Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Improving photo-z estimation under covariate shift with StratLearn

Roberto Trotta, Chiara Moretti (SISSA)
Max Autenrieth, David van Dyk (Imperial)
David Stenning (Simon Fraser U.)
Riccardo Serra (U Trieste)

Spoke 3 Technical Workshop, Trieste October 9 / 11, 2023

Scientific Rationale: Beating Malmquist bias with Machine Learning

SISSA DATA SCIENCE Machine Learning for the Natural Sciences

- **Selection effects** are ubiquitous in astronomy: e.g., brighter objects are more likely to be observed (Malmquist 1922)
- The problem with (supervised) machine learning:
- If the training set is systematically different from the test set because of Malquist bias or other 'selection effects', **generalization will be poor.**
- Our general-purpose, principled solution: **StratLearn**

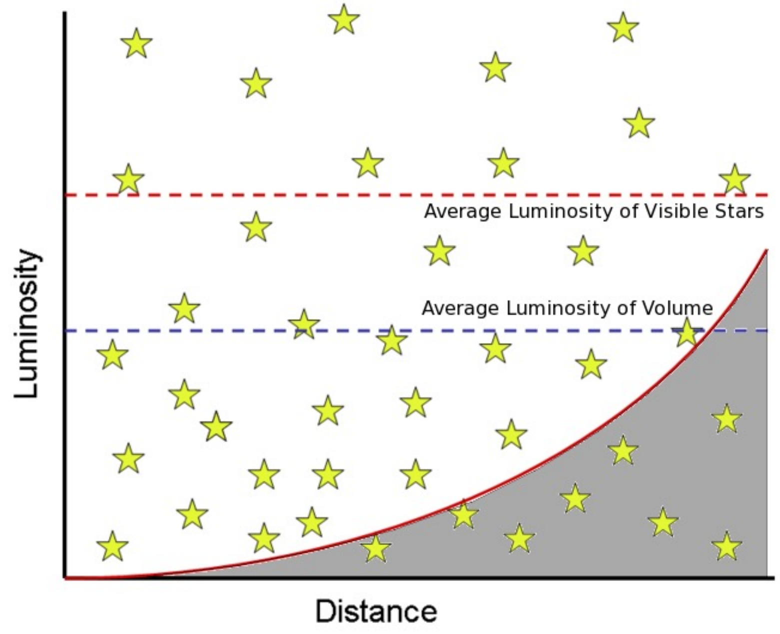


Image: Wikimedia

Covariate shift

Given a feature space, X, and a label space, $Y\,(K > 1$ classes/dependent variables)

n_s labelled samples $\{x_i^s, y_i^s\}$ from the source (s) domain

n_t unlabelled samples from the target (t) domain, $\{x_i^t\}$.

Task: predict $\{y_i^t\}$

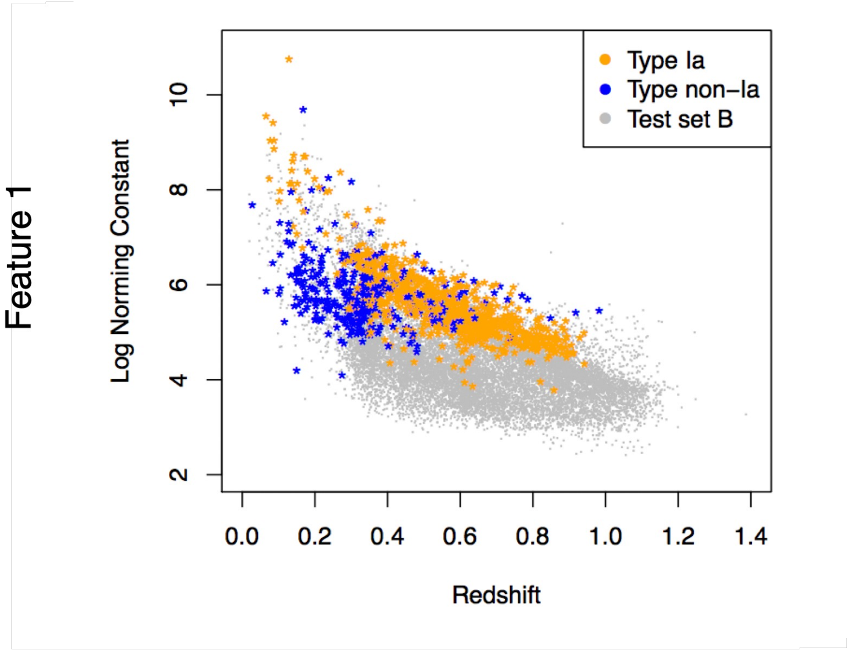
Features: redshift & apparent mag
Label: Ia or non-Ia

