: the DESI QSO target density increases with declination. This could be due to the poorer PSF depth in the  $z$  band in this region. This is likely caused by imaging depth decreasing at higher declination due to increasing airmass that was not compensated for by additional exposure time in the MzLS

observing strategy. Since the  $z$  band plays a crucial role in the QSO selection, the discriminating power between stars and DESI QSO targets is reduced at higher declinations.

(iv) The DES footprint is the least contaminated region, as expected from the high quality of the photometry in this region. Note that DES region is even so contaminated in the region of the crossing of the Sagittarius Stream.

## 3 METHODOLOGY

All the density and feature maps discussed in this section will be pixelized using HEALPIX<sup>5</sup> (Gorski et al. 2005) with  $N_{\text{side}} = 256$ . All HEALPIX operations are done using the HEALPY<sup>6</sup> package (Zonca et al. 2019). The choice of  $N_{\text{side}}$  (i.e. the size of the pixelization) is justified in Section 5.2.3.

### 3.1 Features

The aim of the systematic mitigation is to correct for spurious density fluctuations in the target selection without suppressing the cosmological signal contained in the targets' clustering. We restrict our study to 'features' directly linked to the observational properties such as imaging quality and physical properties that could alter the observations as the stellar density. We are careful to avoid incorporating parameters that relate to specific positions in the sky. For instance, we do not use the Modified Julian Date (MJD) as a feature (as in e.g., Rezaie et al. 2020), since the date of an observation directly translate into the position observed on that date.

#### 3.1.1 Observational features from DR9

We use 10 observational features to describe the systematic effects in the DESI QSO target selection. Whether our features are sufficient to describe spurious density fluctuations will be verified after the mitigation procedure by checking the isotropy of the corrected target selection map (see Section 4.3). The impact of mitigating systematics on the observational features themselves will be presented in Section 4.1.

The feature maps are generated using the script `bin/make_imaging_weight_map` from the DESITARGET package.<sup>7</sup> The maps, which we describe below, are extracted from the random catalogues provided as part of DR9 of the Legacy Imaging Surveys,<sup>8</sup> with the exception of the stellar density map.

(i) Stellar density (stardens) ( $\text{deg}^{-2}$ ): density of point sources from *Gaia* Data Release 2 (DR2; Gaia Collaboration 2018) in the magnitude range:  $12 < \text{PHOT\_G\_MEAN\_MAG} < 17$ .

(ii)  $E(B - V)$  (mag): Galactic extinction from Schlegel, Finkbeiner & Davis (1998) as modified by Schlafly & Finkbeiner (2011).

(iii) PSF depth ( $1/\text{nanomag}^2$ ) in  $r$ ,  $g$ ,  $z$ ,  $W1$ , and  $W2$ : PSF depth is the  $5\sigma$  point-source magnitude depth.<sup>9</sup> The dependence of target selection on PSF depth is governed by two competing effects. On the one hand, the number of resolved objects increases with PSF

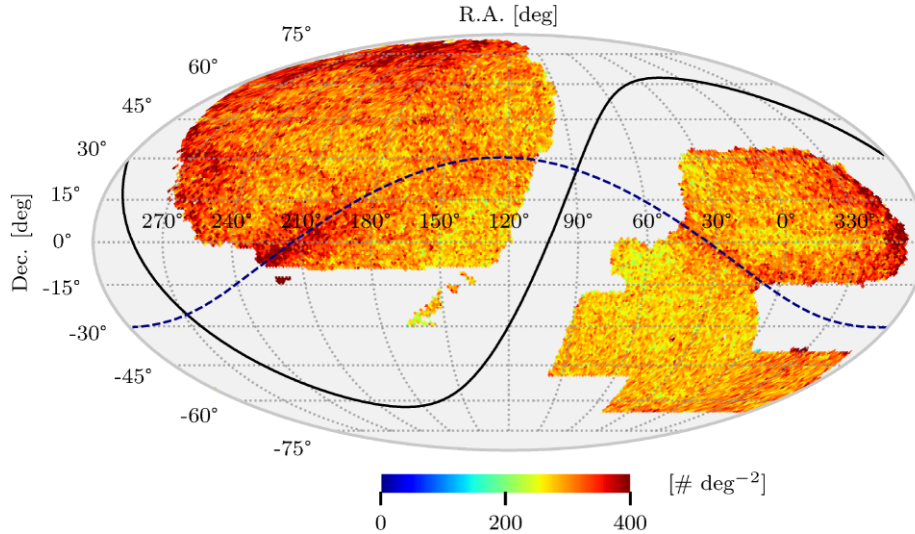
<sup>5</sup><http://healpix.sf.net>

<sup>6</sup><https://healpy.readthedocs.io/en/latest/>

<sup>7</sup><https://github.com/desihub/desitarget>

<sup>8</sup><https://www.legacysurvey.org/dr9/files>

<sup>9</sup>For a  $5\sigma$  point source detection limit in band  $x$ ,  $5/\sqrt{x}$  gives the PSF depth as flux in nanomaggies and  $-2.5(\log_{10}(5/\sqrt{x}) - 9)$  gives the corresponding magnitude (see <https://www.legacysurvey.org/dr9/catalogs/>).



**Figure 3.** Density map of the DR9 QSO target selection. The solid black line indicates the Galactic plane and the blue dashed line indicates the plane of the Sagittarius Stream.

depth. On the other hand, the determination of the flux is better for brighter objects, which means that contaminants are more easily rejected, resulting in a reduced number of targets. In this study, the  $z$  depth does not limit the target selection (cf. Section 2.2): it only affects the measurement of the  $z$  flux. Therefore, the target density decreases with increased  $z$  PSF depth. In contrast, fluxes in  $W1$  and  $W2$  are obtained with forced photometry. This allows fluxes to be measured for fainter objects that are detected at marginal significance in the *WISE* imaging (see Dey et al. 2019). Because such fluxes are noisy, the resulting colours scatter in regions of insufficient depth. Corresponding objects are rejected by the target selection not because they are not quasars but because their colours are not in the region populated by QSOs. Conversely, when the PSF depth in  $W1$  or  $W2$  is higher, the colours are more precise, and a greater number of targets are selected as genuine QSOs.

(iv) PSF size (arcsec) in  $r$ ,  $g$ , and  $z$ : inverse-noise-weighted average of the full width at half-maximum of the PSF, also called the delivered image quality. A small value of PSF size corresponds to a good image resolution, which leads to more precise fluxes and improved target selection.

Fig. 4 shows five of the 10 observational feature maps from DR9. It is important to note that these 10 features are correlated, and the level of correlation depends on the imaging footprint (North, South, DES). Stellar density and  $E(B - V)$  are positively correlated across the entire sky but  $E(B - V)$  contains additional information. This can be seen by comparing the top two left-hand panels in Fig. 4 in the region near  $(RA, Dec.) = (0^\circ, 15^\circ)$ . The  $W1$  and  $W2$  PSF depths are highly correlated in all of the three imaging footprints of interest. However, we include both as our analysis will show that the  $W2$  PSF depth is particularly heavily weighted by the RF in the North, indicating that this feature has extra information that is not contained in the  $W1$  PSF depth. For the optical observational features, two cases are of particular note. In the North, the  $g$  and  $r$  bands were collected by the same camera and so their observational features are positively correlated, but these features are not correlated with the  $z$  band, which was independently obtained by MzLS. In the South and in DES, the three bands were collected with the same camera and therefore are all positively correlated. Finally, the  $W1/W2$  PSF depths are more correlated with the other features in the DES region,

as compared to the North and the South, since all the feature maps are more uniform in the DES footprint.

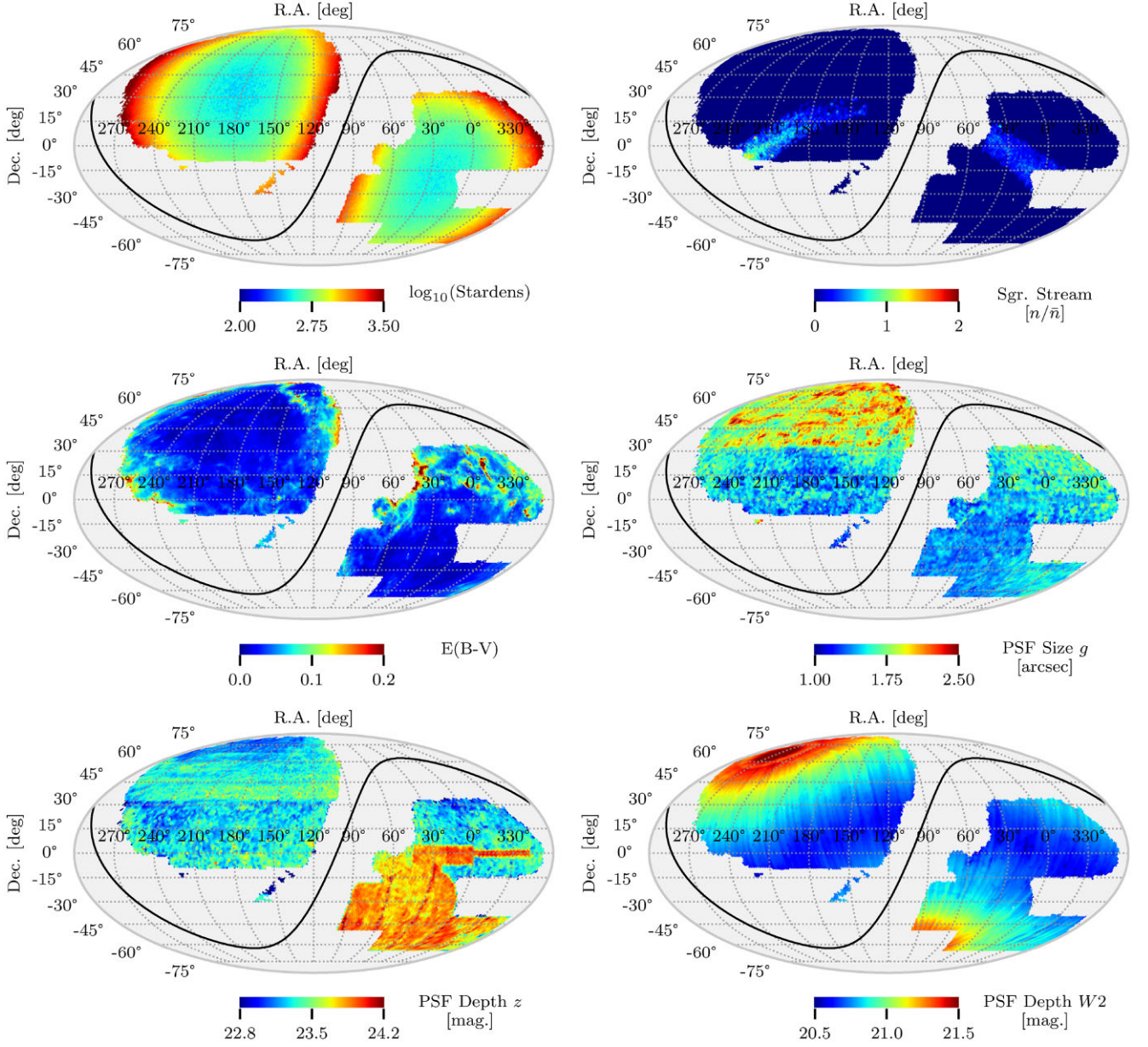
### 3.1.2 Sagittarius stream model

The target density map (Fig. 3) shows a significant excess in the Sagittarius Stream region. The Sagittarius galaxy is one of the closest galaxies orbiting around the Milky Way. The gravitational forces create two tidal arms called streams wrapping the Milky Way with the same orbit (Newberg et al. 2002; Majewski et al. 2003).

The Sagittarius contamination (see the blue dashed line in Fig. 3) occurs mainly in the South footprint but also touches the DES footprint. However, none of the feature maps discussed so far contain a pattern matching this contamination. Thus, an additional feature is required in our analysis. To model this feature, we use the Sagittarius Stream catalogue derived in Antoja et al. (2020). This catalogue is built from the *Gaia* DR2 catalogue, identifying stars in the Stream via their proper motions. Matching the Antoja et al. catalogue to the SDSS DR16 QSO catalogue (Lyke et al. 2020) on position, we find that some of the stars are actually known QSOs. To generate the Stream feature shown in Fig. 4 (top right-hand panel), we remove the known QSOs and apply an  $r > 18$  cut to only select faint stars. A fainter cut would better match the QSO selection, but the Antoja et al. catalogue does not contain a sufficient number of objects faint in  $r$ -mag to apply a fainter cut.

