astro-ph seminar 2021.08.26 (大澤)

Improving the reliability of photometric redshift with machine learning

Razim et al. accepted in MNRAS

ABSTRACT

In order to answer the open questions of modern cosmology and galaxy evolution theory, robust algorithms for calculating photometric redshifts (photo-z) for very large samples of galaxies are needed. Correct estimation of the various photo-z algorithms' performance requires attention to both the performance metrics and the data used for the estimation. In this work, we use the supervised machine learning algorithm MLPQNA to calculate photometric redshifts for the galaxies in the COSMOS2015 catalogue and the unsupervised Self-Organizing Maps (SOM) to determine the reliability of the resulting estimates. We find that for $z_{\rm spec} < 1.2$, MLPQNA photo-z predictions are on the same level of quality as SED fitting photo-z. We show that the SOM successfully detects unreliable $z_{\rm spec}$ that cause biases in the estimation of the photo-z algorithms' performance. Additionally, we use SOM to select the objects with reliable photo-z predictions. Our cleaning procedures allow to extract the subset of objects for which the quality of the final photo-z catalogs is improved by a factor of two, compared to the overall statistics.

Background

測光データから redshift を求める (photo-z)

- spectrum template を用いる (SED) template SED の知識に依存
- SED → z の interpolation function を構成 (ML) (SED, spec-z) のペアが十分な数必要.

ML では 2 種類の error source が存在
interpolation function の不定性が大きい, reference としている spec-z の値が怪しい (1~5%)

⇒ Self-Organizing Map (SOM) を用いて上記 2 つの不定性を評価する.

Data

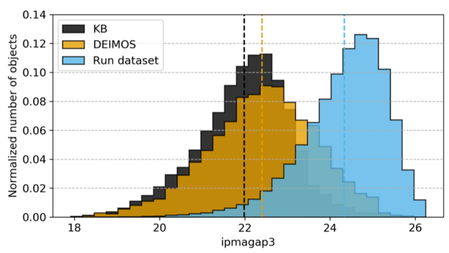


Figure 1. Normalized magnitude distributions for the KB, DEIMOS and run datasets in the `ipmagap3` after standard cleaning but before any SOM filtering. Dashed lines show the mean value of the distribution.

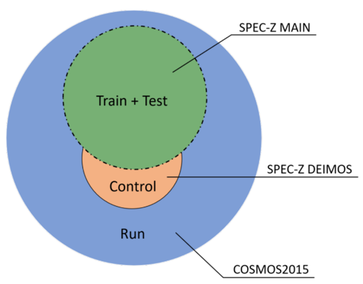


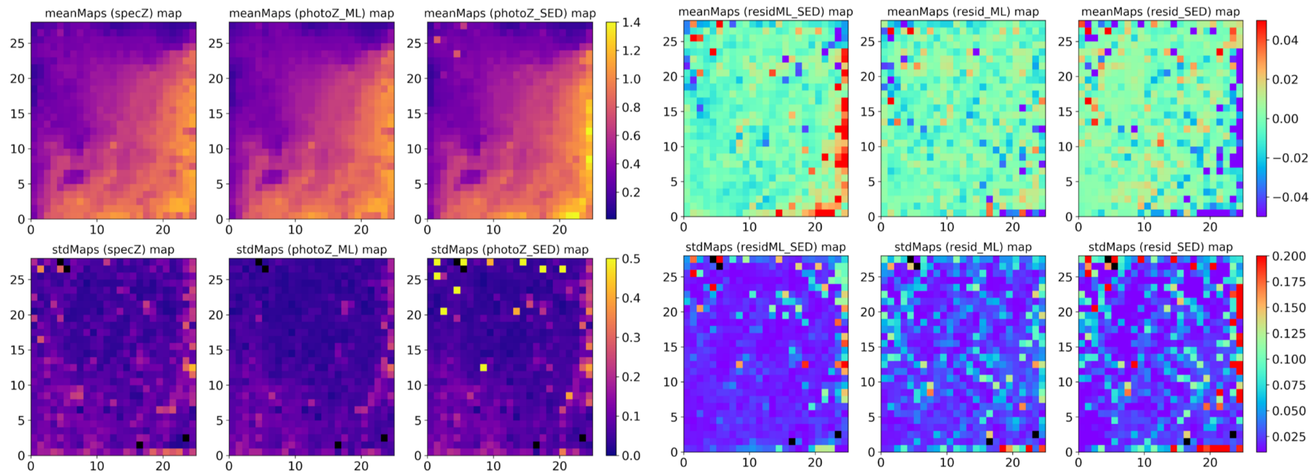
Figure 2. Venn diagram of the catalogues used in this work.