Covariate shift occurs when:

$$p_s(y\,|\,x) = p_t(y\,|\,x)$$

and $p_s(x) \neq p_t(x)$

I.e., the training set is non-representative of the test set.



Revsbech, RT, van Dyk (2018)

:e domain
ured dots)

et domain
y dots)

Technical Objectives, Methodologies and Solutions

"STACCATO": Revsbech, RT, van Dyk (2018): StratLearn, Autenrieth et al (2023)

Propensity scores

e(xᵢ) = probability for object *i* to be selected into the source domain, using the *whole features* set:

$$e(x_i) \equiv P(s_i = 1 \,|\, x_s, x_t)$$

Key idea ("StratLearn"):

subdivide ("stratify") target and source data in *k* subgroups according to quantiles of their propensity scores. Then supervised learning in each stratum ("stratified learner")

Propensity scores as balancing scores

Rosenbaum & Rubin (1983, 1984) show that, conditional on their propensity scores, the *k* subgroups ("strata") have approximately balanced covariate distribution, i.e.

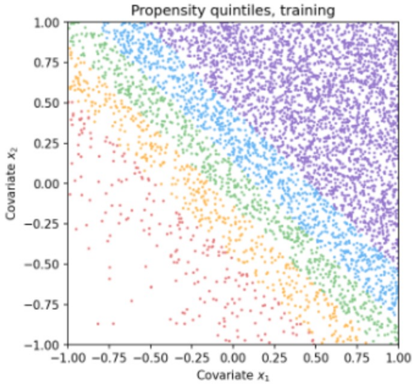
$$p_{s_j}(x) \approx p_{t_j}(x) \text{ for } j = 1,\ldots,k$$

Since $p_s(y\,|\,x) = p_t(y\,|\,x)$, it follows that

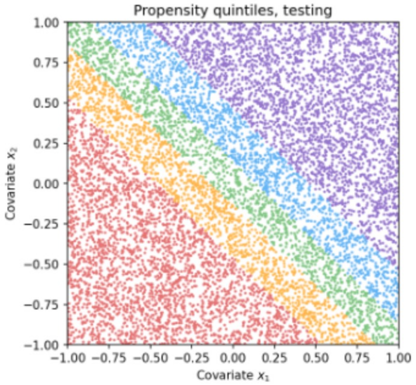
$$p_{s_j}(x, y) \approx p_{t_j}(x, y) \text{ for } j = 1,\ldots,k$$

Hence covariate shift approximately disappears.

Toy example: classification in 2D



Training

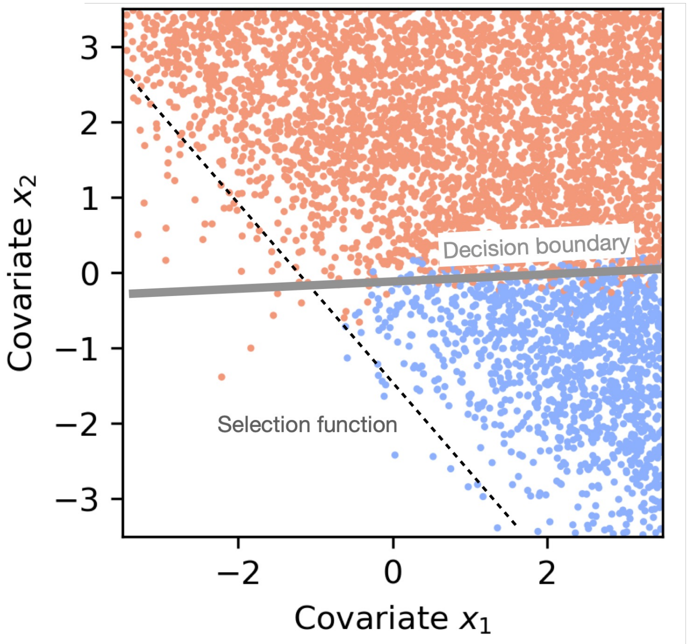


Test

Partitioning in propensity score quintiles groups ("strata")