### 3.1.3 Three imaging footprints

The three footprints of the Legacy Imaging Surveys (North, South, and DES; as defined in Section 2.1) exhibit distinct imaging properties. As shown in Fig. 5, while the PSF size in the  $z$  band is similar in the North and South, the  $r$ -band depth distributions are very different in the North, South, and DES footprints. For instance, a PSF depth of 24.7 in  $r$  leads to a small overall target overdensity in DES (cf. Fig. 9), while it corresponds to an underdensity in the South (cf. Fig. 8). The imaging properties in the South footprint are similar in the North (NGC) and South Galactic Cap (SGC). There is therefore no reason to split the South footprint in two.



**Figure 4.** Maps of the most important features used in our systematics mitigation. The difference between the three imaging footprints highlighted in Fig. 1 is clearly visible in the *PSF depth z* and *PSF size g* feature maps.

In addition, the selection threshold in the QSO target selection is set independently in each of these three footprints. We therefore model systematic effects in the three footprints independently.

### 3.2 Systematic model

We will identify each position on the sky with the corresponding pixel number  $i$  from a HEALPIX pixelization. For each imaging footprint, we only consider pixels that contain at least one object from the random catalogues. The density of selected targets inside a pixel  $i$  is noted  $N_i$ . It is derived from the observational target density as

$$N_i^{\text{obs}} = f_i N_i, \quad (1)$$

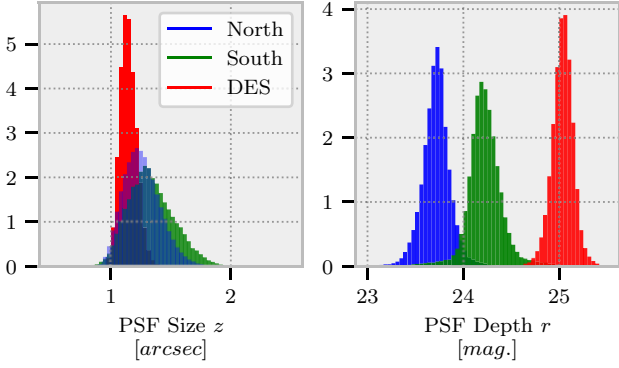
where  $f_i$  is the observed fractional area of pixel  $i$  calculated as the number of random targets inside the pixel divided by the nominal density of randoms, and  $N_i$  is the true number of quasars in the pixel

in the absence of any systematic effects.  $N_i$  is related to the mean target density  $N_0$  by

$$N_i = N_0 (1 + \delta_i), \quad (2)$$

where  $\delta_i$  is the overdensity that contains the cosmological information.

$N_0$  is unknown and has to be estimated from the target selection as the mean of the pixel density over the footprint weighted by  $f$ :  $\hat{N}_0 = \langle N_i \rangle_i = \langle f_i^{-1} N_i^{\text{obs}} \rangle_i$ . Since the target selection is contaminated by stars, we choose different regions known to be less contaminated to perform this estimation in the three footprints. We use the same region used to tune the probability selection during the quasar target selection, which was a box of RA, Dec.  $\in [120^\circ, 240^\circ] \times [32.2, 40^\circ]$  for the North footprint, a box of RA, Dec.  $\in [120^\circ, 240^\circ] \times [24^\circ, 32.3]$  for the South footprint, and the entire footprint for DES. Note



**Figure 5.** Density distribution of the PSF size  $z$  and PSF depth  $r$  for the three imaging footprints shown in Fig. 1. These two histograms show that a joint analysis on the three footprints is difficult since some features share the same support and others do not.

that even these regions are contaminated by stars, such that the actual QSO target density is smaller.

Systematic effects will be taken into account using an additional term  $F_i$  such that

$$N_i = N_0 (1 + \delta_i) F_i. \quad (3)$$

The aim of the imaging systematics mitigation is to describe  $F_i$  as a function of a set of observational features. These features must not depend on the sky position, to avoid suppressing the cosmological signal. We assume that the imaging systematics can be completely explained by our set of observational features. This assumption can be validated by the uniformity of the target selection density map after mitigation (cf. Section 4.3).

The features in the pixel  $i$  are denoted  $s_i$  that is an  $n$ -dimensional vector where  $n$  is the number of features ( $n = 11$  in our case: 10 observational features from Section 3.1.1 and the Sagittarius Stream feature from Section 3.1.2).  $F$  should depend only on these features and not on the pixel number. We thus rewrite  $F_i$  as  $F(s_i)$ , and  $N$  now also depends on the observational features:

$$N_i(s_i) = N_0 (1 + \delta_i) F(s_i). \quad (4)$$

We denote  $A(S)$  to be the subset of pixel numbers that belongs to the same region  $S$  in the space of the feature maps. Averaging over many pixels will suppress the density contrast:  $\langle \delta_i \rangle_{i \in A(S)} = 0$ . Thus, if  $S$  is sufficiently large, the contamination signal is given by

$$F(S) = \frac{\langle N_i(S) \rangle_{i \in A(S)}}{N_0}. \quad (5)$$

This averaging is controlled by hyperparameters of the different regression methods. For instance in the particular case of the RF (see Section 3.3.3), the averaging is controlled by the minimum number of objects in a leaf.

The systematic correction will be modelled by a weight to apply in each pixel, defined as

$$w_i^{\text{sys}} = \frac{1}{F(s_i)}. \quad (6)$$

The correction is only efficient on scales at least as large as the typical size of the pixel at  $N_{\text{side}} = 256$ , i.e. about  $\theta = 0''.22$ . Hence, the correction is constant within each pixel.

The regression is performed using only reliable pixels, which we choose to be pixels with  $f_i > 0.9$ , and is then applied to all pixels. This criterion removes pixels that contain too few targets, which

could bias the regression. The pixels that are excluded represent only about 3.8 per cent of the DR9 footprint, lying mainly at the edges of the footprint and in the region south of Dec.  $< -10^\circ$  in the NGC. For studies with a smaller pixel size, e.g.  $N_{\text{side}} = 512$ , the nominal density of the randoms would have to be increased to limit the Poisson noise when determining  $f_i$ .

### 3.3 Regression methods

We test different methods, utilizing the same feature set, to perform the regression presented above. The initial correction, obtained with a linear regression, turns out to be insufficient (as illustrated in Figs 7–9), necessitating non-linear regression approaches. We therefore test two classical machine learning methods based on SCIKIT-LEARN (Pedregosa et al. 2011), namely, the RF and NN methods.

#### 3.3.1 Linear

The contamination function  $F$  is described as a linear function of the observational features:  $F(s_i) = a_0 + \sum_{j=1}^{11} a_j s_{ij}$ . The coefficients  $a_j$  will be estimated with a least-square minimization using the IMINUIT package (Dembinski et al. 2020). The  $\chi^2$  used for the minimization is defined as

$$\chi^2 = \sum_i \frac{1}{N_i} \left( F(s_i) - \frac{N_i}{N_0} \right)^2 + c_{\text{reg}} (\langle F(s_i) \rangle_i - 1)^2, \quad (7)$$

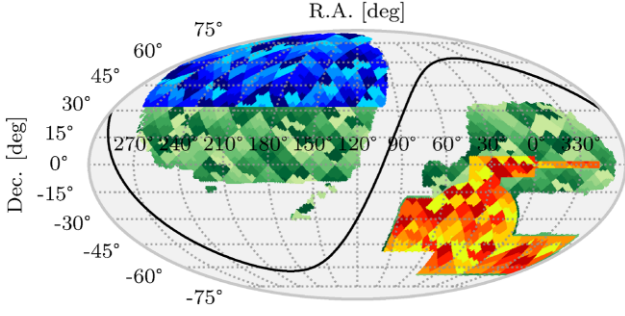
where  $\sqrt{N_i}$  is an estimate of the error for the object count inside a pixel and  $c_{\text{reg}}$  is a penalty term to regularize the regression. Since the distribution of  $N_i/N_0$  is not symmetric around 1, the higher number of pixels with  $N_i/N_0 < 1$  forces the contamination function to not be centred around 1. We therefore use a penalty term to flatten the density map around the chosen mean density. The value of the penalty term depends on the number of pixels used to build  $\chi^2$ . In our configuration, we use  $c_{\text{reg}} = 2 \times 10^6$  and we check that as long as  $c_{\text{reg}}$  is sufficiently large, its value does not change the result of the regression.

#### 3.3.2 K-fold training

Machine learning methods tend to overfit the data and have to be used carefully. Since we cannot create a training sample independent of the data set, we have to use a  $K$ -fold training to avoid overtraining.

A  $K$ -fold training is a method that splits the data into  $K$  folds. For each fold, the method is trained with the remaining  $K - 1$  folds then the regression is performed on the isolated fold. This method guarantees that no data used for the regression are used for the training. A small value of  $K$  ensures no overfitting but the mitigation will be inefficient since the training set would not contain enough information. A high value of  $K$  ensures an efficient mitigation but is more prone to overfitting. In our case, we choose to work with 6 folds in the North and the DES footprint and with 12 folds in the South. Each fold has an area of about  $830 \text{ deg}^2$ .

The locations of the folds have to be carefully chosen, since the contamination has distinct causes. For example, if all the borders of the Galactic plane in the South footprint were to be put in the same fold, the machine learning algorithm would not be able to explain related contamination with the  $K - 1$  remaining folds, since all the relevant information would have been removed from the training set. Therefore, we construct folds from small patches of the sky to spread the information across all folds. Such folds can be constructed easily



**Figure 6.** Distribution of folds across the three imaging footprints. There are six folds in the North (blue region), 12 in the South (green region), and six in DES (red region). Folds were split into small patches so the specific effects that contaminate target selection are always spread across several folds. The area of each patch is  $\sim 52 \text{ deg}^2$ .

in the HEALPIX ‘nested’ scheme using the SCIKIT-LEARN function `GroupKFold`. The size of the patch matters, as patches that are too small lead to overfitting since all the information is present in all the folds. We use patches (groups in SCIKIT-LEARN language)

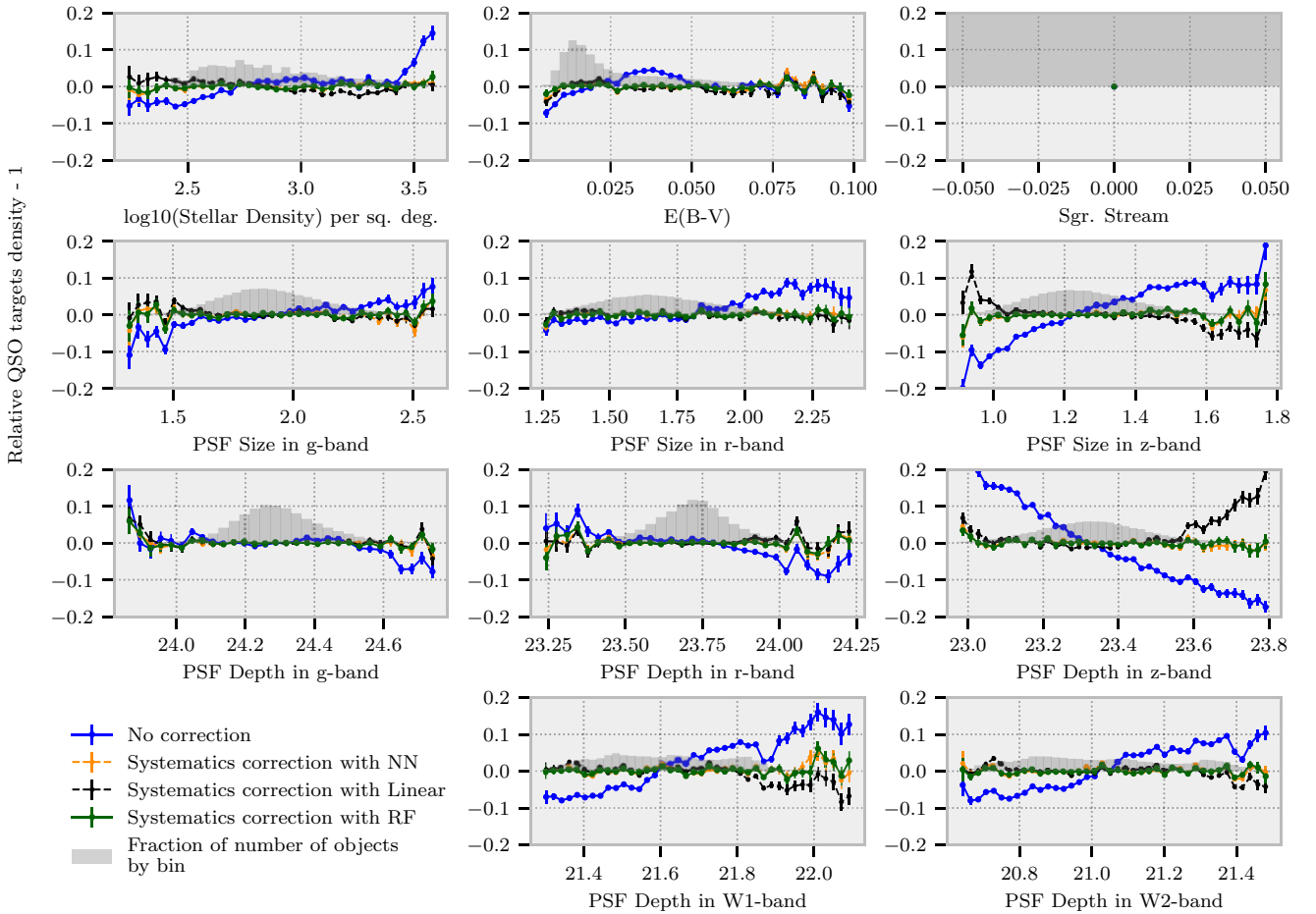
of 1000 pixels, which correspond to an area of  $\sim 52 \text{ deg}^2$  for each patch.