Dataset	Size	spec-z range	Median spec-z	`imagap3` range	Median `imagap3`
KB	19893	[0.020; 1.2]	0.57	[17.91; 26.19]	22.02
DEIMOS	2288	[0.025; 5.68]	0.89	[18.48; 25.90]	22.93
Run	228361	NA	NA	[18.14; 26.25]	24.49

Exp ID	Description	bands	Mean	$\sigma_{\Delta z/(1+z)}$	NMAD	$\eta_{0.15}$
exp001	SDSS-like, aper2	u,B,r,ip,zpp	0.000	0.053	0.025	1.999
exp002	SDSS-like, aper3	u,B,r,ip,zpp	-0.001	0.051	0.027	1.807
exp003	SDSS-like, aper_auto	u,B,r,ip,zpp	-0.003	0.055	0.027	2.364
exp004	SDSS-like, aper_ISO	u,B,r,ip,zpp	-0.004	0.059	0.028	2.346
exp005	broad bands, aper2 and aper3	u,B,r,ip,zpp	-0.001	0.051	0.025	1.756
exp006	broad bands, aper3	B,H,J,Ks,V,Y,ip,r,u,zpp	-0.002	0.048	0.018	2.138
exp007	broad bands + one narrow band, aper3	B,H,J,Ks,V,Y,ip,r,u,zpp,IB574	-0.002	0.048	0.019	1.642
exp008	all bands, aper3	Ks, Y, H, J, B, V, ip, r, u, zp, zpp, IA484, IA527, IA624, IA679, IA738, IA767, IB427, IB464, IB505, IB574, IB709, IB827, NB711, NB816, Hw, Ksw, yH	-0.002	0.049	0.017	1.877

Table 4. Results of the experiments performed with the trained MLPQNA on the blind test set of the KB, after having just applied the standard KB cleaning procedure.

Spec-z filtering by SOM



SOM: 多次元データをデータの類似性を元にして 2次元マップに圧縮する.

→ 各 cell の中で統計量を計算することで怪しい spec-z のデータを抽出できる.

training data を元にして SOM を作成.
各 cell で outlier を検出して remove した.
(ML の再学習はしていない? 記述がない)

		ML photo-z				SED photo-z			
Filtering	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
No filtering	5967	0.048	0.019	-0.0023	1.64	0.094	0.011	-0.0041	2.23
$\lvert K_{\rm spec}\rvert \le 1$	4311	0.025	0.017	0.0002	0.19	0.052	0.010	0.0006	0.70
$\lvert K_{\rm ML}\rvert \le 1$	4071	0.045	0.018	-0.0020	1.28	0.077	0.010	-0.0026	1.65
$\lvert K_{\rm SED}\rvert \le 1$	4133	0.043	0.018	-0.0020	1.06	0.061	0.010	-0.0021	1.43
trainMapOccupation > 5	5167	0.041	0.018	-0.0021	1.18	0.058	0.010	0.0001	1.49
$K_{\rm spec}$ + trainMapOccupation	3761	0.022	0.016	-0.0002	0.05	0.05	0.009	0.0018	0.35
$K_{\rm ML}$ + trainMapOccupation	3587	0.039	0.017	-0.0015	1.06	0.062	0.010	-0.0003	1.17
$K_{\rm SED}$ + trainMapOccupation	3624	0.038	0.017	-0.0016	0.94	0.038	0.010	0.0002	0.99

Table 7. Statistics for ML and SED fitting photo-z calculated for the test dataset after different types of filtering. The upper part of the table presents the statistics calculated for the dataset without spec-z and photo-z outliers. The lower part reports the effects of photometric filtering with occupation map.

