Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **486,** 1377–1391 (2019)
Advance Access publication 2019 March 27

doi:10.1093/mnras/stz856

Star formation rates for photometric samples of galaxies using machine learning methods

M. Delli Veneri,[1]★ S. Cavuoti,[1,2,3]★ M. Brescia,[1] G. Longo[2,3] and G. Riccio[1]

[1] *INAF – Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131, Napoli, Italy*
[2] *Department of Physics 'E. Pancini', University Federico II, via Cinthia 6, I-80126, Napoli, Italy*
[3] *INFN section of Naples, via Cinthia 6, I-80126, Napoli, Italy*

Accepted 2019 March 19. Received 2019 March 19; in original form 2018 November 16

ABSTRACT

Star formation rates (SFRs) are crucial to constrain theories of galaxy formation and evolution. SFRs are usually estimated via spectroscopic observations requiring large amounts of telescope time. We explore an alternative approach based on the photometric estimation of global SFRs for large samples of galaxies, by using methods such as automatic parameter space optimisation, and supervised machine learning models. We demonstrate that, with such approach, accurate multiband photometry allows to estimate reliable SFRs. We also investigate how the use of photometric rather than spectroscopic redshifts, affects the accuracy of derived global SFRs. Finally, we provide a publicly available catalogue of SFRs for more than 27 million galaxies extracted from the Sloan Digital Sky Survey Data Release 7. The catalogue will be made available through the Vizier facility.

Key words: methods: data analysis – techniques: photometric – catalogues – galaxies: distances and redshifts – galaxies: photometry.

1 INTRODUCTION

During the last few year, multiwavelength surveys have led to a remarkable progress in producing large galaxy samples that span a huge variety of galaxy properties and redshift. All together, these data provided us with reliable information for many hundred thousand galaxies (Abazajian et al. 2009; Salvato et al. 2009, 2011; Cardamone et al. 2010; Matute et al. 2012; Marchesi et al. 2016) and have triggered similar improvements in the determination of physical parameters crucial to understand and constrain galaxy formation and evolution. Among these parameters, the global star formation rate (SFR; Madau & Dickinson 2014) provides a luminosity-weighted average across local variations in star formation history and physical conditions within a given galaxy.

Broadly speaking, SFR estimators are usually derived from measured fluxes, either monochromatic or integrated over some specific wavelength ranges, selected in order to be sensitive to the short-lived massive stars present in a given galaxy. In the literature, there is a large variety of such estimators spanning from the UV/optical/near-IR range (∼0.1–5 μm), which probes the stellar light emerging from young stars, to the mid/far-IR (∼5–1000 μm), which instead probes the stellar light reprocessed by dust (Kennicutt 1998; Kennicutt & Evans 2012). Other estimators rely on the gas ionized by massive stars (Calzetti et al. 2004; Hong et al. 2011), hydrogen recombination lines, forbidden metal lines, and in the millimetre range, the free–free (Bremsstrahlung) emission (Schleicher & Beck 2013). Finally, other estimators can, at least in principle, be derived in the X-ray domain, from X-ray binaries, massive stars, and supernovae via the non-thermal synchrotron emission, following early suggestions by Condon (1992).

An ample literature, however, shows that the correct derivation of SFRs from optical/FIR broad-band data is a highly non-trivial task, due to the complex and still poorly understood correlation existing between the SFR and the broad-band photometric properties integrated over a whole galaxy (Rafelski et al. 2016; Fogarty et al. 2017; Cooke et al. 2018; Pearson et al. 2018).

Each estimator is sensible to a specific and different SFR time-scale and thus a proper understanding of the SFR phenomenology requires a combination of different estimators; in particular, UV and total IR radiations are sensible to the longer time-scales, $\sim 10^8$ yr, while the ionizing radiation is sensitive to the shortest time-scales, $\sim 10^6$ yr. Furthermore, optical and UV estimators often need corrections to account for dust presence and, for this reason, they are not used on their own, but in combination with other estimators (Calzetti et al. 2007). Another methodology suitable to estimate SFRs for large samples of objects is the so-called spectral energy distribution (SED) template fitting, which compares an observed galaxy spectrum with a large data base of template spectra, generated by stellar population synthesis models (Conroy 2013). This method, however, suffers from the age–dust–metallicity degeneracy and, in order to reliably measure ages and hence SFRs, high-quality data are required and, due to the choice of template

★ E-mail: micheledelliveneri@gmail.com (MDV);
stefano.cavuoti@gmail.com (SC)

spectra, severe biases are often introduced in the resulting ages. In a seminal paper Wuyts et al. (2011), SFRs for galaxies at $z_{spec} \sim 3$ were derived using all the methods previously explained, finding that all estimators agree with no systematic offset, providing that an extra attenuation towards H II regions is included when modelling the H α SFRs. Nevertheless, the same paper also concluded that, at high redshift, nebular emission lines may introduce a systematic uncertainty affecting the derived specific SFRs by a factor of 2. This work takes place in the framework of the new discipline of Astroinformatics, which aims at allowing the scientific exploitation of large data sets produced by the modern digital, panchromatic and multi-epoch surveys, using a variety of techniques largely derived from, but not restricted to, the statistical learning domain. In this framework, a new viable approach to obtain SFR estimates for large samples of objects was recently presented by Stensbo-Smidt et al. (2017), who transformed the SFR estimation into a machine learning (ML) non-linear regression problem. With this method, the only prerequisite is the availability of a sufficient amount of objects with well-measured SFRs, to be used as the training/validation sample. We follow a similar approach and use exactly the same data in order to compare our results with those in Stensbo-Smidt et al. (2017). A parallel and independent ML approach was used in Bonjean et al. (2019) to solve the SFR regression problem with three main differences with respect to our approach: (1) they use shallow-IR instead of our optical features, (2) they employ a classical feature selection technique [embedded in their Random Forest (RF) model], and (3) they include spectroscopic information into the training parameter space. In particular, we investigate how effective ML-based methods can be in deriving SFRs in large samples of galaxies, paying special attention to *feature selection*, i.e. to the selection of the most suitable parameter space. As we shall demonstrate, the selection of the optimal set of features, in addition to a more accurate prediction, can also be used to derive an insight into the physics of the phenomenon (Brescia et al. 2017).

In Section 2, we introduce the data and in Section 3 all algorithms and ML methods used. In Section 4, we describe our campaign of experiments and related results. Finally, in Section 5, we discuss the results and draw some conclusions.

2 DATA

Since we were also interested in comparing our results with those presented in Stensbo-Smidt et al. (2017), the same data, derived from the Sloan Digital Sky Survey Data Release 7 (SDSS-DR7), have been used (Abazajian et al. 2009). Such data release has also been used by Brinchmann et al. (2004) to derive reliable SFRs for a subsample of $\sim 10^6$ galaxies, through a full analysis of the emission and absorption line spectroscopy, available in the SDSS spectroscopic data set (hence not based on the H_α flux alone). The reliability of this study was confirmed in Salim et al. (2007), who carried out an independent study using optical photometry from the SDSS and near UV measurements from GALEX, thus bypassing some uncertainties inherent the spectroscopic H_α aperture corrections. The local SFRs (normalized to $z = 0.1$) from the two studies (Brinchmann et al. 2004; Salim et al. 2007) turned out to agree within the errors.

The final catalogue contains several types of magnitudes[1]: *psfMag*, *fiberMag*, *petroMag*, *modelMag*, *expMag*, and *deVMag* in the *u*, *g*, *r*, *i*, and *z* bands; it includes also the *spectroscopic redshift* (z_{spec}), the *photometric redshift* (*photoz*), derived using an hybrid combination of a template fitting approach with an empirical calibration using objects with both observed colours and spectroscopic redshift (Csabai et al. 2007), as well as the *average specific star formation rate* (hereafter SFR). Starting from this data set, we performed a pre-processing, in which the following constrains were applied to improve the reliability of the final knowledge base:

(i) We required high-quality estimations of *SFR*, i.e. objects for which the quality flag is equal to 0 (see Brinchmann et al. 2004 for further details).

(ii) We required high-quality spectroscopic redshifts (i.e. with *zWarning* = 0; see Abazajian et al. 2009 for further details).

(iii) All objects affected by missing information, namely objects with at least one feature having a '*Null value*', were removed from the knowledge base, since our chosen ML methods are not capable of handling missing features.

The final knowledge base consists of 603 680 galaxies, respectively, 362 208 for training and 241 472 as blind test set, extracted through a random shuffling and split procedure. Furthermore, for each magnitude type we derived the related colours, i.e. $u - g$, $g - r$, $r - i$, and $i - z$, thus reaching a total of 56 features, 55 photometric (magnitudes, colours, and *photoz*) and one spectroscopic (z_{spec}). Finally, we added the SFR, used as target variable. The distribution of spectroscopic redshifts and SFRs for the knowledge base is shown in Fig. 1.

3 THE METHODS

In this work, we make use of two supervised ML methods: RF (Breiman 2001) and Multi-Layer Perceptron trained by the Quasi Newton Algorithm (MLPQNA; Brescia et al. 2012). Furthermore, in order to optimize their performances, we apply *k*-fold cross-validation (cf. Kohavi 1995) and a novel feature selection model called *Parameter handling investigation LABoratory* (ΦLAB; Brescia et al. 2018). These methods are shortly described in the following sections.