Learning happens within each stratum separately

Training



Test

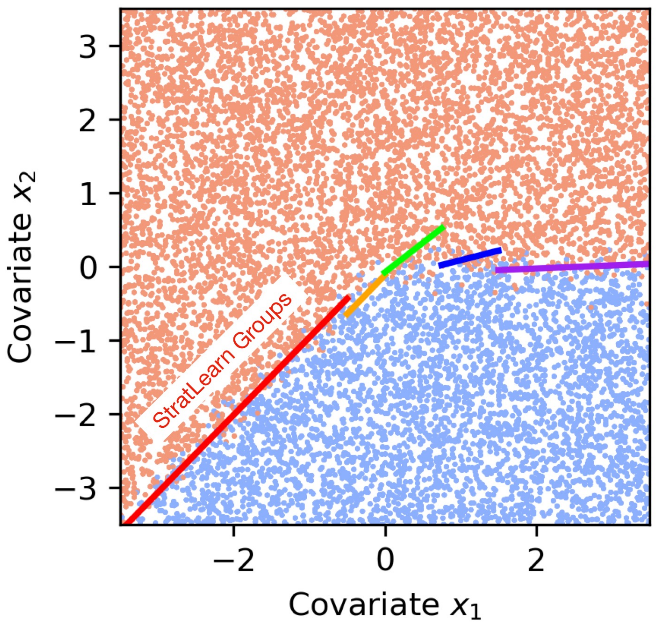


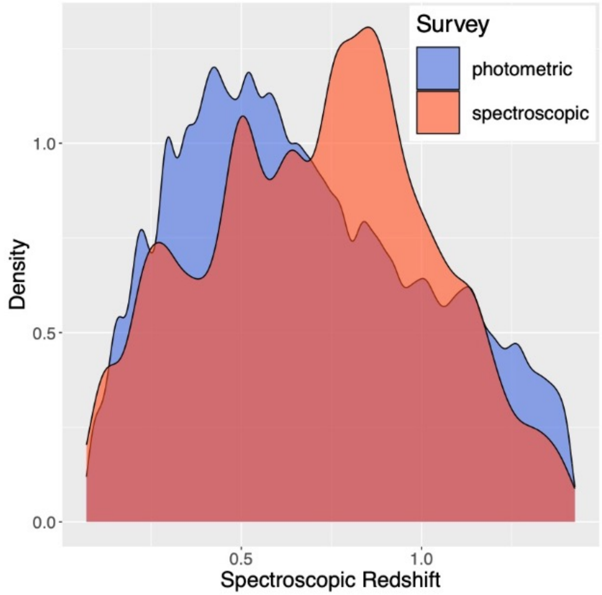
Figure: Riccardo Serra

Accomplished Work

MICE data (simulation study, Wright et al. (2020)):

Photometric data:

- ~12 Million photometric target samples (from shear-measurement weighted photometric distribution)
- 9 (noisy) magnitudes, bands u,g,r,i,Z,Y,J,H,KS



Aim: Assign photometric target objects to tomographic bins, and obtain unbiased mean redshift within each bin.

Table 1: Tomographic redshift bins.

