Sensitivity of Machine Learning Photo-z Estimates to Training Set Degradations

Natalia Stylianou, University of Oxford

Research Objectives:

- Photo-z are key for almost all Rubin extragalactic and cosmological science, hence their accuracy is crucial.
- In this work, we use the photometric redshift code GPz to examine 2 realistically complex training set imperfections scenarios for machine learning based photometric redshift calculation.

How do we estimate photo-z with Machine Learning?

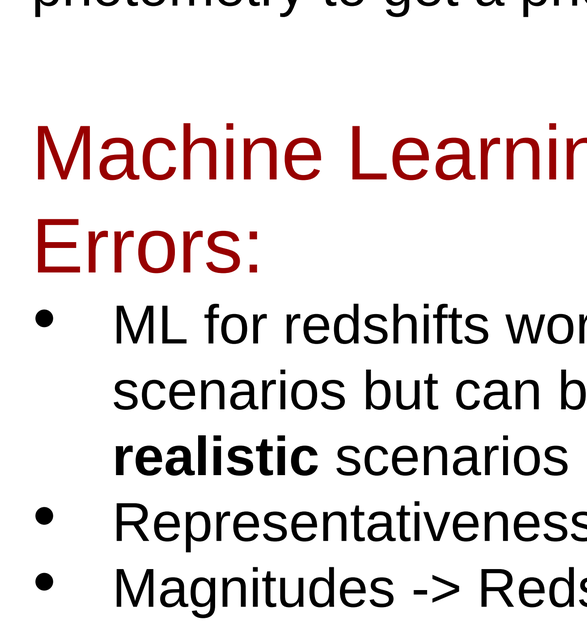
We take a sample of galaxies, with both the photometry and the spectroscopy and use that as our training set and train the mapping from the photometry onto a redshift. We then apply this to galaxies where we only have the photometry to get a photo-z estimate.

Machine Learning – Systematic Errors:

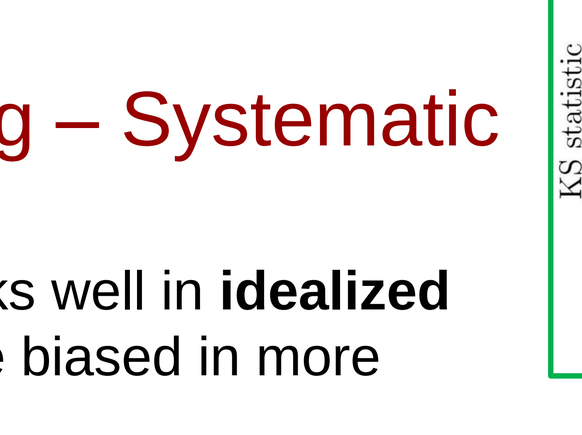
- ML for redshifts works well in **idealized** scenarios but can be biased in more **realistic** scenarios
- Representativeness of the training data
- Magnitudes -> Redshift is not a one-to-one relation; it is more of a one-to-many relation

As an attempt to test more realistic scenarios we use <u>2 training set imperfections</u>:

spectroscopic errors **and** sample incompleteness



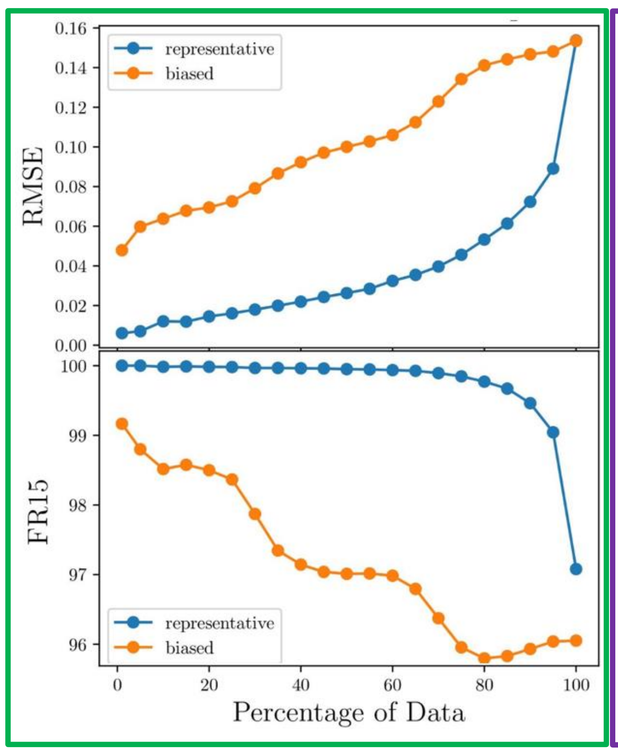
[OII] emission lines are mistaken for Hα lines, and vice versa