3.1 Random Forest

The RF (Breiman 2001) operates by generating an ensemble of decision trees during the training phase, based on different subsets of input data samples. For each decision tree, a random subset of input features is selected and used to build the tree. By imposing a sufficient number of trees (depending on the parameter space complexity and input data amount), all given features will, with high probability, be examined within the produced forest (Hastie, Tibshirani & Friedman 2009). In our experiments, we make use of the RF implementation from the PYTHON library scikit-learn (Pedregosa et al. 2011). For our purposes, we heuristically choose an ensemble of 1000 trees, trying to reach a good trade-off between performance and training computing time. Each tree was created by a random shuffling of the full set of features available and with a minimum split at each node equal to two.

3.2 MLPQNA

The MLPQNA is a model in which the learning rule is based on the Quasi Newton rule, one of the Newton's methods aimed at finding the stationary point of a function and based on an approximation of the Hessian of the training error through a cyclic gradient

[1] http://classic.sdss.org/dr7/algorithms/photometry.html

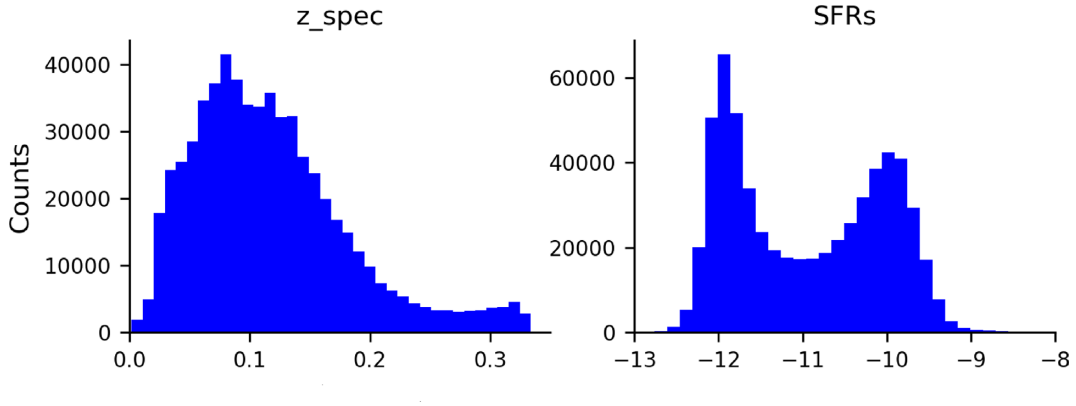

Figure 1. Spectroscopic redshift (left-hand panel) and SFR (right-hand panel) distributions of the knowledge base.

calculation. MLPQNA makes use of the known L-BFGS algorithm (Limited memory – Broyden Fletcher Goldfarb Shanno, Byrd, Nocedal & Schnabel 1994). Our multilayer perceptron architecture consists of two hidden layers with, respectively, $2N + 1$ and $N - 1$ neurons, where N is the number of input features. All further details of the MLPQNA implementation, as well as its performance in different astrophysical contexts, have been extensively discussed elsewhere (Brescia et al. 2012, 2014b; Cavuoti et al. 2013, 2015, 2017; Brescia, Cavuoti & Longo 2015; D'Isanto et al. 2016). With respect to the RF, our actual implementation of the MLPQNA model is generally more computationally intensive and thus some of the experiments performed later on in this paper are referred to the RF model only.

3.3 *K*-fold cross-validation

Within the context of the supervised ML paradigm, it is common practice to exploit the available knowledge base by deriving three disjoint subsets: one (training set) to be used for learning purposes, namely to acquire the hidden correlation among input features and the output target; a second (validation set) to check the training status, in particular, to measure the learning level and to verify the absence of any loss of generalization capabilities (a phenomenon also known as overfitting); and the third one, the test set is used to evaluate the overall performance of the trained and validated model. The latter two data sets are blind or, in other words, they do not contain input patterns already used during the training phase (Brescia et al. 2013).

In some cases, especially in presence of a limited amount of samples available within the knowledge base, a valid alternative approach, also applied in this work, is the so-called *k-fold* cross-validation technique (Kohavi 1995). This is an automatic cross-validation procedure, based on k different training sessions, specified as it follows: (*i*) random splitting of the training set into k random subsets, each one composed by the same fraction of the knowledge base; (*ii*) each of the k subsets is then, in turn, used as test set, while the remaining $k - 1$ subsets are used for training/validation.

The purpose of k-fold cross-validation is, in part, to test the model's performance stability on different subsets of the data, thus making sure that a chosen training/test set was neither particular favourable or unfavourable, and to minimize the risk of any training overfitting occurrence. In our case, we heuristically choose $k = 10$, representing a good compromise between computing efficiency and data amount within the folds.

3.4 Feature selection

Not all input features contain the same amount of information for a particular problem domain, and discovering the most informative variables may, on the one hand, drastically reduce the computing time and, on the other hand, it can provide useful insights into the physical nature of the problem. In this work, we used a novel feature selection method, called *Parameter handling investigation LABoratory* (ΦLAB; Brescia et al. 2018).

The choice of an optimal set of features is connected to the concept of *feature importance*, based on the measure of a feature's *relevance*. Formally, the importance of a feature is its percentage of informative contribution to a learning system.

We approach the feature selection task on two complexity levels: (a) the *minimal-optimal feature selection*, which consists of a selection of the smallest parameter space able to obtain the best learning performance; and (b) the *all-relevant feature selection*, able to extract the most complete parameter space, i.e. all features considered relevant for the solution to the problem. The second level is appropriate for problems with highly correlated features, as these features will contain nearly the same information. With a minimal-optimal feature selection, choosing any one of them (which could happen at random if they are perfectly correlated) means that the rest will never be selected.

We investigated the possibility to find a method able to optimize the parameter space, by solving the all-relevant feature selection problem, thus indirectly improving the physical knowledge about the problem domain. The method presented, ΦLAB, includes properties of both embedded and wrappers categories of feature selection (see Guyon & Elisseeff 2003 for an introduction to feature selection). The details of the method are presented in the Appendix A.

3.5 Evaluation metrics

In order to evaluate the performance of our experiments, we use the quantity $\Delta_{\rm SFR}$, defined as

$$\Delta_{\rm SFR} \equiv {\rm SFR}_{\rm photometric} - {\rm SFR}_{\rm spectroscopic},$$

where $\rm SFR_{photometric}$ is the estimated SFR, $\rm SFR_{spectroscopic}$ is the target value obtained from spectroscopy. We indicate also S_m as the blind test set. Then we use the following metrics:

(i) $\mathrm{RMSE} = \sqrt{\frac{1}{|S_m|}\sum_{n\in S_m}[\Delta_{\rm SFR}]^2}$, the root-mean-square error of the residuals.

Table 1. Performance comparison of the RF and MLPQNA models, calculated on the blind test set, using all the 54 photometric features available and the full training set.

Model	RMSE	Median	η
RF	0.252	−0.021	1.99
MLPQNA	0.261	−0.016	1.76

(ii) *Median* (Δ_{SFR}), the median of the residuals.

(iii) $\sigma = \sqrt{\frac{1}{|S_m-1|}\sum_{n\in S_m}[\Delta_{SFR} - \overline{\Delta_{SFR}}]^2}$, the standard deviation of the residuals.

(iv) η, the percentage of catastrophic outliers. According to the definition by Stensbo-Smidt et al. (2017), we consider an outlier to be catastrophic if $\Delta_{SFR} > 3\sigma$. Consequently, the percentage of outliers depends on the value of σ.

The RMSE and σ turned out to be almost identical in all of our experiments; the mathematical relation between the two estimators is: $\text{RMSE} = \sqrt{\left(\overline{\Delta_{SFR}}^2 + \sigma^2\right)}$. This means that the mean of Δ_{SFR} is negligible. We decided to report only the RMSE in each table. Nevertheless, we will use both estimators since the RMSE is used to evaluate the model performance, while the σ is used to compute the fraction of catastrophic outliers.

4 EXPERIMENTS AND RESULTS

In order to optimize the procedure in terms of SFR accuracy, we performed a series of experiments.

As first step we evaluated the performance of our regression models on the entire set of available features. Afterwards we evaluated the usefulness of the *k*-fold cross-validation, by verifying if such time-consuming operation (in our case it extends the training time of the network by almost a factor of ten) is effectively required to minimize overfitting and to check how the models perform on different data sets. In other words, how stable are the results across the whole data sets. Subsequently, we performed a feature selection to optimize the parameter space, indirectly suitable also for a comparison with the feature selection described in Stensbo-Smidt et al. (2017). Then we performed a series of experiments to evaluate the most appropriate size of the training set. After that we analysed the relationship between the photometric redshift quality and the accuracy of SFRs. Finally, we compared the SFR prediction performance between the methods RF and MLPQNA on the best set of features found by ΦLAB.

4.1 RF and MLPQNA performances on the full set of photometric features

As said above, we performed a preliminary performance test using the full set of available features (i.e. the 54 photometric features described in Section 2). The results are summarized in Table 1.

The results of Table 1 show that RF performs better than MLPQNA.

4.2 *k*-fold cross-validation

As a preliminary step of the training phase, and accordingly to what was done in Stensbo-Smidt et al. (2017), we decided to verify if the *k*-fold cross-validation technique is required to avoid overfitting in this particular use-case. Therefore, we replicated the RF and MLPQNA performance tests on the full set of 54 photometric features (see Section 4.1), but this time implementing the *k*-fold cross-validation using $k = 10$.

Table 2. Experiments result with and without *k*-fold cross-validation. The statistics are calculated on the blind test set only.

Model	Cross-validation			No cross-validation		
	RMSE	Median	η	RMSE	Median	η
RF	0.252	−0.021	1.99	0.252	−0.021	2.07
MLPQNA	0.261	−0.016	1.76	0.261	−0.016	1.78

Table 3. Effect of the cross-validation on the experiments of Table 2. Each column represents the standard deviation across 10 different experiments for a statistical estimator. In this case, our spectroscopic SFRs span in the range $\sim[-14, -17]$. This shows how, in this case, the cross-validation can be considered as negligible.