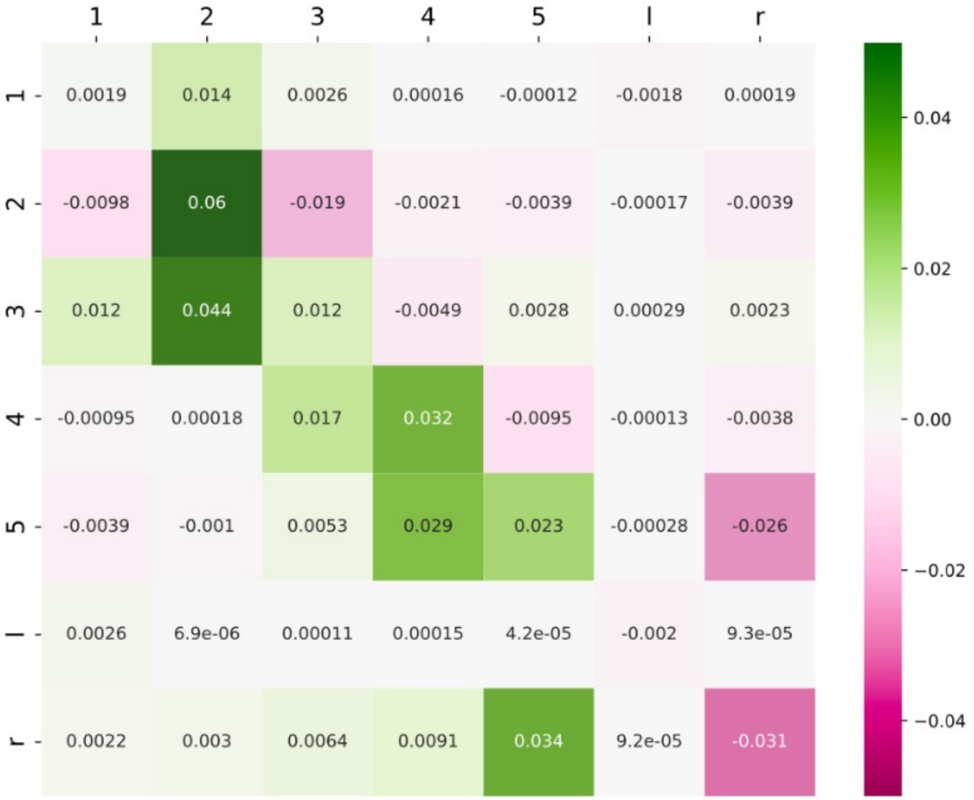
	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Redshift range	(0.1,0.3]	(0.3,0.5]	(0.5,0.7]	(0.7,0.9]	(0.9,1.2]