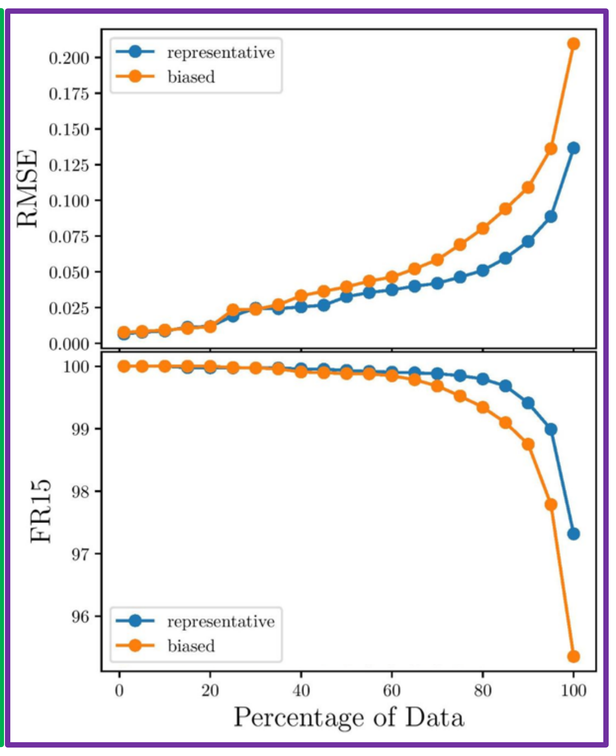


Simulates incompleteness with a selection function inversely proportional to redshift

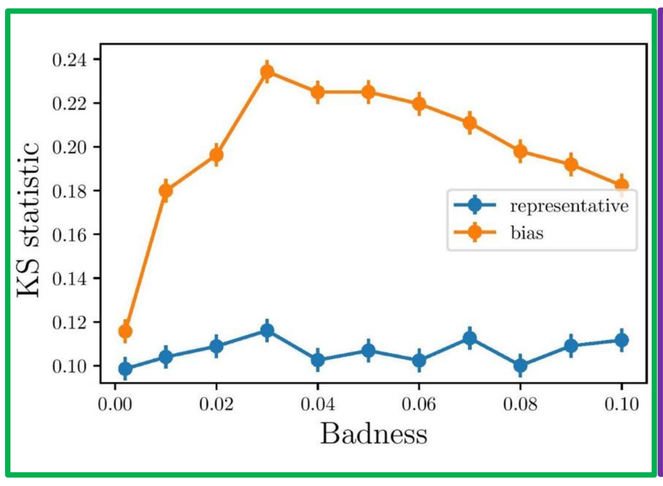
Results:

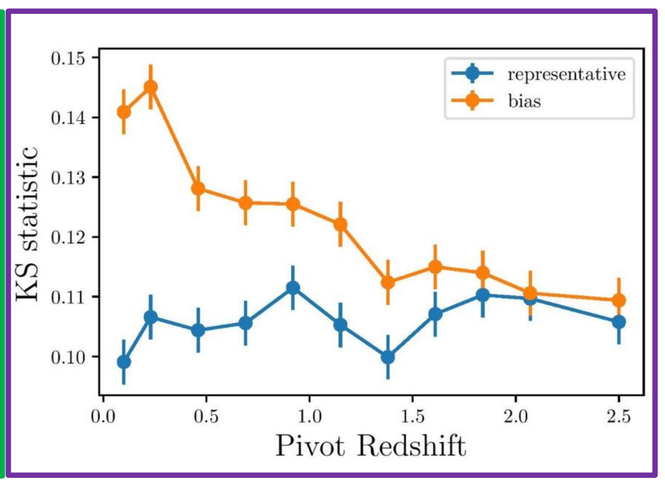
- Photo-z Point Estimates get worse with higher degrees of degradation.





- PIT Metric (assesses how “realistic” a population of photo-z PDFs is compared with the true z) Statistics:





Summary:

- Photo-z can be calculated with ML using spec-z to train the algorithm and can be systematically wrong in various ways.
- We have assessed the impact on photo-z accuracy for two known forms of degradation.
- We found a <u>decrease in photo-z estimates quality with</u>:

  above 1% emission line confusion **and** below 1.5 redshift incompleteness

Check out my paper here

Acknowledgement: I would like to thank my advisors for all their help and support; Alex Malz, Peter Hatfield and co-authors John Franklin Crenshaw and Julia Gschwend. This work was supported by a 2021 LSSTC Enabling Science grant.

Participant in the LSST Corporation’s program for student researchers at the 2022 Rubin PCW.