DIVISION OF
ASTRONOMY AND ASTROPHYSICS
Faculty of Mathematics, Physics and Informatics
Comenius University
Bratislava

PREDICTING THE HI MASS WITH PROBABILISTIC RANDOM FOREST

Mgr. Filip Janák
Dr. Boris Deshev, PhD.
RNDr. Roman Nagy, PhD.

Colloquium 20.11.2024

Presentation outline

- Introduction & Review of our previous work & Motivation
- Probabilistic Random Forest model
- PRF performance
- PRF applications:
  - HI deficiency of ALFALFA galaxies
  - Unaltered HI mass of SDSS galaxies

Introduction

- Huge amount of astrophysical data in recent years
- ML offers a promising solution in data processing and analysis
- Supervised ML
  - Classification, regression
  - **Features** (predictors) and **Target variable**
  - Establish relation/dependence between features and target variables

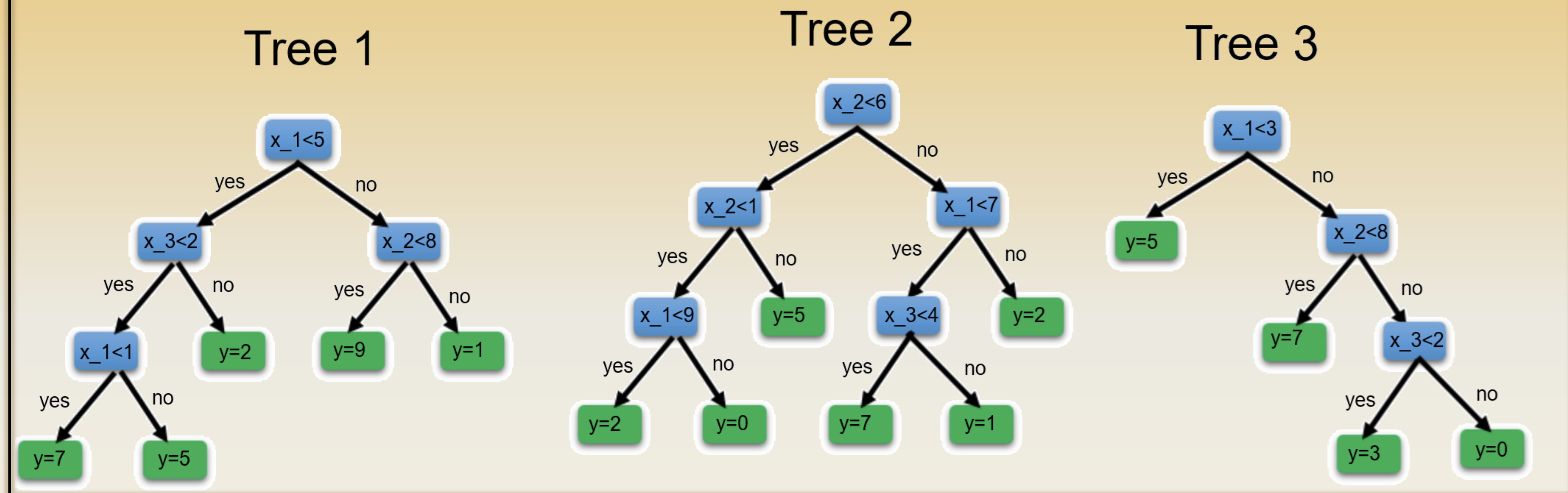
Review of our previous work

- Previous results:
  - Prediction of the unaltered (expected) HI mass of galaxies
  - Establishing the HI deficiency
  - Analysis of the relation between galactic environment and the HI deficiency

- Random Forest (RF)
  - Classification, regression
  - RF consist of decision trees
  - Each tree is built using a random subsample of training set
  - Each tree consists of **nodes**
    - Split nodes (**feature**, **threshold**)
    - Leaf nodes (**value of the target variable**)