Slide credit: Max Autenrieth

Results

Tomographic bin assignment improved on all metrics wrt to the standard approach

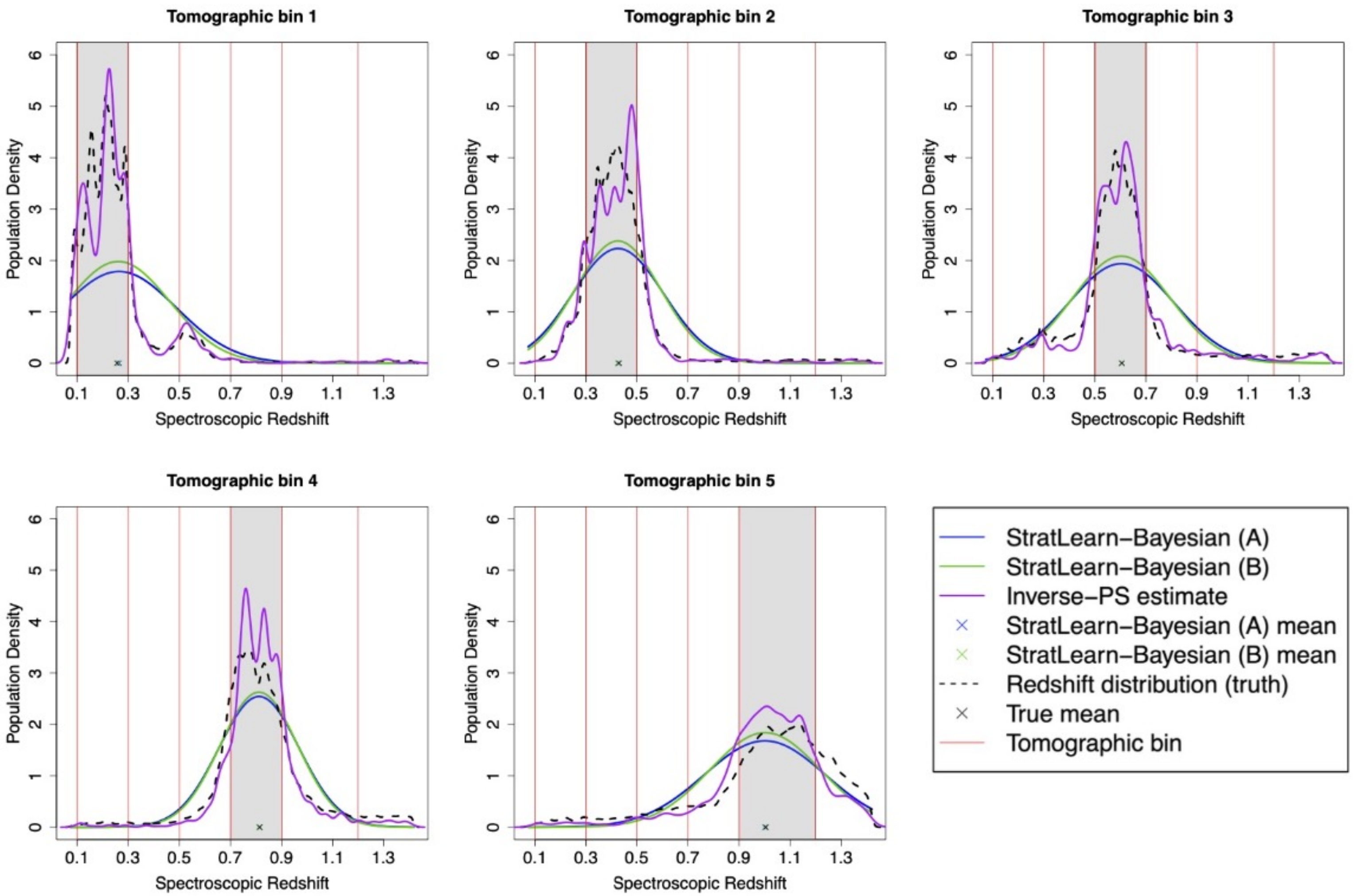
	StratLearn	z_B
Performance metric	mean (sd)	mean (sd)
Accuracy	0.622 (0.003)	0.526 (-)
Balanced Accuracy	0.718 (0.003)	0.706 (-)
Sensitivity	0.502 (0.006)	0.493 (-)
Specificity	0.934 (0.001)	0.918 (-)
Kappa	0.439 (0.006)	0.415 (-)



Slide credit: Max Autenrieth

Results

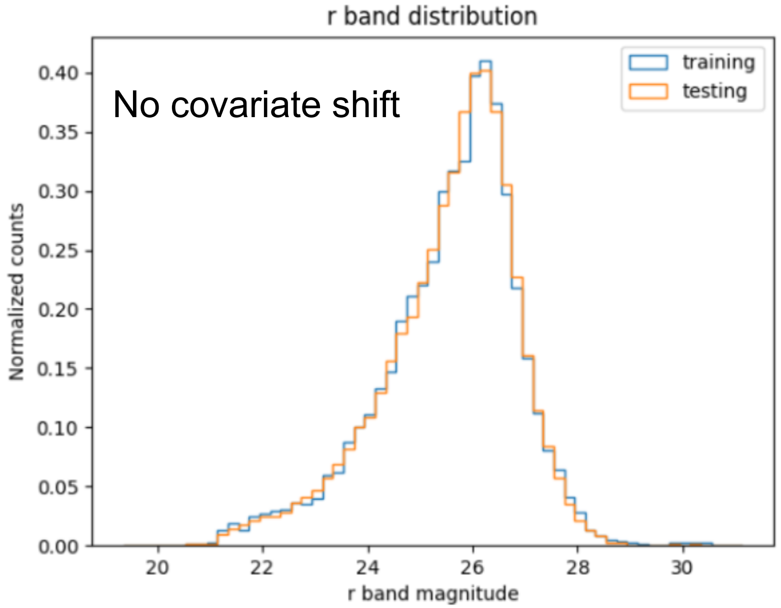
Estimation of the binned redshift population density via Inverse-Propensity score weighting – for use in downstream weak lensing analysis