Occupation Map

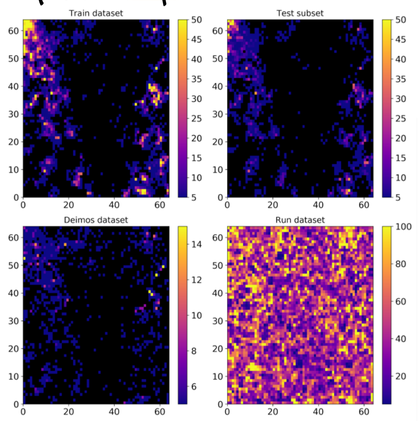


Figure 5. Occupation maps for all datasets projected on SOM trained with run dataset.

今回は "run" データをベースとして SOM を構成.
スペクトルの測定があるデータが run SOM において どこに分布するかを見ている (occupation map).

→ occupation map の値が小さい所では.
ML の精度が低いだろう, という推測に基づく.

		ML photo-z				SED photo-z			
Filtering	Num objects	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$	$\sigma_{\Delta z}$	NMAD	Mean	$\eta_{0.15}$
No filtering	2255	0.099	0.032	0.0347	10.86	0.142	0.014	-0.0082	5.06
$\lvert K_{\rm spec}\rvert \le 1$	1075	0.035	0.020	0.0018	1.02	0.127	0.011	-0.0103	2.88
$\lvert K_{\rm ML}\rvert \le 1$	1209	0.095	0.033	0.0392	12.57	0.090	0.013	-0.0031	4.55
$\lvert K_{\rm SED}\rvert \le 1$	1183	0.085	0.029	0.0245	8.96	0.078	0.013	0.0005	3.30
trainMapOccupation>5	1382	0.058	0.023	0.0127	2.10	0.059 ?	0.012	0.0085	2.68 ?
$K_{\rm spec}$ + trainMapOccupation	758	0.025	0.018	0.0017	0.13	0.031	0.010	0.0038	0.92
$K_{\rm ML}$ + trainMapOccupation	724	0.064	0.022	0.0136	1.93	0.060	0.011	0.0080	2.62
$K_{\rm SED}$ + trainMapOccupation	741	0.046	0.020	0.0063	1.48	0.044	0.011	0.0075	1.89

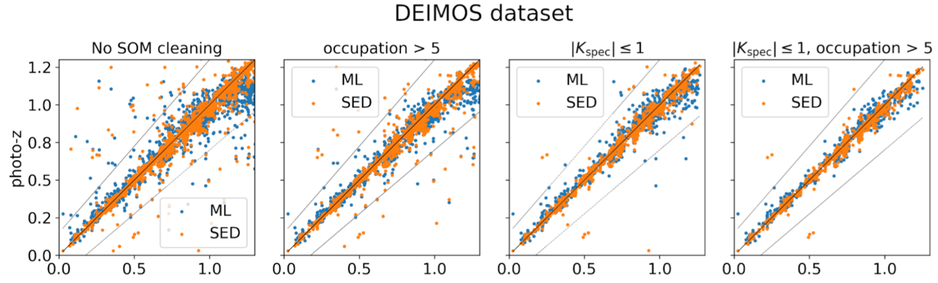



Table 8. Statistics for ML and SED fitting photo-z for the DEIMOS dataset after different types of filterings. Upper part of the table describes statistics after spec-z and photo-z outlier removal, while the lower part reports the effects of photometry filtering with occupation map.