The patch distribution is shown in Fig. 6, which demonstrates that no particular region is described by a single colour. We checked that the estimated weight remains stable if we slightly vary the number of folds or the size of the patch, and used mocks (cf. Appendix A) to ensure that no overfitting occurs.

### 3.3.3 Random forest

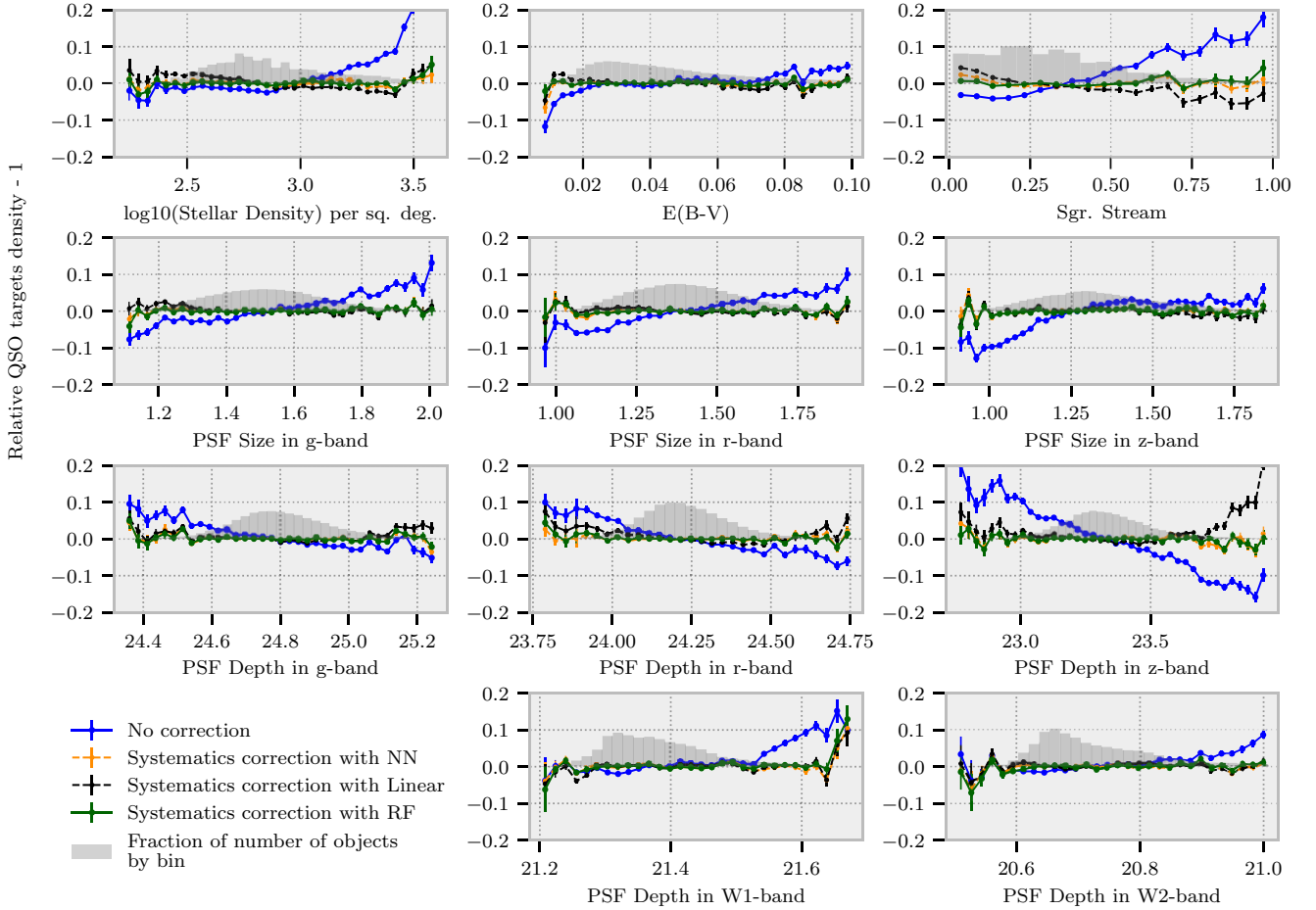
The first machine learning method used in our regression analysis is the well-known random forest (RF) algorithm. It is easy to parametrize and gives a helpful classification of the features as a function of their importance during the regression.

For the regression, we use the same set of hyperparameters for each footprint and we do not normalize the data set. We choose a forest of 200 trees and we fix the minimum number of samples contained inside a leaf at 20. This means the average to estimate  $F(S)$  (cf. equation 5) is computed with at least 20 pixels. The mean number of pixels in each leaf is 80. We checked that the minimum sample size in a leaf has no strong impact on the regression during the  $K$ -fold training.



**Figure 7.** Relative QSO target density in the North footprint as a function of each observational feature. The relative QSO target density is a mean value after rejecting outliers, i.e. pixels with a coverage lower than 90 per cent. The blue lines depict the raw DESI QSO target selection. The green (respectively yellow/black) lines depict the QSO target selection after correcting for systematic effects using the RF (respectively NN/linear) regression. The histogram represents the fraction of objects in each bin for each observational feature, and the error bars are the estimated standard deviation of the normalized target density in each bin. The three methods all successfully flatten the relative QSO target density as a function of each observational feature. However, the linear method is less efficient than the other two methods.





**Figure 8.** Same as Fig. 7, but for the South imaging footprint.

### 3.3.4 Neural network

We also use the multilayer perceptron (MLP) algorithm, a fully connected type of neural network (NN). Finding the best hyperparameters for an NN is difficult. We base our choice on the result of a grid search performed on simulated QSO samples (i.e. mocks). The mocks used for this study are described in Appendix A.

For the regression, we use the same set of hyperparameters for each footprint, but when dedicated mocks for DESI become available, different hyperparametrizations will be possible. We use an MLP with two hidden layers, comprising 10 neurons for the first layer and eight for the second (i.e. a (10, 8) formalism). The data are normalized on each fold dividing each feature by the standard deviation estimated in the considered fold. We hyperparametrize the MLP with a sigmoid activation function, a batch size of 1000, and a tolerance of  $1 \times 10^{-5}$ . We use the Adam solver (Kingma & Ba 2014) to perform the minimization during the training.

## 4 SYSTEMATICS MITIGATION

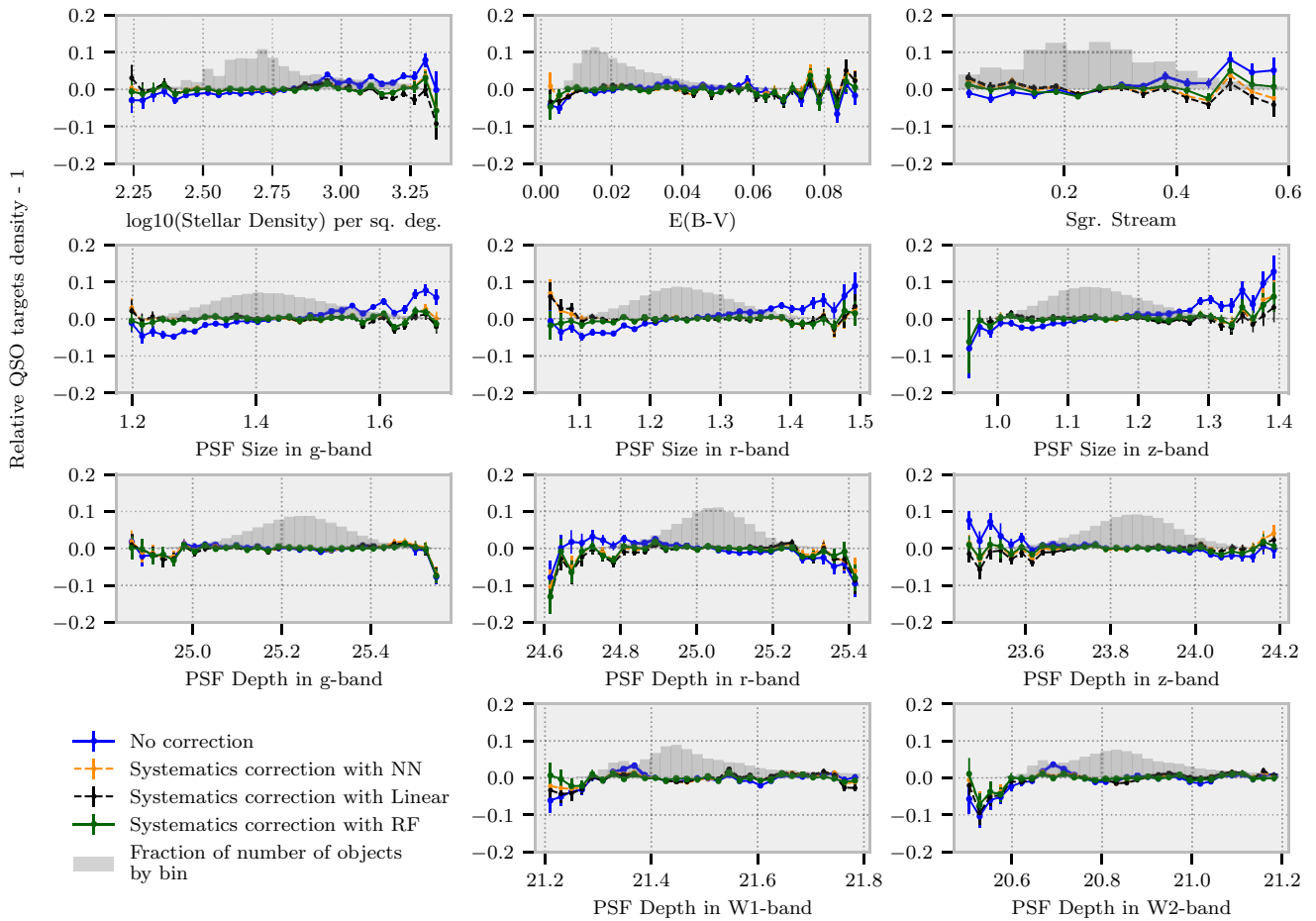
We apply the method presented in Section 3 to correct for observational systematics in the DESI QSO target selection. We first describe the origin of these systematics and explain the role of the most important features. We then present the target density after it has been corrected by our systematic mitigation method.

### 4.1 Systematic plots

To illustrate our method, we plot the relative QSO target density as a function of each observational feature. We will refer to these plots as ‘systematic plots’. In our systematic plots, we centre the relative density around 0. The goal of the correction is to obtain a relative density that is independent of the value of the observational feature. We produce systematic plots for the North (see Fig. 7), for the South (see Fig. 8), and for the DES (see Fig. 9).

Below, we give a brief description of the systematic plots in each footprint.

(i) North (Fig. 7):  $z$  and  $W1/W2$  are crucial for the DESI QSO target selection. The plot of the relative density as a function of the PSF depth  $z$  or the PSF size  $z$  shows that when the  $z$  observational feature values are for bad observational conditions (small PSF depth and high PSF size), the discriminating power of the target selection algorithm is poorer and the relative density higher. In addition to the importance of the  $z$  band for the target selection, the  $z$  band benefits from better image quality (smaller PSF size and higher PSF depth) than  $g$  or  $r$ . The fluctuations of the relative density as a function of the  $g/r$  observational features are therefore weaker but they follow the same general pattern. As explained in Section 3.1.1, the impact of the  $W1/W2$  PSF depths differs from the  $g/r/z$  depths since the depth of the  $WISE$  colours is crucial for selecting QSOs. Hence, the number of targets increases with  $W1/W2$  PSF depth. In addition, the  $z$  PSF



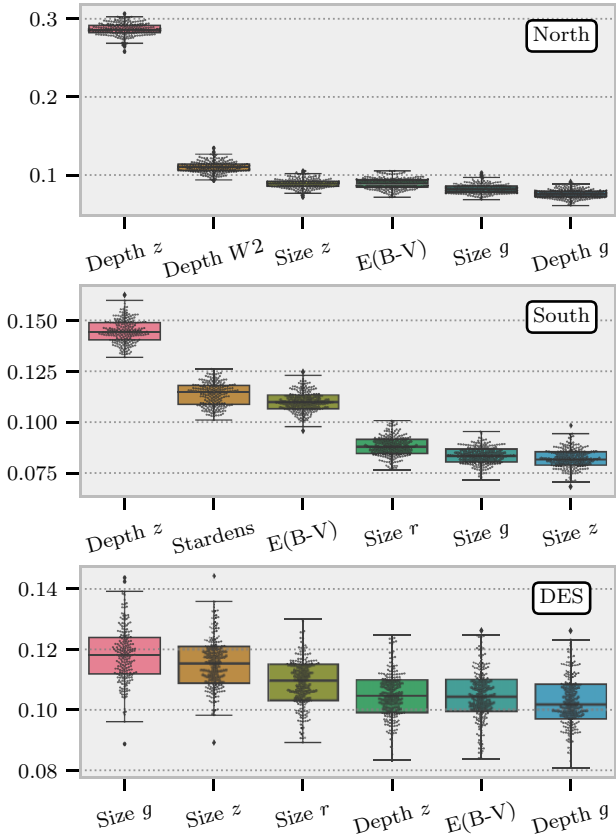
**Figure 9.** Same as Fig. 7, but for the DES imaging footprint.

depth is smaller in regions where the  $W1/W2$  PSF depth is larger, as shown in the bottom panels of Fig. 4. The combined effect increases the target density in this region.