RANDOM FOREST

Features: x_1, x_2, x_3
Target variable: y



- Why RF?
  - RF provides very high accuracy for our type of our problems
  - RF delivers good control and transparency of the created model
  - RF model includes intrinsic feature importances based on the variance reduction which are available for the model itself

PRF - Motivation

- Measurement uncertainties (errors) are inherent in every observation
- ML models (Random Forest) overlooks measurement errors
- Errors may provide valuable additional information
- PRF = RF incorporating errors

- Reis et al. 2019
- Probabilistic Random Forest Classifier

The Astronomical Journal, 157:16 (12pp), 2019 January
https://doi.org/10.3847/1538-3881/aaf101

CrossMark

Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets

Itamar Reis, Dalya Baron, and Sahar Shahaf
School of Physics and Astronomy Tel-Aviv University Tel Aviv 69978, Israel; itamarreis@mail.tau.ac.il, dalyabaron@gmail.com
Received 2018 October 11; revised 2018 November 7; accepted 2018 November 12; published 2018 December 20

Abstract

Machine learning (ML) algorithms have become increasingly important in the analysis of astronomical data. However, because most ML algorithms are not designed to take data uncertainties into account, ML-based studies are mostly restricted to data with high signal-to-noise ratios. Astronomical data sets of such high quality are uncommon. In this work, we modify the long-established Random Forest (RF) algorithm to take into account uncertainties in measurements (i.e., features) as well as in assigned classes (i.e., labels). To do so, the Probabilistic Random Forest (PRF) algorithm treats the features and labels as probability distribution functions, rather than deterministic quantities. We perform a variety of experiments where we inject different types of noise into a data set and compare the accuracy of the PRF to that of RF. The PRF outperforms RF in all cases, with a moderate increase in running time. We find an improvement in classification accuracy of up to 10% in the case of noisy features, and up to 30% in the case of noisy labels. The PRF accuracy decreased by less then 5% for a data set with as many as 45% misclassified objects, compared to a clean data set. Apart from improving the prediction accuracy in noisy data sets, the PRF naturally copes with missing values in the data, and outperforms RF when applied to a data set with different noise characteristics in the training and test sets, suggesting that it can be used for transfer learning.

Key words: methods: data analysis – methods: statistical

PRF Regressor – basic concepts

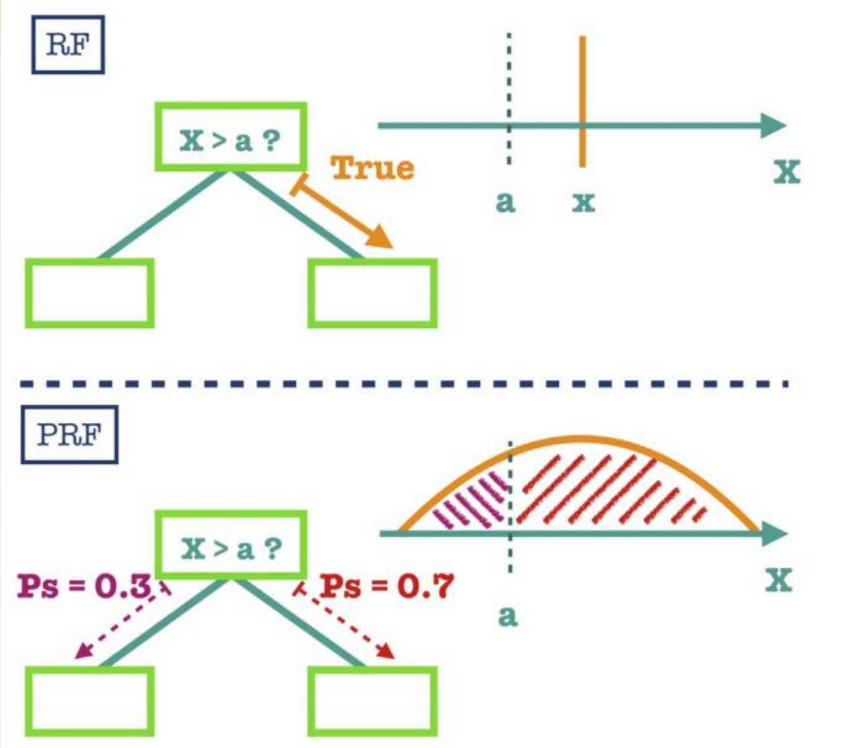
- Each entry is treated as a **distribution**
  - Splits are not deterministic
- Normal distribution
- Entry **value** corresponds with the **mean** μ of the distribution
- Entry **value error** corresponds with the **standard deviation** σ of the distribution