Model	σ_{RMSE}	σ_{Median}	σ_σ	σ_η
RF	0.001	0.00003	0.001	0.041
MLPQNA	0.002	0.00051	0.002	0.002

The results of the experiment can be seen in Table 2 where we compare the RF and MLPQNA performances with and without *k*-fold cross-validation. Experiments with cross-validation, while increasing the computing time by almost an order of magnitude, do not show any significant improvement in terms of accuracy. In Table 3, the standard deviations of the used statistical estimators computed over the ten folds are shown. As it can be seen, the results show that the cross-validation contribution is negligible, thus confirming that the information in the Knowledge Base is well distributed and, as a consequence, that both models are capable to work in a stable way across different data sets, as well as the fact that they are intrinsically robust against overfitting. For such reasons, we decided to perform all further experiments without the *k*-fold cross-validation technique.

4.3 Feature selection results

To perform the feature selection, we made use of our model ΦLAB using the full knowledge base available (see Section 2). The 34 features selected by the method are shown in Fig. 2 and listed in Table 4.

Concerning the excluded features, as it can be seen from Fig. 2, ΦLAB marks as 'unimportant' all the *z*- band magnitudes, five out of six *g*-band magnitudes (retaining only *fiberMag_g* but with a very low ranking), three out of six *u*-band magnitudes, four out of six *i*-band magnitudes and two in the *r* band. Conversely, all colours were retained with the exception of *devMag_u-g* and *petroMag_u-g*.

In particular, from Fig. 2, we can notice that all the exponential and de Vaucouleurs magnitudes are excluded (while their colours are retained) in favour of the *modelMag*.[2] For the other types of magnitudes only two or three are dropped (*i* and *z* for *fiberMag*, *g*, *z*, and *i* for *modelMag*, *u*, *g*, and *z* for the *petroMag* and *g*, *i*, and *z* for the *psfMag*). All together this leads to a total of 22 rejected features.

The optimized parameter space identified by ΦLAB (i.e. the 32 selected features of Fig. 2, excluding the two redshifts) was employed to perform a comparison between the two ML regression models used to estimate the SFR, starting from the same knowledge

[2] http://classic.sdss.org/dr7/algorithms/photometry.html

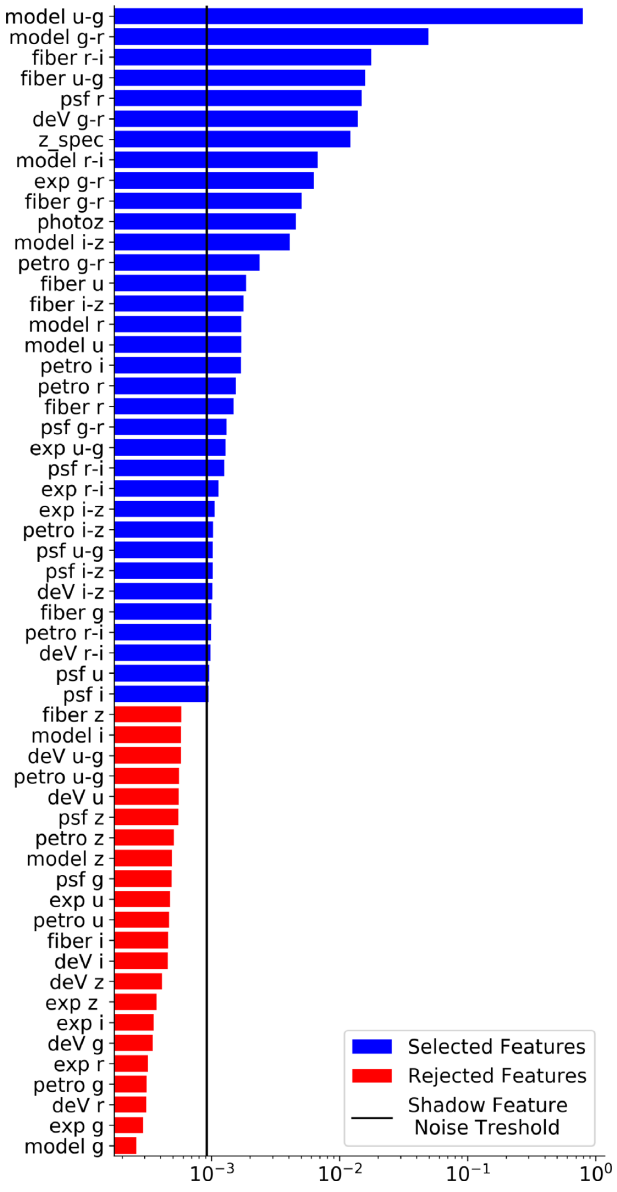

Figure 2. Feature importance percentages derived by applying the ΦLAB method to the full knowledge base and parameter space available, described in Section 2. In blue are marked the selected features, while in red those rejected by the method. The vertical black line is the noise threshold computed through the *shadow feature* technique embedded in the ΦLAB algorithm (see Section 3.4 for details). The noise threshold corresponds to an importance value of ∼0.062 per cent.

Table 4. List of features selected by ΦLAB running on the full knowledge base available.

Feature	model	fiber	psf	exp	petro	deV
$u-g$	✓	✓	✓	✓		
$g-$r	✓	✓	✓	✓	✓	✓
$r-$i	✓	✓	✓	✓	✓	✓
$i-$z	✓	✓	✓	✓	✓	✓
u	✓	✓	✓			
g		✓				
r	✓	✓	✓		✓	
i			✓		✓	
		z_{spec}			*photoz*	
redshift		✓			✓	

Table 5. Comparison between MLPQNA and RF models using the 32 photometric features identified by ΦLAB. Both models have been applied to the same training and blind test sets.

ID	RMSE	Median	η
$RF_{\Phi LAB}$	0.252	−0.021	2.03
$MLPQNA_{\Phi LAB}$	0.248	−0.017	1.99

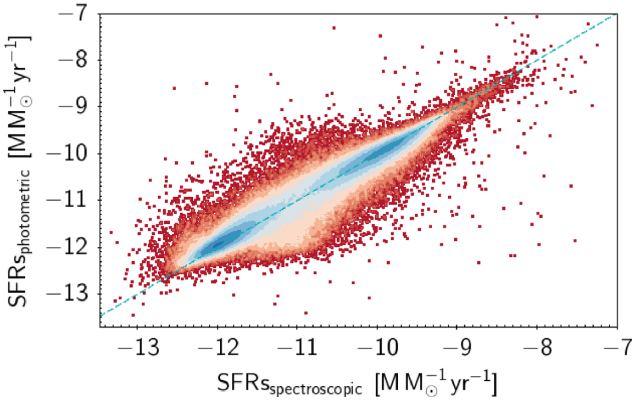

Figure 3. $SFRs_{spectroscopic}$ versus $SFRs_{photometric}$ scatter plot related to the $MLPQNA_{\Phi LAB}$ experiment, selected to produce the final SFR catalogue (see Appendix B).

base. Table 5 reports the results, while the distribution of photometric versus spectroscopic SFRs for MLPQNA is shown in Fig. 3. The MLPQNA obtains the best performance (∼1.5 per cent better accuracy than the RF on the same data). However, this comes at the cost of a much higher computational time, since using 32 features the RF takes ∼0.05 per cent of the computational time required by MLPQNA and this ratio further decreases for an increasing number of features. In spite of this, we decided to use the MLPQNA model to produce the SFRs catalogue presented in Appendix B.

In principle, a robust feature selection method should be able to identify the most relevant features in a way as independent as possible from the specific ML model used to subsequently approach the regression problem. Furthermore, in order to verify that the selected feature space is the best choice, a supplementary set of regression performance tests should be performed by using alternative subsets of features. In what follows we discuss these two aspects.

In order to verify the independence of the feature selection on the two regression methods, we iteratively trained the RF and MLPQNA, using always the entire training set, starting with just one feature and adding, at each iteration, a new feature (in the order of importance selected by ΦLAB). until all the 32 photometric features selected by ΦLAB were used. Fig. 4 shows the RMSE as function of the number of used features for both RF and MLPQNA methods. As it can be seen, the RMSE decreases steadily with the number of features in both cases, reaching the minimum value when the all 32 features are considered.

To further investigate the capability of the ΦLAB method to identify the optimal parameter space, we performed the following additional experiments with the RF:

(i) *RND*: We performed 10 experiments all using the same number of features (32) found by ΦLAB, but randomly extracted from the original parameter space (excluding the redshifts). These

experiments were performed in order to compare, fixed the number of features selected by ΦLAB, the performances achieved by the best all-relevant features experiment ($RF_{\Phi LAB}$ experiment) with those obtained via a random extraction.

(ii) *B+W (Best plus Worst)*: This experiment was performed in order to confirm the lack of relevance of the rejected features and also to investigate why the method rejected some features which at least should have conveyed relevant information. Therefore, we used the best 10 features selected by ΦLAB (excluding redshift) plus the 22 features rejected by ΦLAB, in order to maintain fixed to 32 the amount of used feature.