(ii) South (Fig. 8): the plots as a function of the  $z$  and  $W1/W2$  PSF depths at high values exhibit similar behaviour as in the North. However, the effect is less significant since it is mainly visible in the RA, Dec.  $\in [210^\circ, 270^\circ] \times [15^\circ, 30^\circ]$  region. In comparison to the North, the plots as a function of the  $g/r/z$  features all show similar trends to each other since these bands were collected with the same camera. The  $z$ -band dependence is the strongest, as expected since it is one of the most important bands for the QSO selection. The plots as a function of the  $W1/W2$  PSF depth exhibit seemingly unexpected behaviour at low values of the PSF depth, where the relative density is almost flat. As explained in Section 2.3, the excess of stars near the Galactic plane or the Sagittarius Stream counteracts any expected decrease of targets due to lower values of the  $W1/W2$  PSF depth (cf., in particular, the RA, Dec.  $\in [120^\circ, 150^\circ] \times [-10^\circ, 15^\circ]$  region). The plots in the top panel of Fig. 8 (stellar density, dust, and Sagittarius Stream) indicate a higher relative density due to the presence of stars. These features explain the stellar contamination near the Galactic plane and inside the Sagittarius Stream. Some stars have similar colours to QSOs and are therefore selected as QSO targets. More stars thus enter the QSO selection in regions of higher stellar density.

(iii) DES (Fig. 9): the fluctuations of the relative density are much lower than in the two other footprints. The observational features for  $g/r/z$ , and especially for the  $z$  band, are better in DES than in both the South and the North (cf. the median value of these features in Table 1 and the distribution of the PSF depth size  $z$  and the PSF depth  $r$  in Fig. 5). This results in excellent discrimination of QSOs from stars in the target selection process. So, DES is the least contaminated region and exhibits smaller fluctuations in relative density as a function of the observational features.

In Figs 7–9, we also plot the three different regression methods we apply for each footprint: RF in green, NN in yellow, and linear in black. These plots show that the mitigation works well and correctly flattens each feature. The difference in efficiency between the linear and the non-linear methods is particularly obvious for the North and South footprints, where the contamination is stronger. For example, the linear correction fails when the relative density as a function of the PSF depth  $z$  is large, as shown in Fig. 7. This illustrates the non-linearity of the contamination and justifies the use of the machine learning methods. Both the RF and the NN perform well in correcting the non-linear systematics, although small differences can be found between these two methods. A more detailed comparison between the RF and NN will be done in Section 5.2.2.



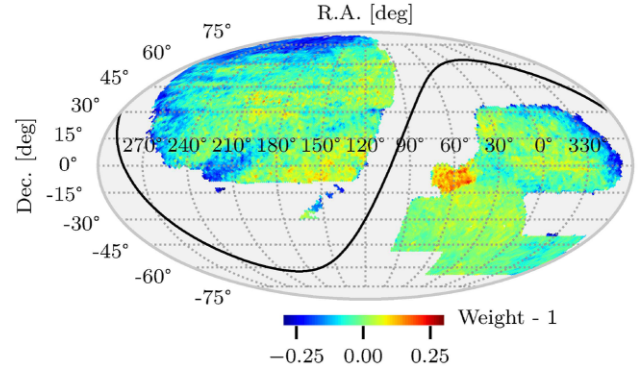
**Figure 10.** Feature importance calculated using the RF method. Only the six most important features are plotted for each footprint. Each dot represents the feature importance value found with one decision tree. The box plot is assembled across all the trees of the forest. The values of the importance cannot be compared between the three regions.

#### 4.2 Importance features

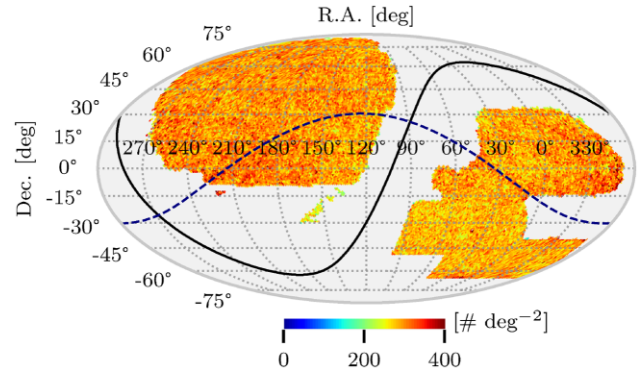
The RF algorithm includes a specific tool called *importance features*. Importance is a measure of which features most affect the regression. The importance features metric for the RF implemented in SCIKIT-LEARN is based on the mean decrease in impurity (MDI), which is also called the Gini importance. The metric corresponds to averaging (over all the trees of the forest) the impurity reduction of each node weighted by the ratio of the training set passing through each node. The *permutation importance* was also tested and yields similar result in our data set than the Gini importance.

A feature with a higher importance value is more discriminating than a feature with a lower importance value, so the importance values are useful to identify the observational features that lead to contamination, to first order. It is worth noting that a feature with a low importance value is not necessarily useless and can still improve the training compared to a case where we remove it. For instance, the Sagittarius Stream feature is necessary to correct the overdensity in the Sagittarius region but it is not a high-value importance feature since it is useful only for a small number of pixels that vary in a manner that is quite different from other pixels in the footprint. This well-known bias can be circumvented using the *permutation importance*, where the Sagittarius Stream feature is ranked as one of the most important in the South.

In Fig. 10, we plot the importance for each feature. We only plot the six most important features for each footprint, since the



**Figure 11.** Distribution of the systematic correction weights (weight  $- 1$ ) across the Legacy Imaging Surveys footprint.



**Figure 12.** Density map of the QSO target selection after mitigating systematics using the RF method. The density map is quasi-uniform compared to the initial density map shown in Fig. 3. The solid black line depicts the Galactic plane, and the blue dashed line depicts the plane of the Sagittarius Stream.

other features have about the same value as the sixth most important feature. We recover the expected most important features described in Section 4.1.

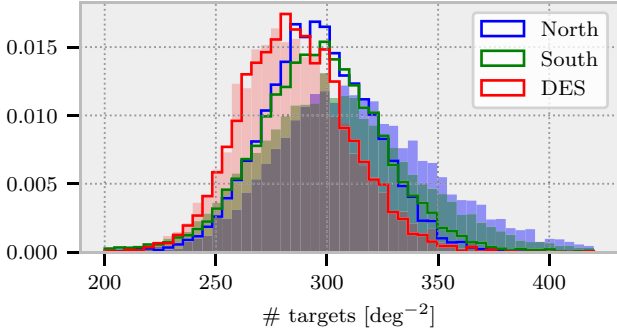
(i) North: the importance feature analysis makes it clear that the PSF depth  $z$  and the PSF depth  $W2$  play a key role in the contamination of the target selection in the North.

(ii) South: the South region is strongly contaminated by stars from the Galactic plane as highlighted by the importance of the stardens and  $E(B - V)$  features. The Sagittarius Stream feature has a low importance even if it is crucial to correct the Sagittarius Stream region (as explained above). The PSF depth  $z$  plays an important role in this overdensity, as already noted. Surprisingly, the  $W2$  PSF depth does not appear as an importance feature, even though it explains the underdensity near the anti-Galactic pole.

(iii) DES: no clear importance feature emerges, as expected given that the DES region is the least contaminated, most uniform, region.

#### 4.3 Quasar target selection after correction

We construct a pixel weight map to correct the over- or underdensities of the target selection. Fig. 11 shows the weight map obtained with the RF regression. We then multiply the density of the DESI QSO target selection by the weight on a pixel-by-pixel basis. Fig. 12 shows the corrected QSO selection. The map is almost completely



**Figure 13.** Density distribution of the QSO target selection for non-corrected (full histogram) and corrected (line) cases in the three photometric footprints. After the correction, the width of the histogram is smaller in each photometric footprint. As mentioned in Section 2.2, the density in DES is lower than in the North or in the South.

uniform at this resolution. The largest overdensities vanish after mitigation, which confirms that the set of observational features that we considered suffices to explain most of the observed large-scale target density variations.

The density distribution of the QSO target selection is shown in Fig. 13. The full histograms are before applying the systematic mitigation and the lines are for after. The systematic mitigation acts on the width of the histograms removing the over- or underdensity pixels. This effect is smaller in DES than in the North and in the South that confirms the visual inspection of the Figs 3 and 12. The difference in mean density between DES and the two other photometric footprints was mentioned in Section 2.2

## 5 ANGULAR CORRELATION AND CLUSTERING

Our systematic mitigation method is able to correct fluctuations in density as a function of our observational features. In this section, we measure the impact of the systematic mitigation on the angular correlation function.

### 5.1 Two-point correlation function

#### 5.1.1 Definition

The three-dimensional two-point correlation function  $\xi(\mathbf{r})$  describes the excess probability to find a pair of objects inside two infinitesimal volumes  $dV_1$  and  $dV_2$  separated by  $\mathbf{r}$ :

$$dP(\mathbf{r}) = \bar{n}^2 (1 + \xi(\mathbf{r})) dV_1 dV_2, \quad (8)$$

where  $\bar{n}$  is the mean density. It is related to the contrast density by  $\xi(\mathbf{r}) = \langle \delta(\mathbf{x})\delta(\mathbf{x} + \mathbf{r}) \rangle_x$ . The cosmological principle ensures that  $\xi$  depends only on  $r$ .

The same definition can be extended to the two-dimensional case, i.e. the angular correlation where volumes are replaced by solid angles and distances by angular distances:

$$dP(\theta) = n_0^2 (1 + \omega(\theta)) d\Omega_1 d\Omega_2, \quad (9)$$

where  $n_0$  is the mean angular density.

#### 5.1.2 Two-point correlation function estimator

We estimate the angular correlation function  $w(\theta)$  using the Landy–Szalay (LS) estimator derived in Landy & Szalay (1993):

$$\hat{w}(\theta) = \frac{aDD - bDR + RR}{RR}, \quad (10)$$

where DD, DR, and RR refer to the weighted pair counts data–data, data–random, and random–random with an angular separation  $\theta$ . The normalization terms are

$$a = \frac{\sum_{i \neq j} w_i^R w_j^R}{\sum_{i \neq j} w_i^D w_j^D} \quad \text{and} \quad b = \frac{\sum_{i \neq j} w_i^R w_j^R}{\sum_i w_i^D \sum_j w_j^R}, \quad (11)$$

where  $w_i^D$  (respectively  $w_i^R$ ) is the weight for the data (respectively random). The LS estimator is known to be unbiased and to have minimal variance in the limit of an infinitely large random catalogue with a volume greater than the scales considered, and for weak correlations ( $w(\theta) \ll 1$ ).

We use the package CUTE<sup>10</sup> (Alonso 2012) to perform the estimation. CUTE is a fast implementation written in C and using OpenMP and MPI.

### 5.2 Angular correlation of the DR9 quasar target selection

The systematic mitigation method is checked to avoid overfitting. This is done in Appendix A by applying the mitigation method to contaminated mocks. Then, we compute the angular correlation of the DR9 quasar target selection corrected by the systematic weights that we calculated in Section 4.