PRF – model training

- Each tree is built using a random subsample of training set (with or without replacement)

- Each tree consist of **nodes**
  - Split nodes (**feature**, **threshold**, left subtree, right subtree, variance reduction)
  - Leaf nodes (**value**, **error**/sigma)

- Creating a split node:
  - Iterate over all possible thresholds (+ middle values) and features (subsample of features) to find the **best split**



Reis et al. 2019

PRF – model training

- Split: each entry propagates to both (left and right) nodes with some probability (unless the probability is less than some threshold value)

- Best split maximizes ***variance reduction***

- $VAR_{red} = VAR_{parent\ node} - \left(w_{left\ node} VAR_{left\ node} + w_{right\ node} VAR_{right\ node}\right)$ (1)
- $w_{left\ node} + w_{right\ node} = 1$ (2)
- $VAR = \sum_{i=1} p_i \left(\mu_i^2 + \sigma_i^2\right) - \left(\sum_{i=1} p_i \mu_i\right)^2$ (3)

PRF – model training

- Tree is built recursively
- Split nodes are created until one of the stopping conditions is met – then a leaf node is created:
- Leaf value: $\sum_{i=1} p_i \, \mu_i$ (4)
- Leaf error: $\sqrt{\sum_{i=1} p_i \left(\mu_i^2 + \sigma_i^2\right) - \left(\sum_{i=1} p_i \, \mu_i\right)^2}$ (5)

PRF – model predicting

- Prediction from single tree:
  - Each entry propagates through the whole tree, reaching „each“ leaf node with some probability (in reality there is also some limiting probability threshold)
  - We aggregate predictions from all leaf nodes using Eq. (4) and (5)

- We aggregate predictions from all trees using Eq. (4) and (5)