The results of these experiments are reported in Table 6. The experiment reaching the best performance is $RF_{\Phi LAB}$, thus confirming the reliability of the ΦLAB method in optimizing the parameter space by selecting the all-relevant subset of features best suited to solve the regression problem. Nevertheless the ΦLAB and B+W experiments show a very similar performance. Such behaviour seems to indicate that most of weak relevant and rejected features bring the same amount of contribution to solve the regression problem and that ΦLAB rejects those features considered as redundant. For the reasons already mentioned and related to the computational cost of MLPQNA, these experiments were performed using the RF only. Anyway, the fact that MLPQNA using the 32 features selected by ΦLAB ($MLPQNA_{\Phi LAB}$ experiment) achieves better performances than when the entire set of 54 photometric features is used (Table 1), indirectly confirms the reliability of the set of features selected by ΦLAB.

4.4 Completeness analysis of the training set

In order to investigate the complexity and completeness of the data set, we performed three experiments using, as training sets, the full data and two randomly extracted samples from the original training set, consisting of 36 000 and 100 000 objects, respectively. We used the 32 features selected by ΦLAB for these experiments. As shown in Table 7, the RF performance, always calculated on the same blind test set (241 472 objects), worsens less than that for MLPQNA with the shrinking of the training set size (see Table 8). Therefore, we use the full amount of data available in the training set in all further experiments.

4.5 Redshifts and analysis of dependence from photo-*z* accuracy

Looking at the feature importance ranking computed by ΦLAB in Fig. 2, as it could be expected, the spectroscopic redshifts (z_{spec}) carries crucial information to estimate the SFRs. Due to the intrinsic uncertainty carried by photometric redshifts, this feature (label *photoz*) has a lower rank (11th out of 56) and does not seem to carry any particular information contribution to boost the prediction performance. Even if the photoz does not improve the accuracy of the SFR estimation, the presence of both features within the parameter space selected by ΦLAB can be justified by considering that photoz is seen as a noisy version of the more accurate *z*_spec.

In order to evaluate the single contribution of both types of redshift, we performed a set of experiments, reported in Table 9, by imposing, respectively, a parameter space composed by all 54 photometric features available without any redshift (experiment PHOT), and the same parameter space in which we alternately added the *z*_spec (experiment ZSPEC) and *photoz* (experiment ZPHOT). As it can be seen by looking at the statistical results of Table 9, the inclusion of *z*_spec obtains, as expected, better performances, while the presence of *photoz* seems to be negligible in terms of prediction improvement. However, although the z_{spec} appears as a relevant feature, we dropped it from the used parameter space, since we were interested in predicting SFR via photometric information only.

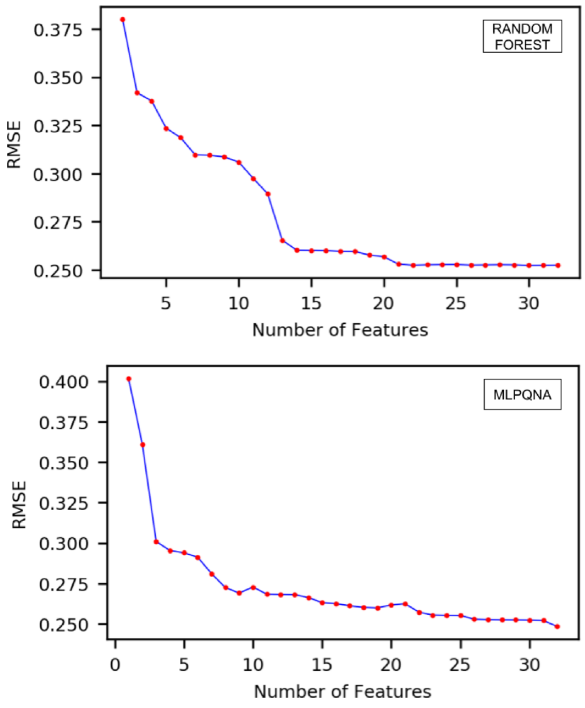

Figure 4. Performance variation of the RF (upper panel) and MLPQNA (lower panel) models with respect to the number of features used in the training. On the *y*-axis, we report the RMSE value computed on the blind test set, while on the *x*-axis the incremental number of features included in the training.

Table 6. Performance of the RF model, calculated on the blind test set, applied to different subsets of features. $RF_{\Phi LAB}$ uses the 32 features selected by ΦLAB, RND uses a set of 32 features randomly extracted from the original parameter space (best value over the 10 extractions), while B+W uses the best 10 features plus the 22 excluded by ΦLAB. In all such four parameter spaces, both spectroscopic and photometric redshifts were excluded.

ID	RMSE	Median	η
$RF_{\Phi LAB}$	0.252	−0.021	2.03
RND	0.269	−0.018	1.87
B+W	0.253	−0.022	2.03

Table 7. RF performance against training set size variation. As features we used the best 32 found by the ΦLAB method and as target the given SFRs. The statistics is calculated on the blind test set.

Number of training objects	RMSE	Median	η
36 000	0.278	−0.022	1.99
100 000	0.265	−0.022	1.97
362 208	0.252	−0.021	2.03

Table 8. MLPQNA performance against training set size variation. As features we used the best 32 found by the ΦLAB method and as target the given SFRs. The statistics is calculated on the blind test set.

Number of training objects	RMSE	Median	η
36 000	0.337	−0.015	1.53
100 000	0.281	−0.017	1.62
362 208	0.248	−0.017	1.99

Table 9. RF performance over the full set of features. The experiment named PHOT (which contains only magnitudes and colours) is performed using all the 54 photometric features (i.e. colours and magnitudes); ZSPEC and ZPHOT are two additional experiments, performed by adding to the M+C parameter space, respectively, the spectroscopic and photometric redshift.

ID	Features	RMSE	Median	η
PHOT	54	0.252	−0.021	1.99
ZSPEC	55	0.232	−0.018	2.00
ZPHOT	55	0.252	−0.021	2.18

Table 10. Prediction results of the RF model applied on the blind test set, obtained, respectively, on the parameter space selected by ΦLAB (ID label $RF_{\Phi LAB}$) and with the addition of the feature z_{spec} (i.e. spectroscopic redshifts, ID label $RF_{\Phi LAB} + z_{spec}$) or *photoz* (i.e. photometric redshifts, ID label $RF_{\Phi LAB} + z_{phot}$).

ID	Features	RMSE	Median	η
$RF_{\Phi LAB}$	32	0.252	−0.021	2.03
$RF_{\Phi LAB}+z_{spec}$	33	0.233	−0.017	2.24
$RF_{\Phi LAB}+z_{phot}$	33	0.252	−0.021	2.04

To further verify the feature selection made by ΦLAB, we repeated the experiments outlined in Table 9 only using the 32 all-relevant features selected by ΦLAB:

(i) $RF_{\Phi LAB}$: experiment using the features identified by ΦLAB (excluding both types of redshift).
(ii) $RF_{\Phi LAB} + z_{spec}$: experiment with the features identified by ΦLAB including the spectroscopic redshift.
(iii) $RF_{\Phi LAB} + z_{phot}$: experiment with the features identified by ΦLAB including the photometric redshift.

These experiments were performed only with the RF, by excluding MLPQNA due to the much longer training time of this model, assuming also a very similar effect of such additional features on both regression models, by considering our previous analysis done on the ΦLAB feature selection (see Section 3.4).

The results, summarized in Table 10, confirm that the spectroscopic redshifts bring a higher contribution than the photometric redshifts to estimate SFRs. However, since the two redshifts should in principle represent the same information, we expect that sufficiently accurate photometric redshifts could replace the spectroscopic information and that any residual prediction error would be dominated by other sources of noise. Therefore, to get an estimate of how accurate photometric redshifts need to be to obtain a SFR prediction with the same accuracy as reached by including spectroscopic redshifts, we decided to proceed through the following steps:

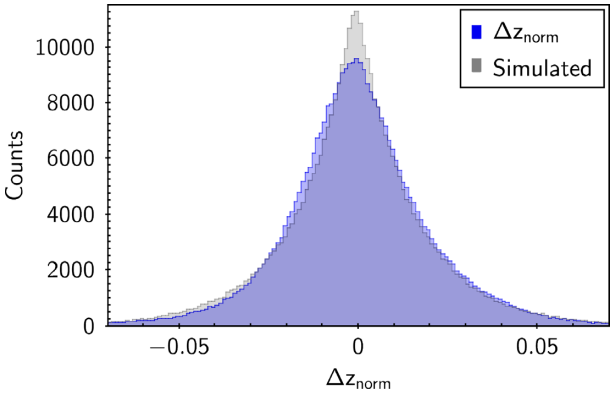

Figure 5. Distribution of redshift residuals Δz_{norm} (coloured in blue) with the superimposed best-fitting Laplacian distribution (coloured in grey).

Table 11. Photometric redshift accuracy estimation experiments. The first four experiments are referred to the SFR $RF_{\Phi LAB} + z_{phot}$ estimations varying the *photoz* measurement precision. While in the last one the photometric redshifts were replaced by spectroscopic redshifts.

Redshift used	RMSE	Median	η
$\sigma = 0.022$	0.249	−0.019	2.08
$\sigma = 0.015$	0.244	−0.019	2.11
$\sigma = 0.007$	0.238	−0.018	2.18
$\sigma = 0.005$	0.236	−0.018	2.21
$RF_{\Phi LAB} + z_{spec}$	0.233	−0.017	2.24

(i) identification of the distribution that fits the Δz_{norm} distribution, where $\Delta z_{norm} = (z_{spec} - photoz)/(1 + z_{spec})$;
(ii) Simulation of several Δz_{norm} distributions of the same shape, but with different accuracy;
(iii) Application of the different Δz_{norm} to the z_{spec} in order to simulate *photoz* with increasing accuracy;
(iv) Testing the SFR estimation using simulated *photoz*.