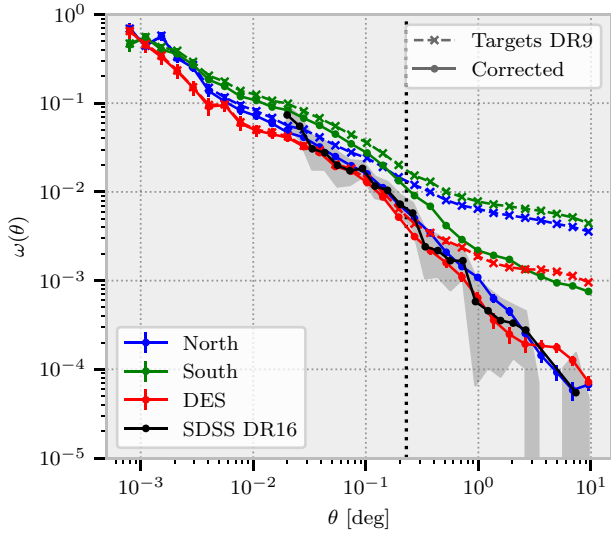
#### 5.2.1 Comparison with SDSS DR16

We calculate the angular correlation function of the raw and corrected (RF method) QSO target densities for the three different imaging footprints. For comparison, we also calculate the angular correlation function of SDSS DR16 quasars. Note that the correlation with SDSS quasars cannot be computed at angles smaller than 62 arcsec due to fibre collisions (e.g. Dawson et al. 2016). For SDSS DR16 quasars, we use the systematic weights provided in Rezaie et al. (2021) based on an NN treatment (see Ross et al. 2020, for the standard treatment). The results are shown in Fig. 14. Note that the error bars shown in Fig. 14 are the standard deviation of the LS estimator (Landy & Szalay 1993) except for the correlation function of SDSS quasars, for which the errors are estimated using the standard deviation across 100 EZ-mocks from Zhao et al. (2021).

After mitigating for systematics, the angular correlation in the North and in the DES region (again, see Fig. 1 for the definition of these regions) is comparable with the correlation computed with the SDSS DR16 sample. The SDSS DR16 sample has been carefully corrected for systematic effects (as derived in previous work), and so is expected to be largely free from any contamination. In the South, even after mitigating for systematics, we do not recover the same level of correlation as for the SDSS DR16 sample. Reasons for this difference are discussed further in Appendix B.

It is worth noting that the correction is larger at larger angles, i.e. the impact of systematic effects is higher on large scales. Therefore, mitigation of photometric systematics is critical for studies that require information from large scales, such as studies of primordial non-Gaussianity.

<sup>10</sup><https://github.com/damonge/CUTE>



**Figure 14.** The angular two-point correlation function of the DR9 QSO targets. The dashed lines represent the nominal DESI QSO target selection. The solid lines are for the corrected QSO target selection with RF method. The black line is the angular correlation function from SDSS DR16 (which cannot be calculated at angles smaller than 62 arcsec, due to fibre collisions). The grey region depicts the error for SDSS DR16 estimated as the standard deviation across 100 EZ-mocks. The black dotted line corresponds to the typical resolution of our regression analyses, i.e. the pixel size at  $N_{\text{side}} = 256$ . The solid green line is discussed in detail in Appendix B.

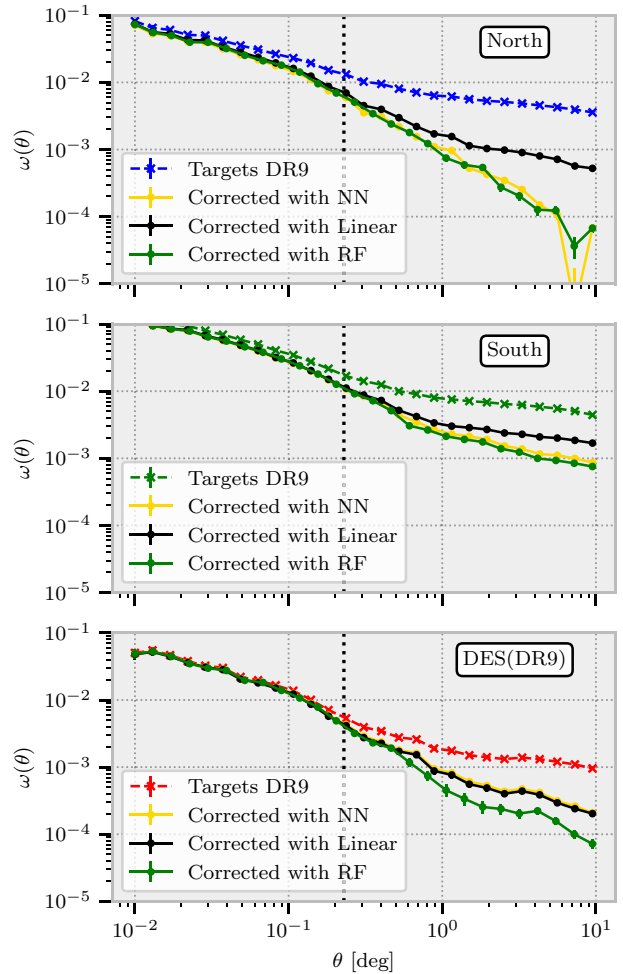
### 5.2.2 Regression method comparison

Fig. 15 shows the difference between the three methods introduced in Section 4 on the North, on the South, and on DES. As expected, the linear method is less effective than the other two in the highly contaminated North (top panel), highlighting the necessity of our more complex machine-learning-based regressions. The two machine-learning-based methods give similar results in the North. None of the three methods properly correct for the contamination in the South (middle panel); this is discussed further in Appendix B.

The information in Fig. 9 suggests that all three methods introduced in Section 4 are quite effective at mitigating systematics in the DES region. However, the bottom panel of Fig. 15 demonstrates that the linear method and the NN method less effectively correct the angular correlation function in the DES region as compared to the RF method. The difference between the linear and the RF methods comes from the non-linear part of the contamination in this region. The systematics plots show that in the highly contaminated regions, the linear method corrects less than the RF.

The difference between the NN and the RF is more subtle since the training information, i.e. the area and the chosen folds in that area, is similar. The explanation comes from the difference between the two algorithms. The NN method is less efficient to correct small, highly contaminated regions. The RF creates boxes in feature space and separates the most contaminated regions from the rest. The estimation of the correction weight is then possible everywhere.

To solve this problem, a regularization term can be introduced to force the NN to also consider small regions. The choice of the regularization value strongly depends on the size of the small regions and on their location in feature space. Without realistic mocks for the DESI QSO sample, this additional hyperparameter cannot be easily optimized. The training time for the NN also varies considerably as a function of the value of the tolerance chosen to sample the NN



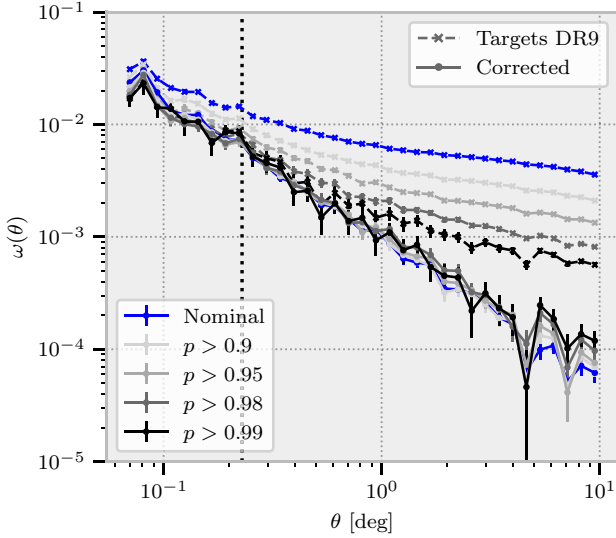
**Figure 15.** The angular correlation function for the three different methods introduced in Section 4 for the three regions highlighted in Fig. 1. The linear method is less effective than the two machine-learning-based methods, especially in the most strongly contaminated region (the North). The correction in the South does not improve properly the correlation for any of the three methods. The RF and NN methods produce similar results in the North but slightly different results in DES. This is because the RF is more robust to sampling high levels of contamination from a small number of pixels.

hyperparameters, whereas, with our parametrization, the RF method is quicker to train. Since the RF correction already obtains good results in DES, we do not need to improve the current NN method. We leave fine optimizing of the NN method for future work when realistic DESI QSO mocks become available.

### 5.2.3 Resolution of the correction

The size of the pixels used when determining weights to mitigate systematics is critical, because it gives the scale at which the correction is most effective. We perform all the analysis above with  $N_{\text{side}} = 256$  corresponding to a characteristic angle of  $\sim 0.22$ , i.e.  $12.6 h^{-1} \text{ Mpc}$  at  $z \sim 1.7$  (Planck 2015; Planck Collaboration XIII 2016). This pixel size is chosen such that there are sufficient targets inside each pixel to learn about contamination without introducing significant Poisson noise in the mitigation.

For the DESI QSO target selection, a pixel with a size of  $N_{\text{side}} = 256$  typically contains a median of  $\sim 16$  QSO targets. At a size of



**Figure 16.** Evolution of the angular correlation function in the North region when the probability threshold of the quasar target selection is increased. The nominal probability in the North is  $p(r) = 0.88 - 0.04 \times \tanh(r - 20.5)$ . The dashed lines depict raw target selections with different probability thresholds. The solid lines show the corrected versions of the same selections. The black dotted line corresponds to the typical resolution of the correction, i.e. the pixel size at  $N_{\text{side}} = 256$ .

$N_{\text{side}} = 512$ , the number of targets decreases to  $\sim 4$ . With such a small number of targets, the per-pixel density is too noisy for machine learning methods to relate fluctuations in density to observational features. For the same reason, we needed to increase the pixel size used in our analysis to  $N_{\text{side}} = 128$  when studying eBOSS mocks, since the density of the eBOSS mocks was lower than the DESI QSO target selection density (see Appendix A).

Other DESI targets, such as ELGs and LRGs, will have a higher density and it will be possible to decrease the size of the pixels used to estimate correction weights by a factor of 2–4.

#### 5.2.4 Systematic checks: restrictive QSO target selection

Since the DESI QSO targets are selected with an RF classification, the stellar contamination depends on the value of the selection threshold. This dependence propagates into the angular correlation, since we broadly expect a lower  $r_0$  for a less contaminated sample.

Fig. 16 shows the angular correlations in the North for DESI QSO targets with different values of the selection threshold (dashed lines) and of the corresponding corrected samples (full lines). The systematic weights are generated with the RF method for each target selection. The nominal probability in the North footprint used in the QSO selection depends on  $r$  and is given by  $p(r) = 0.88 - 0.04 \times \tanh(r - 20.5)$ . This threshold is lower than for the other regions. When the threshold increases, the selection is less contaminated since the RF will select a higher fraction of bona fide QSOs.

The amplitude of the angular correlation in Fig. 16 decreases as the probability threshold is increased, suggesting that the excess of correlation is mostly due to stellar contamination. This also confirms that our method to mitigate the systematics does not overfit the data since when the contamination is removed, the correlation converges to the same level as the SDSS QSO correlation (cf. Fig. 14).

Finally, we note that the excess correlation in the South is *not* reduced when the probability threshold of the RF selection is increased. This is because many stars in the Sagittarius Stream have

selection features that resemble QSOs. These stars are assigned near unit probability by the DESI QSO targeting algorithm, such that they are not removed by a more restrictive selection.

### 5.3 Limber parameters

#### 5.3.1 Limber approximation

The angular correlation function  $w(\theta)$  is related to the 3D one  $\xi(r)$  through

$$w(\theta) = \int_0^\infty \int_0^\infty \xi(r_{12}) \tilde{S}(r_1) r_1^2 dr_1 \tilde{S}(r_2) r_2^2 dr_2, \quad (12)$$

where  $r_{12} = \sqrt{r_1^2 + r_2^2 - 2r_1r_2 \cos(\theta)}$  and  $\tilde{S}(r)$  represents the probability for an object at distance  $r$  to be observed by the survey. It is defined from the selection function  $S(r)$  by

$$\tilde{S}(r) = \frac{S(r)}{\int_0^\infty S(r) r^2 dr}. \quad (13)$$

We will estimate  $S(r)$  by multiplying the quasar luminosity function from Palanque-Delabrouille et al. (2013) by the completeness of the DESI QSO target selection.

Adopting the approximation that the correlation function is non-negligible only for small values of  $r$ , i.e. that the angular correlation function is non-zero only for small values of  $\theta$ , one can assume that  $x = r_2 - r_1$  and  $y = (r_1 + r_2)/2$ , and hence derive Limber’s equation valid only for small angles (for an explicit derivation, see for instance Kurki-Suonio 2019):

$$w(\theta) \simeq \int_0^\infty \tilde{S}(y)^2 y^4 \int_{-\infty}^\infty \xi(\sqrt{x^2 + y^2 \theta^2}) dx dy. \quad (14)$$

Assuming a power law for the correlation function:

$$\xi(r) = \left(\frac{r}{r_0}\right)^{-\gamma}, \quad (15)$$

the angular correlation function becomes

$$w(\theta) = \sqrt{\pi} \frac{\Gamma(\frac{\gamma-1}{2})}{\Gamma(\frac{\gamma}{2})} r_0^\gamma \theta^{1-\gamma} \int_0^\infty \tilde{S}^2(y) y^{5-\gamma} dy, \quad (16)$$

where  $\Gamma$  is the gamma function.