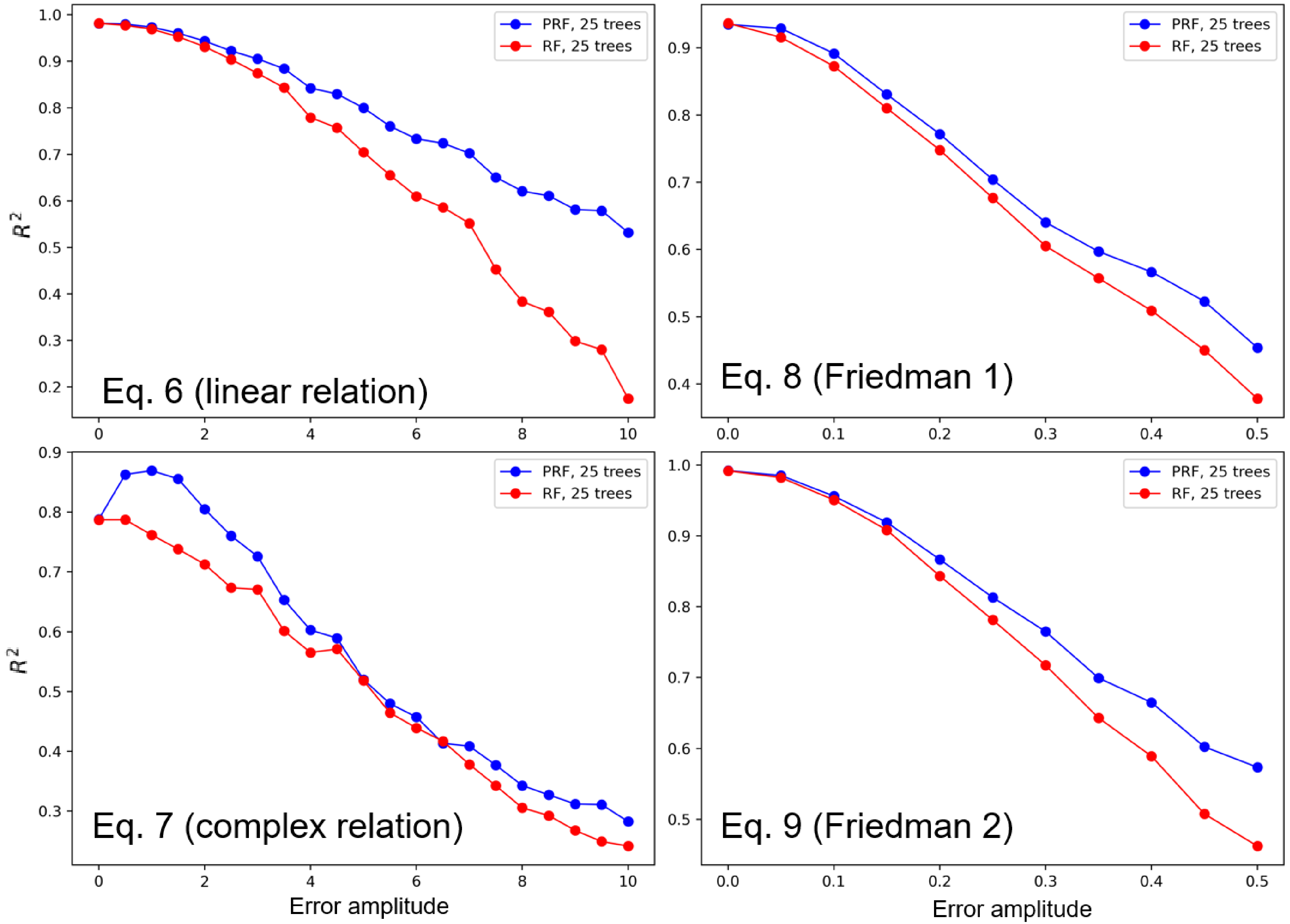
PRF - advantages

- Better performance than RF
- Natural representation of missing values
- Uncertainties of predictions
  - Predictions are distributions
- Converges faster than RF