We started by calculating the Δz_{norm} distribution of the *photoz* used for the $RF_{\Phi LAB} + z_{phot}$ experiment, obtaining a distribution with a bias of −0.000 79 and a σ of 0.022. We then estimated through a Kolmogorov–Smirnov test (Oliphant 2007) the distribution that best fits the Δz_{norm} distribution of the *photoz*. We tried to fit the data with all the continuous distributions implemented in the *scipy.stats* module.[3] A Laplacian distribution with a standard deviation of 0.015 and a bias of 0.0077 was found to be the best fit (see Fig. 5). This distribution was then used to generate random noise that we added to the original z_{spec} distribution in order to simulate the *photoz*'s (and thus its measurement error). The process of noise generation and addition was repeated 10 times in order to compare the resulting SFR estimation statistics and to make sure that the correlation between the simulated error and the corresponding statistics was consistent. Afterwards, we repeated the $RF_{\Phi LAB} + z_{phot}$ experiment using this new photometric redshift distributions finding an average RMSE variation along the 10 extractions of ∼0.001.

As reported in the first row of Table 11, the statistical performance is very similar to the $RF_{\Phi LAB} + z_{phot}$ experiment (Table 10), thus proving that our simulation is able to reproduce the behaviour of photometric redshifts (the slight difference in performance may be

[3] https://docs.scipy.org/doc/scipy/reference/stats.html

due to the presence of systematic errors, ignored by the simulation). We then proceeded to an iterative decrease of the σ of the Laplacian distribution, in order to simulate an increasing quality of *photoz* estimations; at each step we repeated (ten times) the $\mathrm{RF}_{\Phi\mathrm{LAB}} + z_{\mathrm{phot}}$ experiment with the new distribution of *photoz*. The results are reported in Table 11 and show that, in order to obtain an efficiency comparable with the one obtained using the spectroscopic redshifts, an accuracy of at least $\sigma = 0.005$ is required for the *photoz* estimation. We want to underline that this is simply an indication of the photometric redshift accuracy required to become indistinguishable from the SFR prediction accuracy reached with spectroscopic redshifts. This standard deviation value is lower than what can be found in literature; see for instance Brescia et al. 2014b ($\sigma = 0.028$ in the range $0 < z_{\mathrm{spec}} \leq 1$) or Laurino et al. 2011 ($\sigma = 0.015$ in the range $0 < z_{\mathrm{spec}} \leq 0.65$) or Ball et al. 2008 ($\sigma = 0.021$ in the range $0 < z_{\mathrm{spec}} \leq 0.5$), motivated by the smallest redshift range considered in this particular case ($0 < z_{\mathrm{spec}} \leq 0.33$).

4.6 Catastrophic outliers

As already mentioned, due to the higher accuracy, we decided to use the MLPQNA model to create our SFRs catalogue (see Section 4.3), so in order to detect possible issues with the model and gain insights into the nature of the physical problem, we analysed the nature of the catastrophic outliers (i.e. those objects whose SFR prediction error resulted higher than 3σ) distribution relative to the $\mathrm{MLPQNA}_{\Phi\mathrm{LAB}}$ experiment. In Fig. 6, it is shown the distribution of catastrophic outliers in the $\mathrm{SFRs}_{s\mathrm{pectroscopic}}$ versus $\mathrm{SFRs}_{\mathrm{photometric}}$ space, resulting from the $\mathrm{MLPQNA}_{\Phi\mathrm{LAB}}$ experiment reported in Table 6. We estimated the pixel density through a kernel density estimation method (Scott 1992) and coloured the pixels on the basis of their density. As shown in the scatter plot of Fig. 6(a), most of the point are clustered in a small region (highlighted in yellow) hereafter called the *overdensity region* (that is confirmed also using the RF results). In order to understand why these objects are outliers, we selected all the objects belonging to the overdensity region through cuts in their local density. The scatter plots of Figs 6(b) and (c) show highlighted in orange all the objects with a density, respectively, six and eight times higher than the average point density. Depending on the cuts, the overdensity region contains 1877 objects (six times the average density) or 1277 objects (eight times the average density) out of the total number of 4840 objects classified as catastrophic outliers. We then investigated the possibility that these objects could form a cluster in some bi-dimensional projections of the parameter space. We tried all the possible magnitudes, colours, and redshifts combinations without finding any obvious clustering (some of these combinations are shown in Appendix C). We also checked whether the group could correlate with a specific (high) error measure associated with any of the used features, but no any evident correlations were found. The nature of the objects in the overdensity region is still under further investigation.

4.7 Comparison with a recent work

In order to compare our regression models with the k-NN used by Stensbo-Smidt et al. (2017) and their feature selection, we performed an experiment using the full training set and the set of 8 features found by Stensbo-Smidt et al. (2017). In Table 12, we present the statistical results, which show a comparable performance among the three methods, although with a lower RMSE obtained by RF and MLPQNA. Using the features found by ΦLAB, the RF and MLPQNA can achieve even better performance, as shown in Table 6. This is not surprising as k-NN is much more sensitive to the dimensionality of the parameter space (the so-called curse of dimensionality) than other two models. These latter can, therefore, take advantage of the information carried by a larger number of features than a k-NN model.

5 DISCUSSION AND CONCLUSIONS

In this work, based on our preliminary analysis of the problem presented at the ESANN-2018 conference (Delli Veneri et al. 2018), we estimated star formation rates for a large subset of the SDSS-DR7 and produced a catalogue of SFRs derived using photometric features only (magnitudes and colours) and the MLPQNA ML model (see Appendix B) trained on a knowledge base of spectroscopically determined SFRs. By looking at Fig. 3 and the statistics in Table 6, the regression results appear very promising. This is particularly true, considering that the dynamical range of SFR is between -12 and -7, and also that we have $\sim$5000 outliers out of the 242 000 objects of the blind test set and, finally, taking into account the low percentage of outliers ($\sim$2 per cent). However, from the results obtained by varying the size of the model training set (Tables 7 and 8), we think that a larger knowledge base of SFRs would further improve the performances.

Furthermore, the residual scatter is likely to be an artefact of the photometry. The fig. 5(a) in Stensbo-Smidt et al. (2017) shows a scatter plot for the predictions obtained with SED fitting. It appears qualitatively similar to the current work and could suggest that there is a more fundamental limit to the accuracy we can expect from optical photometry only. This is not an obvious issue; for example in Brescia et al. (2014b) it was demonstrated, in the case of estimation of photometric redshifts, that the model performance, over a certain amount of data, does not scale with the size of the training set.

By considering the median estimator, in all our experiments its values are always negative. This is a consequence of the presence of the overdensity described in Section 4.6 and shown in Fig. 4. We intend to perform a deeper investigation on such objects, which will focus on the characterization of objects in the overdensity region in terms of their spectroscopic, morphological, and evolutionary properties.

By applying the ΦLAB method, we found the all-relevant set of features and were able to discard almost half of the initial set of features without any loss in precision over the full set, but with a great gain in computing time. We tested the ΦLAB method several times, confirming the reliability of its feature selection.

Since in future surveys, it is likely that no large spectroscopic samples will be available, we run a simulation to find the minimum accuracy required for photometric redshifts in order to effectively replace spectroscopic estimates, finding that SFRs can be predicted with the same accuracy under the condition to provide photo-z with an error smaller than 0.005 (see Table 11).

From our results on the SDSS-DR7, we think that our ML methods could be applied to other surveys to reliably calculate SFRs. On this note, we intend to expand our photometric knowledge base to the UV, X-ray, and infrared in order to

(i) use the full spectrum to identify and constrain outliers and potential issues in the methods (i.e. AGN selection through X-ray photometry);
(ii) incorporate the UV and infrared information to derive SFRs.

Moreover, we intend to apply our methods to derive photometric SFRs from the ESO-KiDS-DR4 (Kuijken et al., in preparation). We