The integral calculation requires a fiducial cosmology. We choose a  $\Lambda$  cold dark matter ( $\Lambda$ CDM) cosmology following the *Planck* 2015 parameters from Planck Collaboration XIII (2016):  $\Omega_{\text{m},0} = 0.308$ ,  $\Omega_{\Lambda,0} = 0.691$ ,  $\Omega_{\text{b},0} = 0.048$ ,  $h = 0.677$ ,  $\sigma_8 = 0.815$ , and  $n_s = 0.967$ .

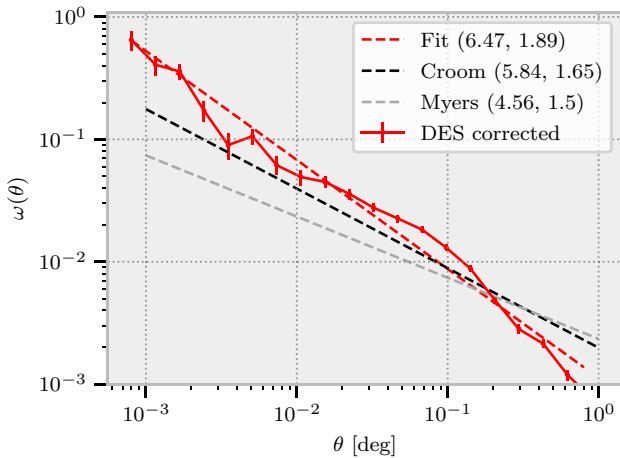
#### 5.3.2 Fitting of Limber parameters

As explained in Section 5.3.1, the power-law parametrization of the correlation function  $\xi(r)$  can be constrained by the angular correlation function  $w(\theta)$  at small angles. We proceed by estimating these power-law parameters and comparing them to previous measurements done with SDSS data by Myers, White & Ball (2009) and with 2dF data by Croom et al. (2005).

The Limber parameters are estimated in the three imaging regions highlighted in Fig. 1 and also in the DES region with only Dec.  $> -30$ , which we will refer to as DES(Dec.  $> -30$ ). This region will almost correspond to the intersection between DES and the expected nominal DESI footprint. The results are shown in Table 2 and fit for the corrected DESI QSO targets in DES region is plotted in Fig. 17. The correlation function is fitted from  $1 \times 10^{-3}$  to 0:8, corresponding to a transverse separation of 0.045–45  $h^{-1}$  Mpc at redshift 1.7. The mean redshift of the DESI QSO targets ( $\sim 1.7$ ) is obtained tanks to

**Table 2.** Limber parameters for the DESI QSO target selection in the imaging footprints depicted in Fig. 1. The errors are estimated using a subsampling method. We provide the measurements for both the non-corrected and corrected cases. The correction is performed with the RF method.  $\bar{z}$  is the mean redshift of the sample.

	$r_p$ min ( $h^{-1}$ Mpc)	$r_p$ max ( $h^{-1}$ Mpc)	$\bar{z}$	$r_0$ ( $h^{-1}$ Mpc)	$\gamma$	$r_0$ ( $h^{-1}$ Mpc)	$\gamma$
Croom	1.0	25	1.35	–	–	$5.84 \pm 0.33$	$1.64 \pm 0.04$
Myers	1.6	40	2	–	–	$4.56 \pm 0.48$	1.5 (fixed)
				DESI QSO targets		Corrected	
North	0.045	45	1.7	$10.15 \pm 0.70$	$1.61 \pm 0.03$	$7.49 \pm 0.57$	$1.89 \pm 0.02$
South	0.045	45	1.7	$12.88 \pm 0.95$	$1.64 \pm 0.03$	$10.33 \pm 0.84$	$1.80 \pm 0.02$
DES (Dec. $> -30$ )	0.045	45	1.7	$6.76 \pm 0.58$	$1.78 \pm 0.04$	$6.15 \pm 0.45$	$1.88 \pm 0.04$
DES	0.045	45	1.7	$7.19 \pm 0.34$	$1.79 \pm 0.03$	$6.47 \pm 0.31$	$1.89 \pm 0.02$



**Figure 17.** Fit of the Limber parameters in the DES region for the corrected DESI QSO targets (red line). The Limber parameters ( $r_0, \gamma$ ) are given in each case. The black (respectively grey) dashed line shows the Limber function measured by Croom et al. (2005) (respectively Myers et al. 2009). The parameters for the other regions are given in Table 2.

the selection function introduced in Section 5.3.1, which corresponds to the estimated redshift distribution of the sample.

The errors on the parameters  $r_0$  and  $\gamma$  are estimated using a subsampling method: 10 (respectively 18, 10, 4) subregions of similar area ( $\sim 450$  deg<sup>2</sup>) are used for the North (respectively South, DES, DES Dec.  $> -30$ ). Subsampling is used to probe the variability of the angular correlation function in different areas of the footprint, where stellar contamination and systematic effects should differ. If the systematics are properly mitigated, the error of each parameter should be reduced, i.e. the corrected angular correlation functions over each subregion should be more similar.

The value  $r_0$ , which parametrizes the amplitude of the correlation function, captures the offset of different measured angular correlation functions. For instance, the fact that the angular correlation in the South is higher than in DES, manifests in a value of  $r_0$  that is higher in the South than in DES. As systematics tend to lead to an increase in amplitude,  $r_0$  tends to be higher when systematics are dominant. For example, the fact that the value of  $r_0$  that we measure in the DES footprint is generally comparable to the values found for previous measurements (Croom et al. 2005; Myers et al. 2009), suggests that systematics have been mitigated well in the DES region.

The parameter  $\gamma$  describes the slope of the angular correlation. A higher value of  $\gamma$  means a steeper slope. It is worth noting that since the relevant correction scales of the systematic weights are

larger than the pixel size used for our analyses ( $\sim 0.2$ ), the angular correlations of the corrected targets have a steeper slope than for the raw target samples. The comparison of  $\gamma$  between the DESI QSO targets and previous measurements is not particularly relevant since any excess correlation caused by systematics on small scales, i.e. below the typical angular resolution of our corrections, cannot be removed with our mitigation procedure.

Though no spectroscopic confirmation of DESI QSO targets is available across a significant fraction of the footprint used in our study, our correction for angular systematics enables us to provide a Limber amplitude parameter  $r_0$  that is consistent with those found by Croom et al. (2005) and Myers et al. (2009) using spectroscopic data. Nevertheless, the fitting of the Limber parameters will be greatly improved by information from the DESI spectroscopic survey. Follow-up spectroscopy will enable us to remove stars from the DESI QSO target sample, so any clustering analyses will be free of the stellar contamination that increases the amplitude of the angular correlation measurements.

## 6 CONCLUSION

The QSO target selection for the 5-yr DESI survey, which is based on the DR9 of the Legacy Imaging Surveys, has recently been finalized. This study illustrates imaging systematic effects and stellar contamination in the DESI QSO selection, with a view to studying and correcting for large-scale density fluctuations that will impact DESI clustering measurements.

After a brief description of the current QSO target selection for DESI, we explain the different effects that cause density variations in the target sample. These effects can be separated into three categories.

(i) Stellar contamination from the Milky Way. This occurs mainly near to the Galactic plane and is caused by the large excess of stars in this region.

(ii) Impact of the depth of the observations in the  $z$  and  $W2$  bands. These two bands are crucial for DESI QSO target selection. Hence,  $z$ -band systematics and above-average  $W2$  PSF depth produce strong overdensities, which are most pronounced in the North of the Legacy Surveys imaging. The impact of the depth in the  $W2$  band is hidden in the South by the numerous stars from the Milky Way.

(iii) Stellar contamination from the Sagittarius Stream. These stars have similar colours to QSOs and thus heavily contaminate the DESI QSO target selection. A map of the Sagittarius Stream enables us to remove the contamination to first order but is insufficient to correct for higher order effects, as demonstrated by the angular correlation function of DESI QSO targets in this region. The part of the remaining correlation that comes from Sagittarius Stream stars cannot currently

be removed with imaging maps alone, either by improving the QSO target selection method or by improving the Sagittarius Stream map. This is because many stars in the Stream are too faint to have been previously observed. These stars prevent us from properly analysing the impact of other imaging features on target selection and the excess correlation observed in this region. However, this analysis will be done once these stars are removed by spectroscopy with DESI.

The QSO target selection, incorporating weights to correct for systematics, was validated by computing the angular correlation function in each imaging region of the DESI Legacy Imaging Surveys. We show that after mitigating for systematics, the angular correlation function from the target selection is comparable to the angular correlation function of SDSS DR16 quasars, except for in the Sagittarius Stream region (as explained above). Our results are very encouraging for the upcoming DESI spectroscopic survey, since our analysis can produce similar angular correlation functions as derived for SDSS spectroscopic data, despite utilizing only information from imaging.

The method described in this paper was built and optimized specifically for the DESI QSO target selection. In particular, the hyperparameters for the NN and the RF methods are specific to DESI QSO targets. Nevertheless, our approach could readily be adapted to any DESI target class. In particular, the form of the  $K$ -fold training was developed independently of the target class.

Once the DESI survey progresses, we will apply the method outlined in this paper to compute imaging systematic correction weights for the DESI spectroscopic QSO catalogue. We expect these corrections to be smaller for future DESI samples, since many first-order systematic effects that are caused by stellar contamination will be removed by spectral information.

## ACKNOWLEDGEMENTS

This research is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under contract no. DE-AC02-05CH11231, and by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; additional support for DESI is provided by the U.S. National Science Foundation, Division of Astronomical Sciences under contract no. AST-0950945 to the NSF's National Optical–Infrared Astronomy Research Laboratory; the Science and Technology Facilities Council of the United Kingdom; the Gordon and Betty Moore Foundation; the Heising-Simons Foundation; the French Alternative Energies and Atomic Energy Commission (CEA); the National Council of Science and Technology, Mexico; the Ministry of Economy of Spain, and by the DESI Member Institutions.

ADM was supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Number DE-SC0019022.

The authors are honoured to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation.

## DATA AVAILABILITY

The Data Release 9 (DR9) of the DESI Legacy Imaging Surveys is available via <https://www.legacysurvey.org/dr9/>. The systematic weights derived in this paper will be shared on reasonable request to the corresponding author.