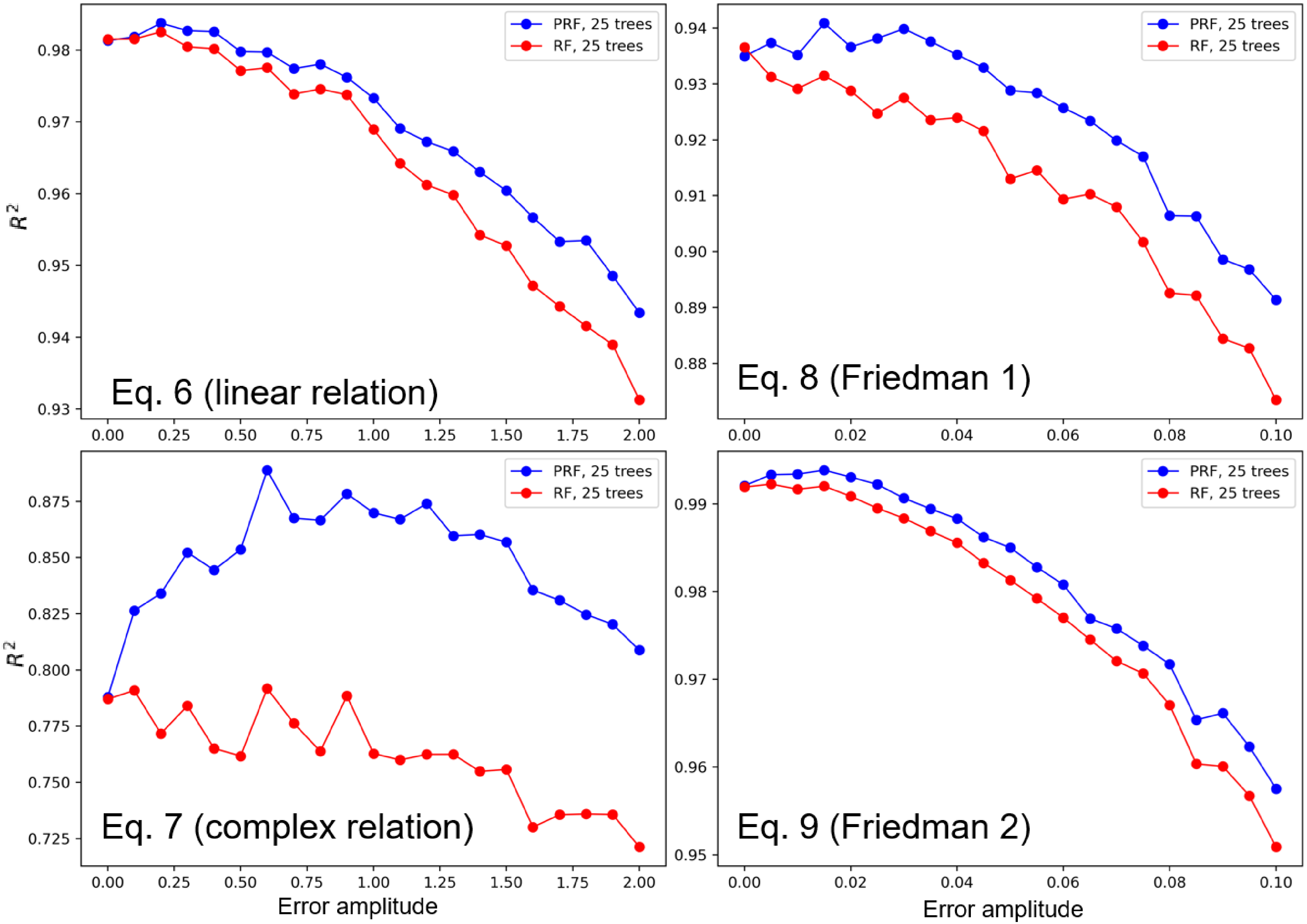
Performance: PRF vs RF

- $y = 0.04x_1 + 0.12x_2 + 0.2x_3 + 0.28x_4 + 0.36x_5$ (6)
  - 10 features, 5 informative
- $y = 20\sin\left(\frac{\pi x_1 x_2}{100}\right) + 10\cos\left(\frac{\pi x_3}{10}\right) + 2x_4 + \left(\frac{x_5}{2}\right)^2$ (7)
  - 6 features, 5 informative
- $y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$ (8)
  - 10 features, 5 informative
- $y = \sqrt{{x_1}^2 + \left(x_2 x_3 - \frac{1}{x_2 x_4}\right)^2}$ (9)
  - 5 features, 4 informative