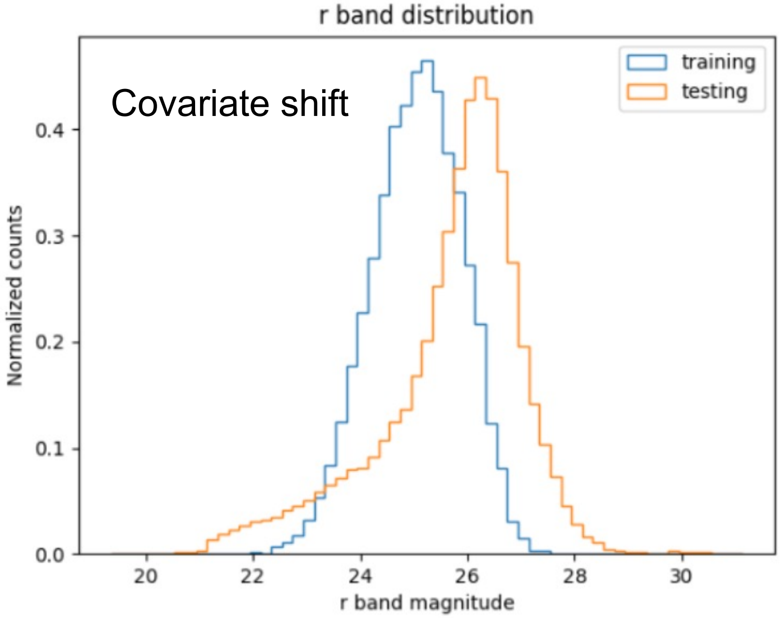


Ongoing work

With Riccardo Serra (Master student, UniTS), Chiara Moretti (postdoc):
Conditional density estimation of photo-z under realistic simulated data and covariate shift scenario.
Eventual target: Application to Euclid data.

Current work: 100,000 objects sampled from the Buzzard Flock catalogue (DeRose et al., 2019), a synthetic catalogue generated by adding galaxies to a dark matter N-body simulations and imposing a realistic set of observational properties and systematics (Stylianou et al., 2022); similar to what can be expected for LSST (*ugrizy* bands)