## REFERENCES

- Ahumada R. et al., 2020, *ApJS*, 249, 3  
 Alonso D., 2012, preprint ([arXiv:1210.1833](https://arxiv.org/abs/1210.1833))  
 Antoja T., Ramos P., Mateu C., Helmi A., Anders F., Jordi C., Carballo-Bello J. A., 2020, *A&A*, 635, L3  
 Bautista J. E. et al., 2021, *MNRAS*, 500, 736  
 Burleigh K. J., 2018, PhD thesis, Univ. California, Berkeley  
 Castorina E. et al., 2019, *J. Cosmol. Astropart. Phys.*, 09, 010  
 Croom S. M. et al., 2005, *MNRAS*, 356, 415  
 Dawson K. S. et al., 2016, *AJ*, 151, 44  
 De Mattia A. et al., 2021, *MNRAS*, 501, 5616  
 Dembinski H. et al. 2020, *scikit-hep/iminuit: v2.0.0 (v2.0.0)*. Zenodo, available at <https://zenodo.org/record/4310361#.YaWm6dBBzIU>  
 DES Collaboration, 2021, *ApJS*, 255, 20  
 DESI Collaboration, 2016, preprint ([arXiv:1611.00036](https://arxiv.org/abs/1611.00036))  
 Dey A. et al., 2019, *AJ*, 157, 168  
 du Mas des Bourboux H. et al., 2020, *ApJ*, 901, 153  
 Elsner F., Leistedt B., Peiris H. V., 2016, *MNRAS*, 456, 2095  
 Elvin-Poole J. et al., 2018, *Phys. Rev. D*, 98, 61  
 Flaugher B. et al., 2015, *AJ*, 150, 150  
 Gaia Collaboration, 2018, *A&A*, 616, A1  
 Gil-Marín H. et al., 2020, *MNRAS*, 498, 2492  
 Gorski K. M., Hivon E., Banday A. J., Wandelt B. D., Hansen F. K., Reinecke M., Bartelmann M., 2005, *ApJ*, 622, 759  
 Ho S. et al., 2012, *ApJ*, 761, 14  
 Hou J. et al., 2021, *MNRAS*, 500, 1201  
 Kalus B., Percival W. J., Bacon D. J., Mueller E. M., Samushia L., Verde L., Ross A. J., Bernal J. L., 2019, *MNRAS*, 482, 453  
 Kingma D. P., Ba J., 2014, preprint ([arXiv:1412.6980](https://arxiv.org/abs/1412.6980))  
 Kitanidis E. et al., 2020, *MNRAS*, 496, 2262  
 Kong H. et al., 2020, *MNRAS*, 499, 3943  
 Kurki-Suonio H., 2019, Spring 2019: Galaxy Survey Cosmology, Part 1. Available at <http://www.courses.physics.helsinki.fi/teor/gsc/>  
 Landy S. D., Szalay A. S., 1993, *ApJ*, 412, 64  
 Leistedt B., Peiris H. V., Mortlock D. J., Benoit-Lévy A., Pontzen A., 2013, *MNRAS*, 435, 1857  
 Leistedt B. et al., 2016, *ApJS*, 226, 24  
 Lyke B. W. et al., 2020, *ApJS*, 250, 8  
 Majewski S. R., Skrutskie M. F., Weinberg M. D., Ostheimer J. C., 2003, *ApJ*, 599, 1082  
 Meisner A. M., Lang D., Schlegel D. J., 2017, *AJ*, 154, 161  
 Mueller E.-M. et al., 2021, preprint ([arXiv:2106.13725](https://arxiv.org/abs/2106.13725))  
 Myers A. D. et al., 2006, *ApJ*, 638, 622  
 Myers A. D., White M., Ball N. M., 2009, *MNRAS*, 399, 2279  
 Myers A. D. et al., 2015, *ApJS*, 221, 27  
 Neveux R. et al., 2020, *MNRAS*, 499, 210  
 Newberg H. J. et al., 2002, *ApJ*, 569, 245  
 Palanque-Delabrouille N. et al., 2013, *A&A*, 551, A29  
 Pedregosa F. et al., 2011, *J. Machine Learning Res.*, 12, 2825  
 Planck Collaboration XIII, 2016, *A&A*, 594, A13  
 Prakash A. et al., 2016, *ApJS*, 224, 34  
 Raichoor A. et al., 2021, *MNRAS*, 500, 3254  
 Rezaie M., Seo H. J., Ross A. J., Bunesco R. C., 2020, *MNRAS*, 495, 1613  
 Rezaie M. et al., 2021, *MNRAS*, 506, 3439  
 Ross A. J. et al., 2011, *MNRAS*, 417, 1350  
 Ross A. J. et al., 2013, *MNRAS*, 428, 1116  
 Ross A. J. et al., 2017, *MNRAS*, 464, 1168  
 Ross A. J. et al., 2020, *MNRAS*, 498, 2354  
 Rybicki G. B., Press W. H., 1992, *ApJ*, 398, 169  
 Schlafly E. F., Finkbeiner D. P., 2011, *ApJ*, 737, 103  
 Schlegel D. J., Finkbeiner D. P., Davis M., 1998, *ApJ*, 500, 525  
 Silva D. R. et al., 2016, *Am. Astron. Soc. Meeting Abstr.*, #228, 317.02  
 Suchyta E. et al., 2016, *MNRAS*, 457, 786  
 Tamone A. et al., 2020, *MNRAS*, 499, 5527  
 Tegmark M., Hamilton A. J. S., Strauss M. A., Vogeley M. S., Szalay A. S., 1998, *ApJ*, 499, 555  
 Wang Y. et al., 2020, *MNRAS*, 498, 3470



Weaverdyck N., Huterer D., 2021, *MNRAS*, 503, 5061  
 Wright E. L. et al., 2010, *AJ*, 140, 1868  
 Yèche C. et al., 2020, *Res. Notes AAS*, 4, 179  
 Zhao C. et al., 2021, *MNRAS*, 503, 1149  
 Zonca A., Singer L., Lenz D., Reinecke M., Rosset C., Hivon E., Gorski K., 2019, *J. Open Source Softw.*, 4, 1298  
 Zou H. et al., 2019, *ApJS*, 245, 4

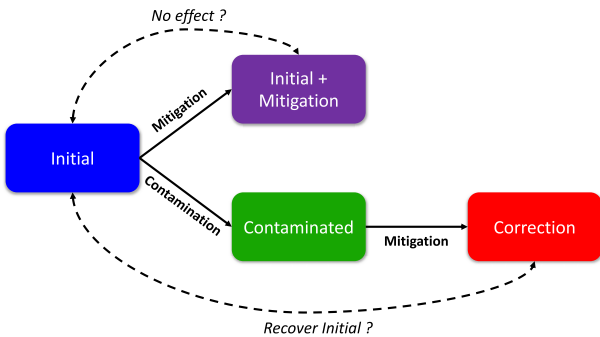
## APPENDIX A: VALIDATION WITH MOCKS

To validate the systematic mitigation method and avoid overfitting, we use a set of 100 QSO EZ-mocks from eBOSS (Zhao et al. 2021). They have a smaller density than the nominal target selection density and a smaller area. However, it will not impact the analysis since the surface and the density are enough for our test. Besides, since our analysis for the legacy survey is different for the three photometric footprints, we will mimic one footprint only, similar to a subregion of the North footprint.

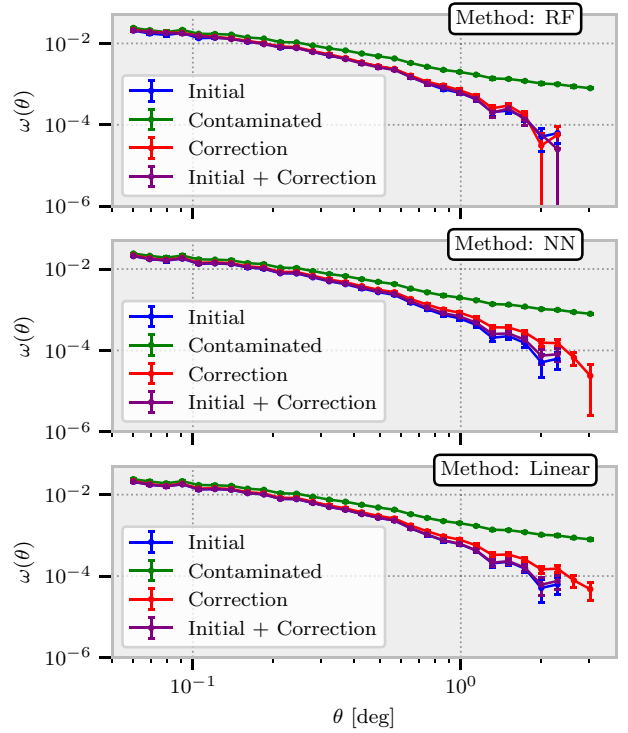
The mocks contain a cosmological signal but no systematic effects. They are contaminated using the inverse weight estimated previously and this contamination will be mitigated with our method using the same observational features.

We present here contamination estimated with the RF method (it is the inverse of the map shown in Fig. 11). We test the method in two different ranges of contamination: one describing a weakly contaminated case (extracted from the RA, Dec.  $\in [100^\circ, 270^\circ] \times [32^\circ, 60^\circ]$  box) where all the methods work well; a second describing a strongly contaminated case (extracted from the RA, Dec.  $\in [120^\circ, 290^\circ] \times [55^\circ, 90^\circ]$  box) where the linear method is inefficient as shown in the systematics plots of Fig. 7. To compare the method efficiency and avoid biasing the results, we also applied an NN-based estimation for the strong contaminated case.

The validation pipeline is shown in Fig. A1. There are two different tests (dashed lines). The first one is to check whether our method overfits the data. We apply the mitigation method to the initial (uncontaminated) mocks such that no correction is expected. The angular correlations of the uncontaminated mocks and of the corrected uncontaminated mocks should be identical. The second test is to validate our mitigation. We apply the mitigation method to the contaminated mocks, expecting to recover the same angular correlation as for the initial mocks.



**Figure A1.** Pipeline validation with mocks. The initial mocks (blue) are uncontaminated and contain a cosmological signal. They are contaminated with systematic weights (green). Then, they are corrected with the mitigation presented above (red). We compare the corrected mock to the initial mock to verify if we recover the correct initial state. As a sanity check, we also applied the mitigation to the initial mock (purple) to verify whether our mitigation technique overfits target density variations. These tests are represented by the two black dashed lines.



**Figure A2.** Weakly contaminated case: mean on 100 eBOSS EZ-mocks of the angular correlation functions. The error bars are the standard deviation between all realizations. Colours follow the same scheme as in Fig. A1. The systematic contamination is done with a RF estimation of a weakly contaminated area (RA, Dec.  $\in [100, 270] \times [32, 60]$ ) of the DR9 of Legacy Surveys. The method does not overfit the data and recovers the initial correlation. Note that the contamination was estimated with the RF method that explains the small residual systematics with the NN and linear correction.

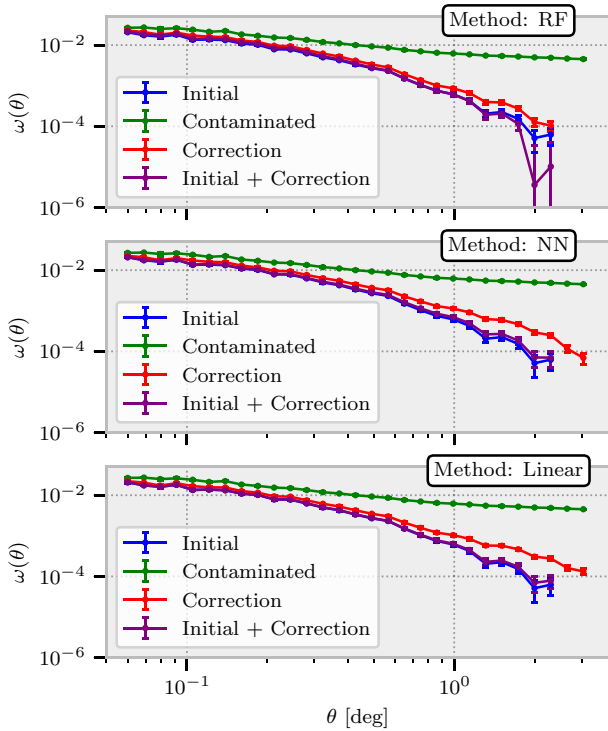
Some subtleties have to be considered. First, the density of these mocks ( $70 \text{ deg}^{-2}$ ) is lower than that of the quasar targets ( $300 \text{ deg}^{-2}$ ). Thus, the size of the correction cannot be the same since the number of objects inside a pixel of  $N_{\text{side}} = 256$  is too small. We therefore downgrade the contamination and the observational feature maps to  $N_{\text{side}} = 128$ .

Secondly, the footprint of these mocks is different from the North footprint. The size of the training sample is here twice smaller as the data. We, therefore, use only three folds instead of six for the DR9 North training. It is worth noting that we need to change the position of the folds to cover the footprint correctly as explained in Section 3.3.2, especially for the strongly contaminated case.

Nevertheless, the training conditions are slightly different and we expect some differences for the second test.

The angular correlation functions for the weakly contaminated case are shown in the Fig. A2. Each correlation function is the average of the angular correlations of the 100 mocks and the error bars are the standard deviation between all the realizations. We show the correlation function up to  $\theta \sim 3.5$ , which corresponds to a comoving angular distance of  $\sim 250 \text{ Mpc } h^{-1}$  at  $z = 1.7$ , the maximum size of the eBOSS analysis. The result for each regression method is shown in the three panels.

The purple lines are the angular correlations of the initial mocks mitigated with our methods. In the three cases, they lie on top of the blue lines that show the angular correlations of the initial mocks. This indicates no sign of overfitting compare to the significance of the correction. The green lines are the angular correlations for the mocks



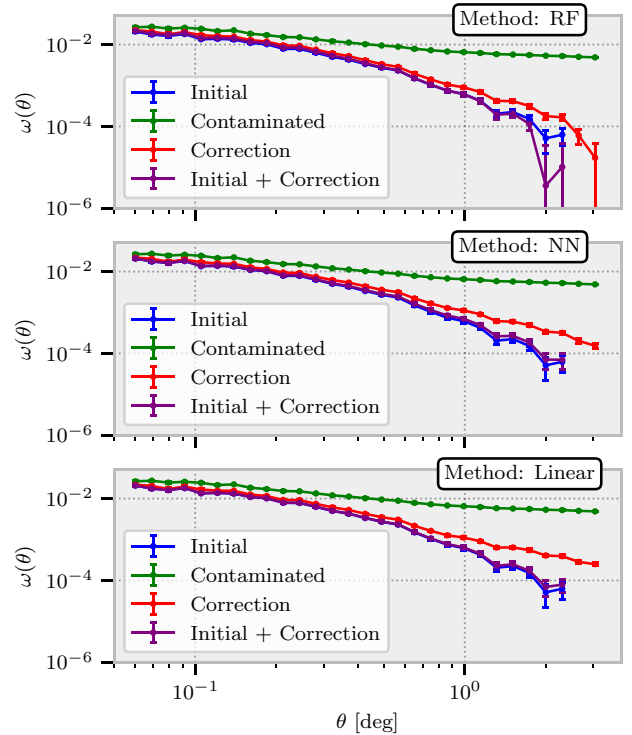
**Figure A3.** Strongly contaminated case: same as Fig. A2 but the systematic contamination depicts strong contaminated area (RA, Dec.  $\in [120, 290] \times [55, 90]$ ). The correction is not perfect since the training sample is smaller than the Legacy Surveys case and cannot contain the same information in each fold. Besides, the correction is done with a higher  $N_{\text{side}}$  that can explain that we do not fully recover the initial state.

after contamination. The red lines are the angular correlations for the contaminated mocks after mitigation. The red line exactly recovers the blue line in the case of the RF method. In the two other cases, the red lines recover more or less the blue lines and the slight difference stems from the contamination was done from an estimation of the systematics by the RF method that benefits the RF case.