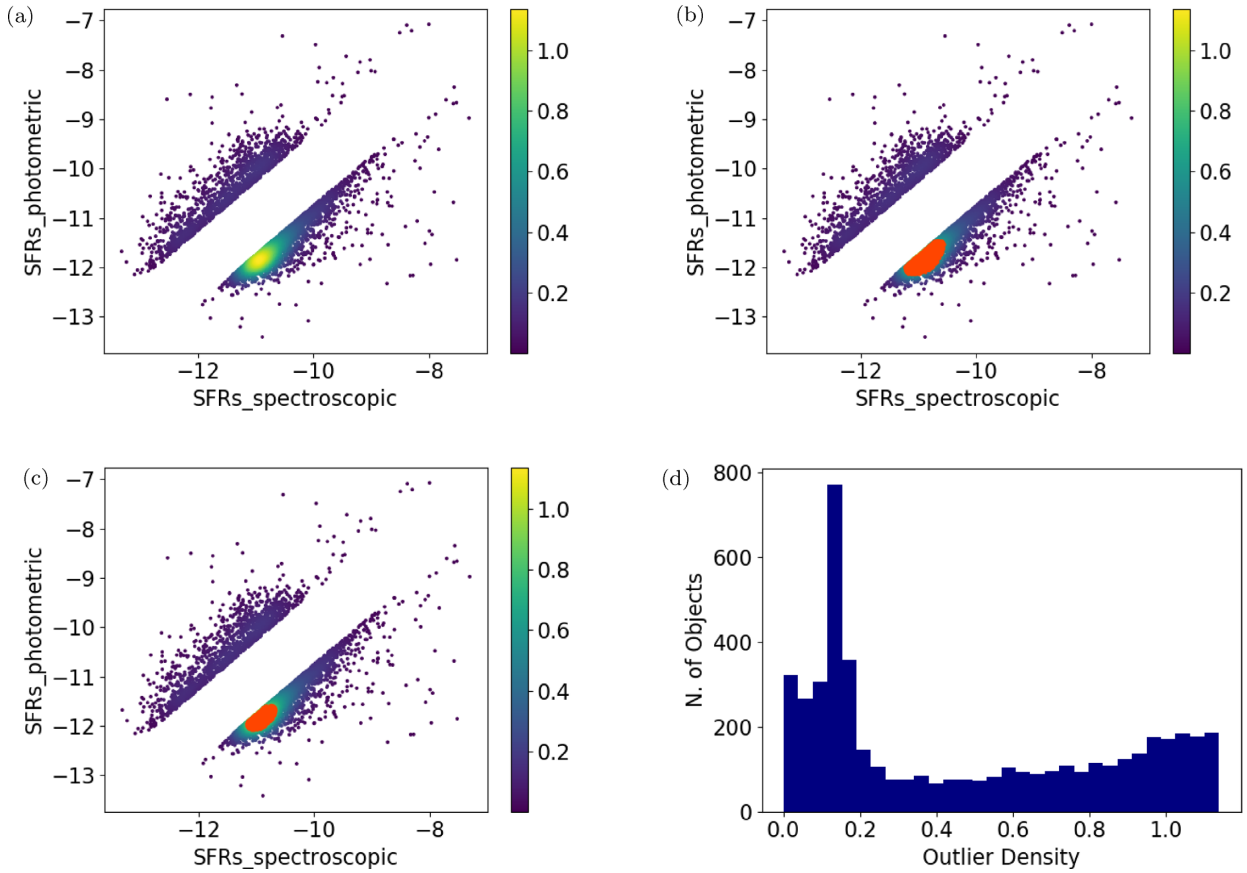

Figure 6. The scatter plot in the top left corner (a) shows the distribution of outliers in the $SFRs_{spectroscopic}$ versus $SFRs_{photometric}$ space with a superimposed density map, while the diagrams int the top right (b) and bottom left (c) corners show highlighted in orange all the objects with a density, respectively, six and eight times higher than the average point density. The histogram in the bottom right corner (d) shows the outliers density distribution.

Table 12. Comparison between our RF and MLPQNA against (Stensbo-Smidt et al. 2017) k-NN using the full train set and the best eight features found by Stensbo–Smidt.

Model	RMSE	Median	η
RF	0.264	−0.020	1.86
k-NN	0.274	0.013	1.85
MLPQNA	0.265	−0.021	1.85

wish to conclude by saying that the natural evolution of this work will be to expand our knowledge base above $z_{spec} = 0.33$. In this case, on one hand, redshifts would have a bigger impact on galaxy emission and thus magnitudes; on the other hand, we should be able to produce high-quality SFRs for larger samples of objects.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referee for extremely valuable comments and suggestions. A special thank goes to Kristoffer Stensbo–Smidt for all the work and the precious suggestions he put in reviewing and commenting this work. MB acknowledges the *INAF PRIN-SKA 2017 program 1.05.01.88.04* and the funding from *MIUR Premiale 2016: MITIC*. GL and MDV acknowledge the UE funded Marie Curie *ITN SUNDIAL* that partially supported this work. SC acknowledges support from the project '*Quasars at high redshift: physics and Cosmology*' financed by the ASI/INAF agreement 2017-14-H.0. Topcat has been used for this work (Taylor 2005). C^3 has been used for catalogue cross-matching (Riccio et al. 2017). DAMEWARE has been used for ML experiments (Brescia et al. 2014a).

REFERENCES

Abazajian K. et al., 2009, ApJS, 182, 543
Ball N. M., Brunner R. J., Myers A. D., Strand N. E., Alberts S. L., Tcheng D., 2008, ApJ, 683, 12
Bonjean V., Aghanim N., Salomé P., Beelen A., Douspis M., Soubrié E., 2019, A&A, 622, A137
Breiman L., 2001, Mach. Learn., 45, 5
Brescia M. et al., 2014a, PASP, 126, 783
Brescia M., Cavuoti S., Paolillo M., Longo G., Puzia T., 2012, MNRAS, 421, 1155
Brescia M., Cavuoti S., D'Abrusco R., Longo G., Mercurio A., 2013, ApJ, 772, 140
Brescia M., Cavuoti S., Longo G., De Stefano V., 2014b, A&A, 568, A126
Brescia M., Cavuoti S., Longo G., 2015, MNRAS, 450, 3893
Brescia M., Cavuoti S., Amaro V., Riccio G., Angora G., Vellucci C., Longo G., 2017, Data Analytics and Management in Data Intensive Domains. XIX International Conference, DAMDID/RCDL 2017, Moscow, Russia
Brescia M. et al., 2018, MNRAS, submitted
Brinchmann J., Charlot S., White S., Tremonti C., Kauffmann G., Heckman T., Brinkmann J., 2004, MNRAS, 351, 1151

Byrd R., Nocedal J., Schnabel R., 1994, Math. Program., 63, 129
Calzetti D. et al., 2007, ApJ, 666, 870
Calzetti D., Harris J., Gallagher J. S. III, Smith D. A., Conselice C. J., Homeier N., Kewley L., 2004, AJ, 127, 1405
Cardamone C. N. et al., 2010, ApJS, 189, 270
Cavuoti S., Brescia M., D'Abrusco R., Longo G., Paolillo M., 2013, MNRAS, 437, 968
Cavuoti S., Brescia M., De Stefano V., Longo G., 2015, Exp. Astron., 39, 45
Cavuoti S., Amaro V., Brescia M., Vellucci C., Tortora C., Longo G., 2017, MNRAS, 465, 1959
Condon J. J., 1992, ARA&A, 30, 575
Conroy C., 2013, ARA&A, 51, 393
Cooke K. C., Fogarty K., Kartaltepe J. S., Moustakas J., Odea C. P., Postman M., 2018, ApJ, 857, 122
Csabai I., Dobos L. M. T., Herczegh G., Józsa P., Purger N., Budavári T., Szalay A. S., 2007, Astron. Nachr., 328, 852
D'Isanto A., Cavuoti S., Brescia M., Donalek C., Longo G., Riccio G., Djorgovski S., 2016, MNRAS, 457, 3119
Delli Veneri M., Cavuoti S., Brescia M., Riccio G., Longo G., 2018, preprint (arXiv:1805.06338)
Fogarty K., Postman M., Larson R., Donahue M., Moustakas J., 2017, ApJ, 846, 103
Guyon I., Elisseeff A., 2003, J. Mach. Learn. Res., 3, 1157
Hara S., Maehara T., 2017a, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. AAAI press, San Francisco, CA, p. 1985
Hara S., Maehara T., 2017b, 31st AAAI Conference on Artificial Intelligence. AAAI press, California, USA, p. 1985
Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edn. Springer, New York, USA
Hong S. et al., 2011, ApJ, 731, 45
Kennicutt R. C. Jr, 1998, ARA&A, 36, 189
Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531
Kohavi R., 1995, Proceedings of the 14th International Joint Conference on Artificial Intelligence – Vol. 2. IJCAI'95. Morgan Kaufmann Publishers Inc., San Francisco, p. 1137
Kursa M., Rudnicki W., 2010, J. Stat. Softw., 36, 1
Laurino O., D'Abrusco R., Longo G., Riccio G., 2011, MNRAS, 418, 2165
Madau P., Dickinson M., 2014, ARA&A, 52, 415
Marchesi S. et al., 2016, ApJ, 817, 34
Matute I. et al., 2012, A&A, 542, A20
Oliphant T. E., 2007, Comput. Sci. Eng., 9, 10
Pearson W. J. et al., 2018, A&A, 615, A146
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Rafelski M., Gardner J. P., Fumagalli M., Neeleman M., Teplitz H. I., Grogin N., Koekemoer A. M., Scarlata C., 2016, ApJ, 825, 87
Riccio G., Brescia M., Cavuoti S., Mercurio A., di Giorgio A. M., Molinari S., 2017, PASP, 129, 024005
Salim S. et al., 2007, ApJS, 173, 267
Salvato M. et al., 2009, ApJ, 690, 1250
Salvato M. et al., 2011, ApJ, 742, 61
Schleicher D. R. G., Beck R., 2013, A&A, 556, A142
Scott D., 1992, Multivariate Density Estimation: Theory, Practice, and Visualization. A Wiley-Interscience Publication, Wiley
Stensbo-Smidt K., Gieseke F., Igel C., Zirm A., Steenstrup Pedersen K., 2017, MNRAS, 464, 2577
Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, Astronomical Data Analysis Software and Systems XIV. Astron. Soc. Pac., San Francisco, p. 29
Tibshirani R. J., 2013, Electron. J. Statist., 7, 1456
Wuyts S. et al., 2011, ApJ, 742, 96