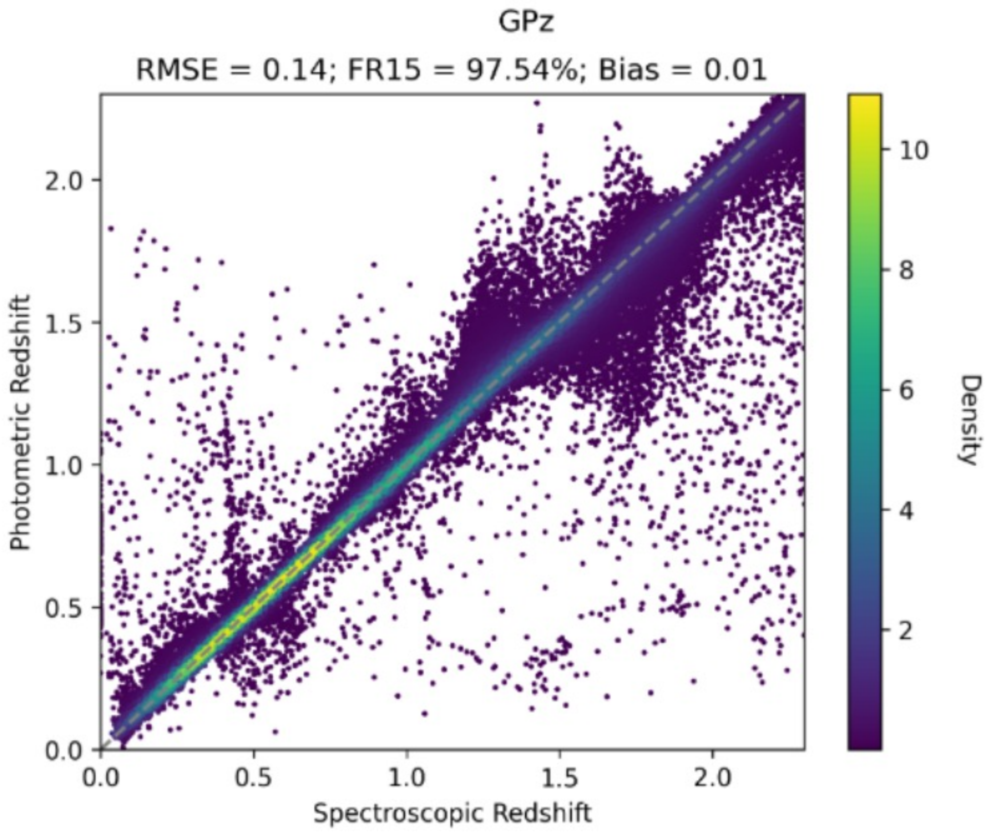


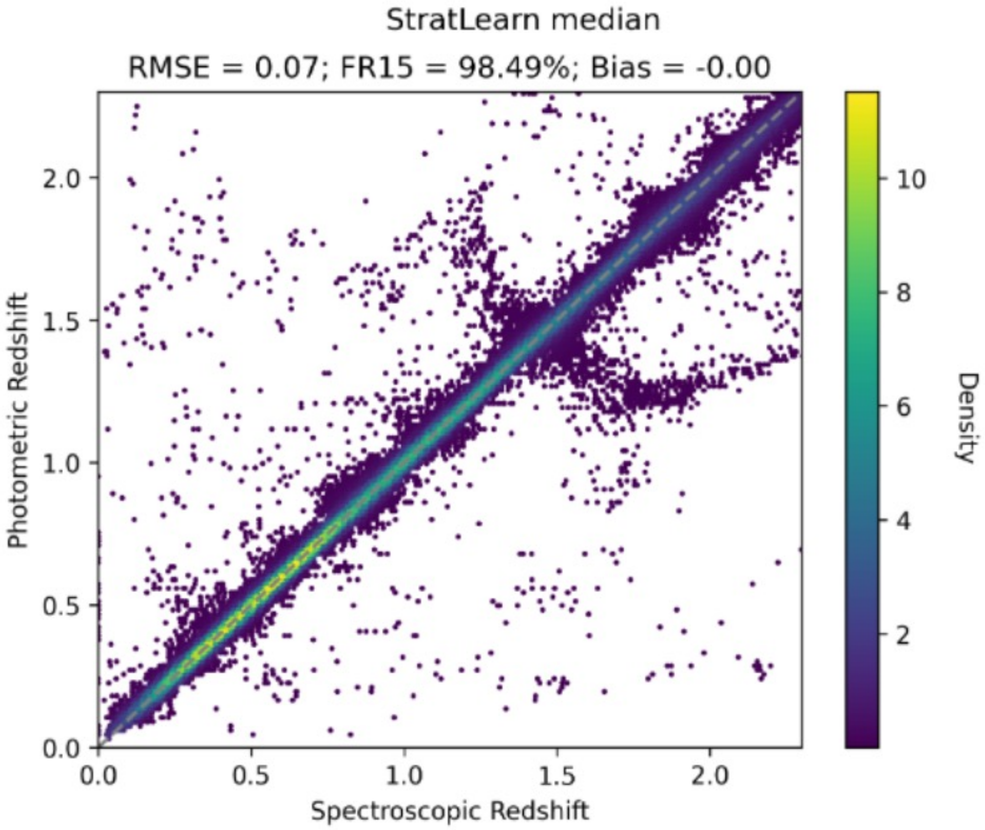
Work by Riccardo Serra & Chiara Moretti

Ongoing work

Using StratLearn:
- Smaller RMSE
- No bias
- Smaller number of catastrophic outliers.

For comparison – standard method





Work by Riccardo Serra & Chiara Moretti

Next Steps and Expected Results (by next checkpoint: April 2024)

- Publication of Weak Lensing calibration methods (MICE sims) – target: Dec 2024
- Publication of photo-z proof-of-concept on sims – target: Feb 2024
- Application to Euclid simulation of photo-z reconstruction – target April 2024
- Application to downstream WL analysis and cosmological parameters – target April 2024
- 3 presentations in academic conferences, 3 in public outreach eventa – target April 2024

Conclusions

- StratLearn offers a principled, all-purpose solution to the problem of unrepresentative training data
- Best-in-class performance for SNIa classification (not shown), WL tomographic binning, photo-z estimation