The angular correlations for the strongly contaminated case are shown in Fig. A3 for the RF-based contamination and in Fig. A4 for the NN-based one. The two figures are similar, indicating that the contaminated method does not benefit the RF or NN mitigation. The result from the same contamination method can be compared together.

Here again, there is no indication of overfitting for the NN and linear method. For the RF method, the purple line is slightly below the blue one above  $1^\circ$  indicating tiny overfitting. Note that the mocks that we used here cover a smaller area than the DESI QSO sample, thereby making our method more prone to overfit for this test case. Optimization of the hyperparameters to reduce overfitting on the DESI footprint will be pursued with DESI mocks. However, this tiny overfitting is much less significant than the level of systematic corrections validating that most of the impact of RF weights on the angular correlation function of the DESI QSO sample is real systematics mitigation and not overfitting.

The angular correlation of contaminated mocks is higher as a consequence of the stronger contamination compared to the previous case. The red lines do not perfectly recover the blue lines even in the case of the RF regression. Indeed, the smaller size of the training sample and the modification of the fold form prevent the NN and RF methods to learn all the information needed to fully



**Figure A4.** Strongly contaminated case: same as Fig. A3 but the systematic contamination that depicts strong contaminated area (RA, Dec.  $\in [120, 290] \times [55, 90]$ ) is estimated with the NN method.

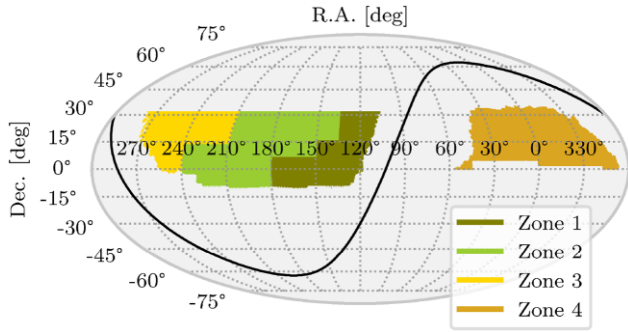
correct this contamination. In addition, the NN method is less efficient to correct small contaminated regions without any additional hyperparametrization like regularization terms. The RF approach is more robust than the NN one to variations in the training sample and less dependent on hyperparametrization. For the linear method, the correction is less efficient: in the region of extreme observational features, the systematic effects are less corrected, as shown in Section 4.1.

This set of mocks and the second test are used to optimize the hyperparameters of the NN and perform the grid search method (Section 3.3.4). Given the subtleties presented above, however, it is expected that the hyperparameters used here are not the best one for the DR9 training. More optimized parameter tuning will eventually be achieved with mocks matching the DESI QSO samples.

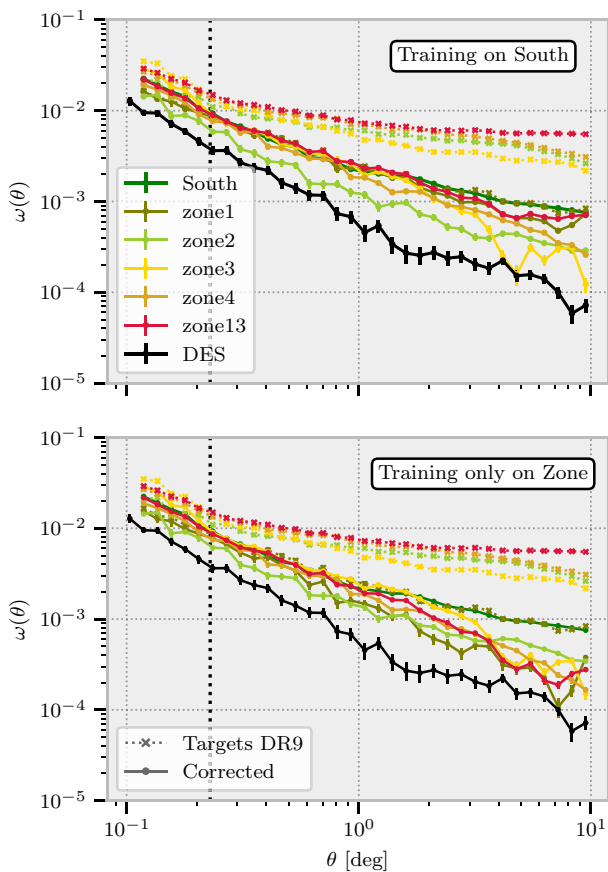
## APPENDIX B: SOUTH FOOTPRINT AND SAGITTARIUS STREAM CONTAMINATION

As shown in Fig. 14, even after the systematic mitigation, we do not recover the same level of correlation in the South footprint as in the two other regions. To analyse the excess of correlation in the South footprint, we divide it into four zones, represented in Fig. B1. These zones are as follows.

- (i) Zone 1 near the antagalactic pole shows less overdensity than the other regions near the Galactic plane. Its lower density is due to the lower value of the PSF depth  $W_2$  in this zone.
- (ii) Zone 2 contains the Sagittarius Stream and it is strongly contaminated by stars.
- (iii) Zone 3 and zone 4 show strong overdensity due to stars from the galactic plane. In addition, zone 4 describes the SGC part of the South footprint and we considered also zone 13 combining zone 1



**Figure B1.** The South region is split in four distinct zones to analyse the stars contamination. Zone 1 near the antagalactic pole shows less overdensity than the three others. Zone 2 contains the Sagittarius Stream. Zone 3 and zone 4 show strong overdensity due to stars from the galactic plane.



**Figure B2.** The angular correlation for the four zones that are represented in Fig. B1. Dotted lines are the correlations of DR9 targets, and the solid ones are for the corrected targets. For comparison, we also plot the correlation in the South region (green line) and in the DES region (black line) that is known as the least contaminated region. In the top panel, corrections are calculated on all the South footprint. In the bottom panel, corrections are calculated in each zone independently.

with zone 3 that describes the NGC without the Sagittarius Stream part of the South.

The angular correlation functions for these different zones are shown in Fig. B2. Here each mitigation is performed with the RF method. The top panel shows the correction done with training on all the South footprint. The bottom panel shows the correction where the

training was done individually on each consider zone. We add also the angular correlation for the corrected targets in South (in green) and in DES (in black) for reference. The correlations of the corrected targets on each zone are lower than on all the South footprint but it is not the average on zones due to the missing cross-terms.

The excess of correlation at small scales cannot be removed by our method. It is due to either stars or non-considered features. The part causes by a non-considered feature is left for a future study with the spectroscopic sample. The excess causes by the stellar contamination will be removed with the spectroscopic data as explained in Section 5.2.4. Hence, we will discuss the excess of correlation at large scales compared to the correlation in DES.

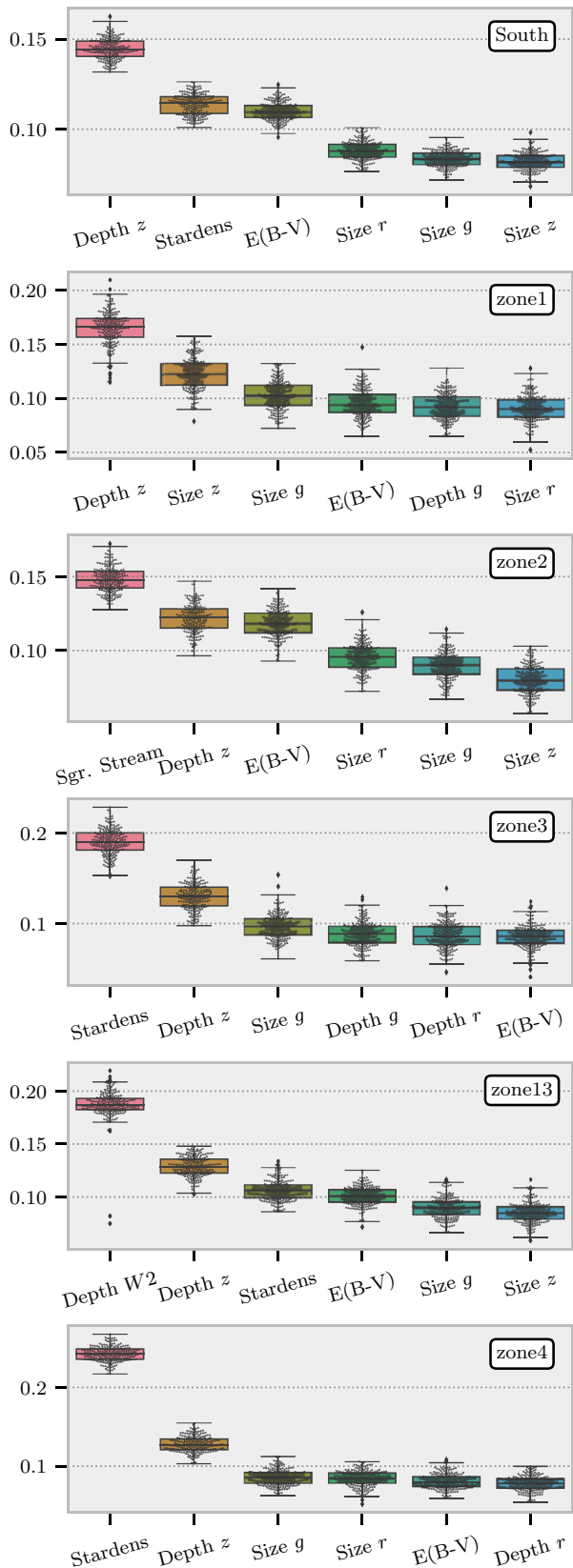
The correction estimated with the training on all South footprint is not sufficient to recover the same level of correlation as in DES after mitigation for all zones. The correction in zone 2 is more efficient than in the other zones since the Sagittarius Stream feature separates the Sagittarius Stream from the rest during the training. This zone does not recover the same angular correlation at large scales as in DES since the Sagittarius Stream feature is built as the spatial average of candidates stars suppressing the angular correlation information contained in this feature.

The angular correlation in zone 1 without systematic mitigation is lower than those in the three other zones since this region does not show strong overdensity. However, the systematic mitigation is inefficient in this region when the training is done on all the South footprint. This is not the case when we apply the mitigation from the training only on zone 1. Our method is unable to extract the correct information for this zone in all the South. Zone 4 and zone 13 are better when the training is performed only on each zone, recovering a level of correlation that is in more agreement with that of DES at large scales.

The inefficiency of the regression in all the South can be explained by the numerous stars in this region that bias the information about the observational systematic effects preventing the regression to learn correctly the true observational features dependency of the relative target density. For instance, the PSF depth  $W_2$  is not characterized as an important feature (cf. Section 4.2), while it explains the lower density observes around the antagalactic pole. Fig. B3 shows the feature importance estimated with the training on the consider zone only. The feature importance for all the zone used is plotted. PSF depth  $W_2$  is almost uniform (cf. Fig. 4 for its distribution) in zone 1 and zone 3 and so does not appear as an important feature when the training is done in each region. However, it is found as the most important feature in the training on zone 13 illustrating that the stars in zone 2 and in zone 4 bias the training, e.g. the fluctuation of the relative density in the function of the PSF depth  $W_2$  in zone 2 is masked by the presence of numerous stars from the Sagittarius Stream.

Even if the training in an individual zone is more effective, especially at large scales than the training in all the South, we do not recover the same level correlation as in DES at intermediate scales. This can be explained still by the presence of numerous stars in each zone. The stellar contamination is mixed with the imaging systematics and the two effects are not easily separable. Then, the excess of correlation can be caused either by the stellar contaminant or by an additional unconsidered imaging feature.

This analysis could be performed once the spectroscopic survey is done since the stellar contamination will vanish. The impact of imaging features can be then studied without any significant bias. If the set of features introduced during this analysis contains all the information, our systematic mitigation method will be able to



**Figure B3.** Same as Fig. 10, but the feature importance plotted is estimated with the training only on the consider zone. The different behaviours on each zone are highlighted by the different features that are qualified as important. In particular, the density of stars from the galactic plane prevents the regression to learn correctly the role of the PSF depth  $W2$  as expected.

correctly learn the true dependence on the observational features as in DES or the North and recover the correct angular correlation.

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.