APPENDIX A: ΦLAB METHOD

ΦLAB is based on the combination of two components: *shadow features* and *Naïve LASSO* statistics. The term *shadow features* arises from the idea to extend the given parameter space with artificial features (Kursa & Rudnicki 2010). Given a data set of N samples, represented through a D-dimensional parameter space, we introduce a shadow feature for each real one, by randomly shuffling its values among the N samples, thus doubling the original parameter space. Shadow features are, thus, random versions of the real ones and their importance percentage can be used as a threshold for when a real feature is containing actual information. Such a threshold is important since feature selection methods only provide a ranking of the features, never an absolute important/not important decision. The second component of ΦLAB is based on the *Naïve LASSO* statistics. The LASSO (Least Absolute Shrinkage and Selection, Tibshirani 2013) performs both a variable selection and a regularization of a ridge regression (i.e. a shrinking of large regression coefficients to avoid overfitting), enhancing the prediction accuracy of the statistical model. The regularization is a typical process exploited within ML, based on the addition of a functional term to a loss function (e.g. a penalty term). LASSO performs the so-called L_1 regularization (i.e. based on the standard L_1 norm), which has the effect of *sparsifying* the weights of the features, effectively turning off the least informative features. In particular, we included two *Naïve LASSO* techniques in ΦLAB. One is the A-LASSO (Alternate-LASSO; Hara & Maehara 2017a), able to find all weakly relevant features that could be removed from the standard LASSO solution. Such method calculates a list of features alternate to those selected by the standard LASSO, each one associated with a calculated score, reflecting the performance degradation from the optimal solution. In ΦLAB, we select only the alternate features that achieve the lowest score difference from the best features, trying to reach the best trade-off between feature selection performance and flexibility in the analysis of the parameter space. Such alternate features smoothly degrade the solution score, but may potentially infer more flexibility, by relaxing the intrinsic stiffness of the best solution system. The second version of the standard LASSO is E-LASSO (Enumerate-LASSO; Hara & Maehara 2017b), which enumerates a series of different feature subsets, considered as solutions with a decreasing level of approximation. The main concept behind is that an optimal solution to a mathematical model is not necessarily the best solution to any physical problem. Therefore, by enumerating a variety of potential solutions, there is a chance to obtain better solutions for the problem domain task. For instance, Hara and Maehara demonstrate that E-LASSO solutions are good approximations to the optimal solution, by also improving the flexibility for the selection of the parameter space, covering a wide spectrum of variations within the problem domain (i.e. by helping to find the all-relevant set of features). The shadow features and *Naïve LASSO* are then combined by selecting the candidate weak relevant features through the shadow feature noise threshold and by extracting the final set of weak relevant features through a filtering process, based on the A-LASSO and confirmed by E-LASSO. To summarize, we find the list of candidate features through the shadow features technique and then we use the LASSO operator to explore the parameter space and verify the effective contribution carried by those features considered as weak relevant to the solution of the problem. The loss function based on L1 regularization is crucial to quantify the degradation of performance when other features subsets are replacing the best one, by also automatically identifying the exact redundancy of some features that the shadow features technique is unable to disentangle in terms of individual importance.

The pseudo-code of the features selection method can be summarized by the following steps (see also Fig. A1):

(i) Let the set $\{x_1, x_2, ..., x_D\}$ be the initial complete parameter space composed by D real features;
(ii) Apply the shadow feature selection (SFS method) and produce the following items:

– $SF = x_{s_1} ... x_{s_D}$, the list of shadow features, obtained by randomly shuffling the values of real features;
– max(IMP[parameter space, SF]) $\forall x \in$ parameter space & $\forall x_s \in SF$, the importance list of all 2D features, original and shadows.
– st: noise threshold, defined as the max$\{$IMP[SF], $\forall x_s \in$ SF$\}$.
– BR $= \{x \in$ parameter space with IMP[x] $\geq$ st$\}$, the set of best relevant real features;
– RF $= \{x \in$ parameter space, rejected by the SFS$\}$, the set of excluded real features, i.e. not relevant;
– WR $= \{x \in$ parameter space with IMP[x] $<$ st$\}$, the set of weak relevant real features.

(iii) At this stage, the complete parameter space is now split into *BR*, *WR,* and *RF*. Now we consider the reduced parameter space, space$_{red}$ = {BR + WR}, obtained by excluding the rejected features. In principle, it may correspond to the original parameter space if there is no rejections by the SFS:

(A) If RF==$\emptyset$ && WR$==\emptyset$, the SFS method confirmed all real features as high relevant, therefore return ALL-RELEVANT (parameter space), i.e. the full parameter space as the optimized parameter space and EXIT.
(B) If RF$\neq\emptyset$ && WR$==\emptyset$, the SFS method rejected some features and confirmed others as high relevant, therefore return ALL-RELEVANT (BR) as the optimized parameter space and EXIT.
(C) If WR$\neq\emptyset$, regardless some rejections, SFS confirmed the presence of some weak relevant features that must be evaluated by LASSO methods, therefore go to step (iv).

(iv) Apply E-LASSO method on the space$_{\rm red}$ = {BR + WR} producing:

– EL_S: a list of M subsets of features, considered as possible solutions, ordered by decreasing score;
(A) If WR $\subseteq$ EL_S, then all weak relevant features are possible solutions, therefore return ALL-RELEVANT(BR + WR) as the optimised parameter space and EXIT.
(B) Else go to step (v);

(v) Apply A-LASSO method on the space$_{red}$ = {BR + WR}(set of candidate features) producing:

– AL_S, a set of T features, each one with a corresponding list of features List(t) considered as alternate solutions with a certain score;
(A) If AL_S $==\emptyset$, then no alternate solutions exist, therefore:

(A.1) If EL_S$==\emptyset$, then return ALL-RELEVANT(BR) as the optimized parameter space and EXIT.

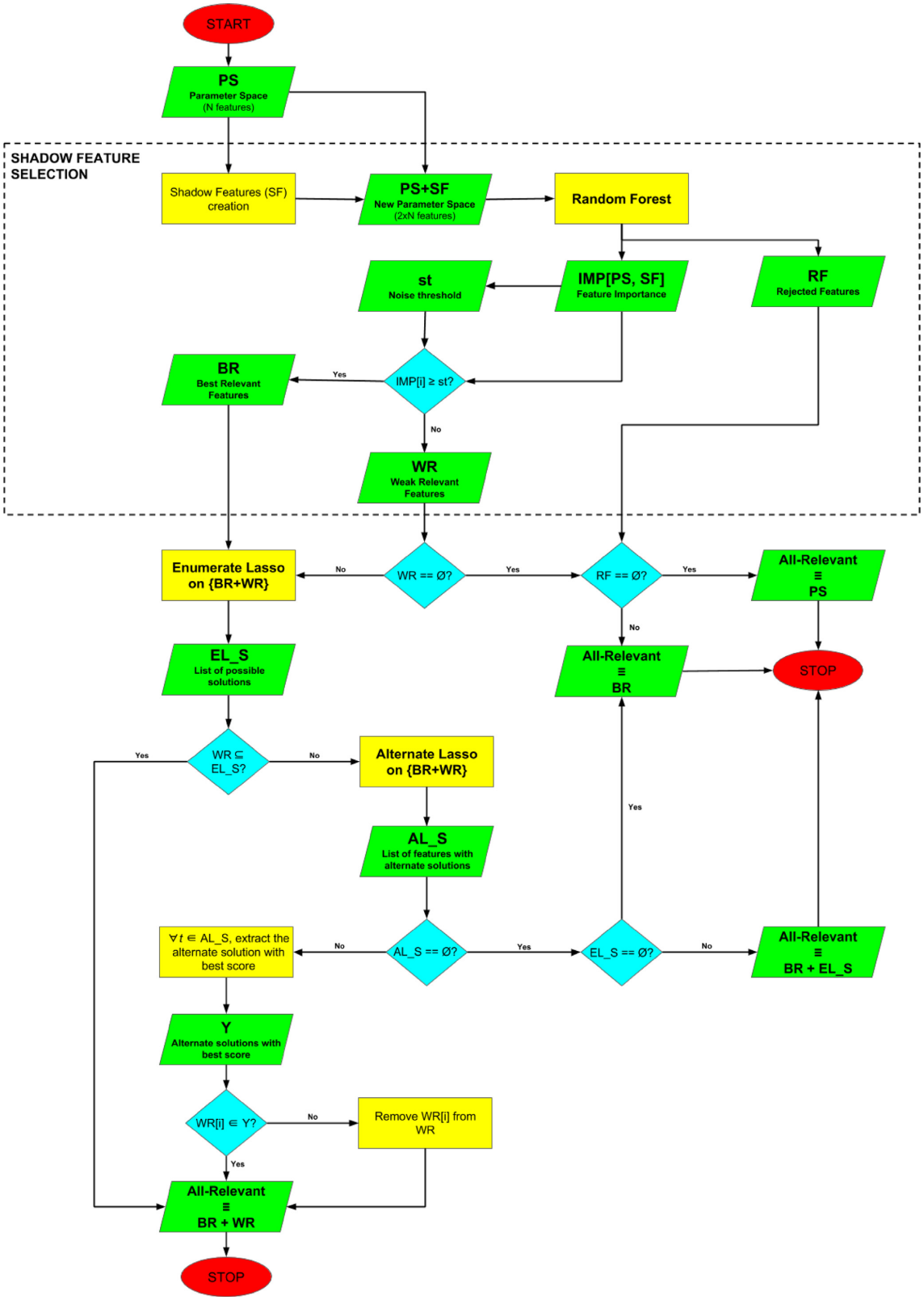

Figure A1. ΦLAB workflow.

(A.2) Else if EL_S$\neq \emptyset$, then return ALL-RELEVANT(BR + EL_S) as the optimized parameter space and EXIT.

(B) Else extract $\forall t \in T$ the alternate solution with Score(as) $=$ min{Score(y), $\forall y \in$ List(t)};
(C) Go to step (vi).

(vi) For each $x \in$WR:

(A) If x is alternate solution of at least one feature $t \in T$, with [t $\in$ BR || t $\in$ EL_S], then retain x within WR set;
(B) Else reject x (by removing x from WR);

(vii) Return ALL-RELEVANT(BR + WR) as the final optimized parameter space and EXIT.

APPENDIX B: CATALOGUE

We produced an SFR catalogue containing SFRs for 27 513 324 galaxies of the SDSS-DR7, which will be available on the Vizier facility. The catalogue is actually accessible at http://dame.na.astro.it/sfr/Catalogue.csv. To produce the catalogue, we started by querying the *Galaxy* View[4] of the SDSS-DR7 for all the needed photometric features of galaxies with a 'good' photometry (see PhotoFlags) and containing no *Missing Values*. We then applied the magnitudes cuts of our knowledge base (in order to keep the photometric features within the ranges of our knowledge base) and cross-matched the resulted data set with the *photoz* catalogue derived by Brescia et al. (2014b), in order to use them as a quality flag. The final catalogue contains the following columns:

(i) Identifiers: *dr9objid*, *objid*, *ra*, *dec*, i.e. respectively, the object identifier in the SDSS DR9 and DR7 and their ascension and declination coordinates;
(ii) Quality flags: *photoz* and *Quality_Flag*, i.e. the *photometric redshifts* measured by Brescia et al. (2014b) and the associated flag. The *Quality_Flag* can assume three values 1, 2, and 3; 1 stands for the best photo-z accuracy, 2 and 3 for decreasing accuracy;
(iii) SFR: It is computed by the MLPQNA model with the 32 best features selected by the ϕLAB method (excluding redshifts).

In order to select only SFRs with high-quality (i.e. only select sources inside the training set parameter space constrains), the user should impose $photoz \leq 0.33$ and $Quality\_Flag = 1$. This is due by considering that in our knowledge base there are only objects with spectroscopic redshift less than 0.33, thus we are able to predict SFRs only for objects within such redshift range. These constraints will select $\sim$6.6 million objects. Since we do not have any spectroscopic redshifts for the catalogue objects, we must use photometric redshifts (where available) to perform these cuts. Nevertheless using photometric redshifts instead of spectroscopic ones may introduce some contamination in the catalogue, i.e. a source may be inside the $photoz \leq 0.33$ cut when in reality it has a spectroscopic redshift higher than 0.33. To estimate the number of such contaminants, we verify that among the 871 784 objects with $photoz \leq 0.33$ and a spectroscopic redshift only $\sim$1.33 per cent resulted to have a true redshift higher that 0.33.

APPENDIX C: BI-DIMENSIONAL PROJECTIONS TO ISOLATE THE OVERDENSITY REGION

In this section, we show some examples of bi-dimensional projections in the parameter space, among the most relevant features, done in order to isolate the objects in the overdensity region. As stated in Section 4.6, no projections were found able to achieve such separation. See Fig. C1.

[4]http://skyserver.sdss.org/dr7/en/help/browser/browser.asp?n=Galaxy&t = U

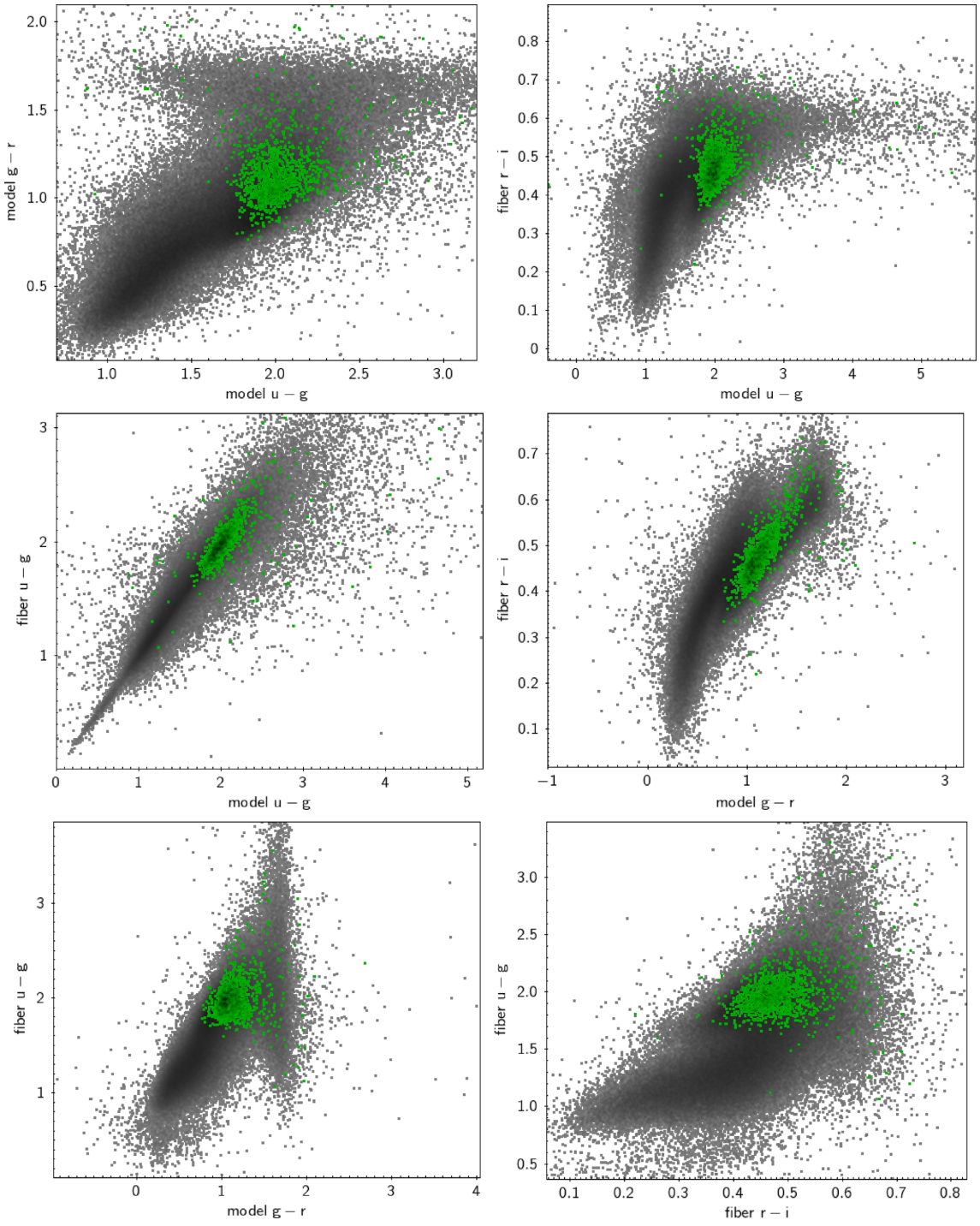

Figure C1. Some examples of bi-dimensional projections of the parameter space, done in order to isolate the objects of the overdensity region shown in Fig. 6. In particular, all the combinations of the most relevant colours are shown. The objects belonging to the overdensity region are highlighted in green colour.

APPENDIX D: EXAMPLE OF QUERIES USED TO OBTAIN GALAXIES FROM THE SDSS-DR7

```
SELECT
   p.objid, p.ra, p.dec,
   p.modelMag_u, p.modelMag_g, p.modelMag_r, p.modelMag_i, p.modelMag_z,
   p.devMag_u, p.devMag_g, p.devMag_r, p.devMag_i, p.devMag_z,
   p.expMag_u, p.expMag_g, p.expMag_r, p.expMag_i, p.expMag_z,
   p.petroMag_u, p.petroMag_g, p.petroMag_r, p.petroMag_i, p.petroMag_z,
   p.fiberMag_u, p.fiberMag_g, p.fiberMag_r, p.fiberMag_i, p.fiberMag_z,
   p.psfMag_u, p.psfMag_g, p.psfMag_r, p.psfMag_i, p.psfMag_z,
   q.objid as dr9objid
  INTO
   mydb.p75p90
  FROM
   Galaxy as p,
   dr9.PhotoPrimaryDR7 as s,
   dr9.Galaxy as q
  WHERE
   p.mode  =  1 AND
   p.dec > =  75 AND p.dec < 90 AND
   s.dr7objid  =  p.objid AND
   s.dr8objid  =  q.objid AND
   p.modelMag_u > -9999 AND p.modelMag_g > -9999 AND
   p.modelMag_r > -9999 AND p.modelMag_i > -9999 AND
   p.modelMag_z > -9999 AND p.devMag_u > -9999 AND
   p.devMag_g > -9999 AND p.devMag_r > -9999 AND
   p.devMag_i > -9999 AND p.devMag_z > -9999 AND
   p.expMag_u > -9999 AND p.expMag_g > -9999 AND
   p.expMag_r > -9999 AND p.expMag_i > -9999 AND
   p.expMag_z > -9999 AND p.petroMag_u > -9999 AND
   p.petroMag_g > -9999 AND p.petroMag_r > -9999 AND
   p.petroMag_i > -9999 AND p.petroMag_z > -9999 AND
   p.fiberMag_u > -9999 AND p.fiberMag_g > -9999 AND
   p.fiberMag_r > -9999 AND p.fiberMag_i > -9999 AND
   p.fiberMag_z > -9999 AND p.psfMag_u > -9999 AND
   p.psfMag_g > -9999 AND p.psfMag_r > -9999 AND
   p.psfMag_i > -9999 AND p.psfMag_z > -9999 AND
   dbo.fPhotoFlags('PEAKCENTER') ! =  0 AND
   dbo.fPhotoFlags('NOTCHECKED') ! =  0 AND
   dbo.fPhotoFlags('DEBLEND_NOPEAK') ! =  0 AND
   dbo.fPhotoFlags('PSF_FLUX_INTERP') ! =  0 AND
   dbo.fPhotoFlags('BAD_COUNTS_ERROR') ! =  0 AND
   dbo.fPhotoFlags('INTERP_CENTER') ! =  0
```

This paper has been typeset from a TEX/LATEX file prepared by the author.