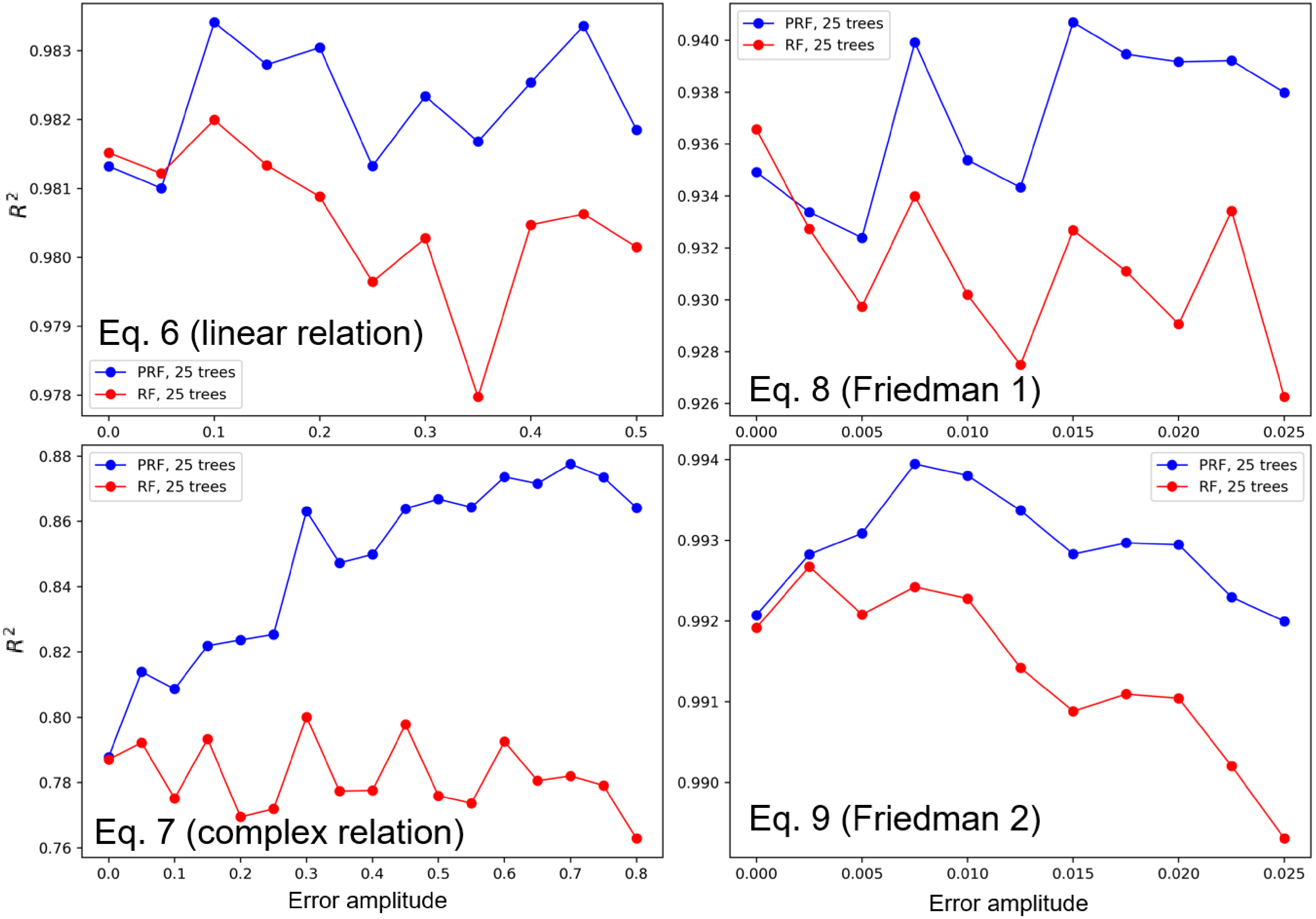
Large error amplitudes



Normal error amplitudes



Small error amplitudes



Performance: PRF vs RF

- PRF usually outperforms RF (Especially when errors are large)

- In some cases the PRF performance increases with error amplitude
  - Only for small error amplitudes
  - This is not caused by increasing noise in the data but due to the fact that the model "knows" that the noise is getting larger $\Rightarrow$ model can smooth more over the data
  - Overstated features errors resolve this problem

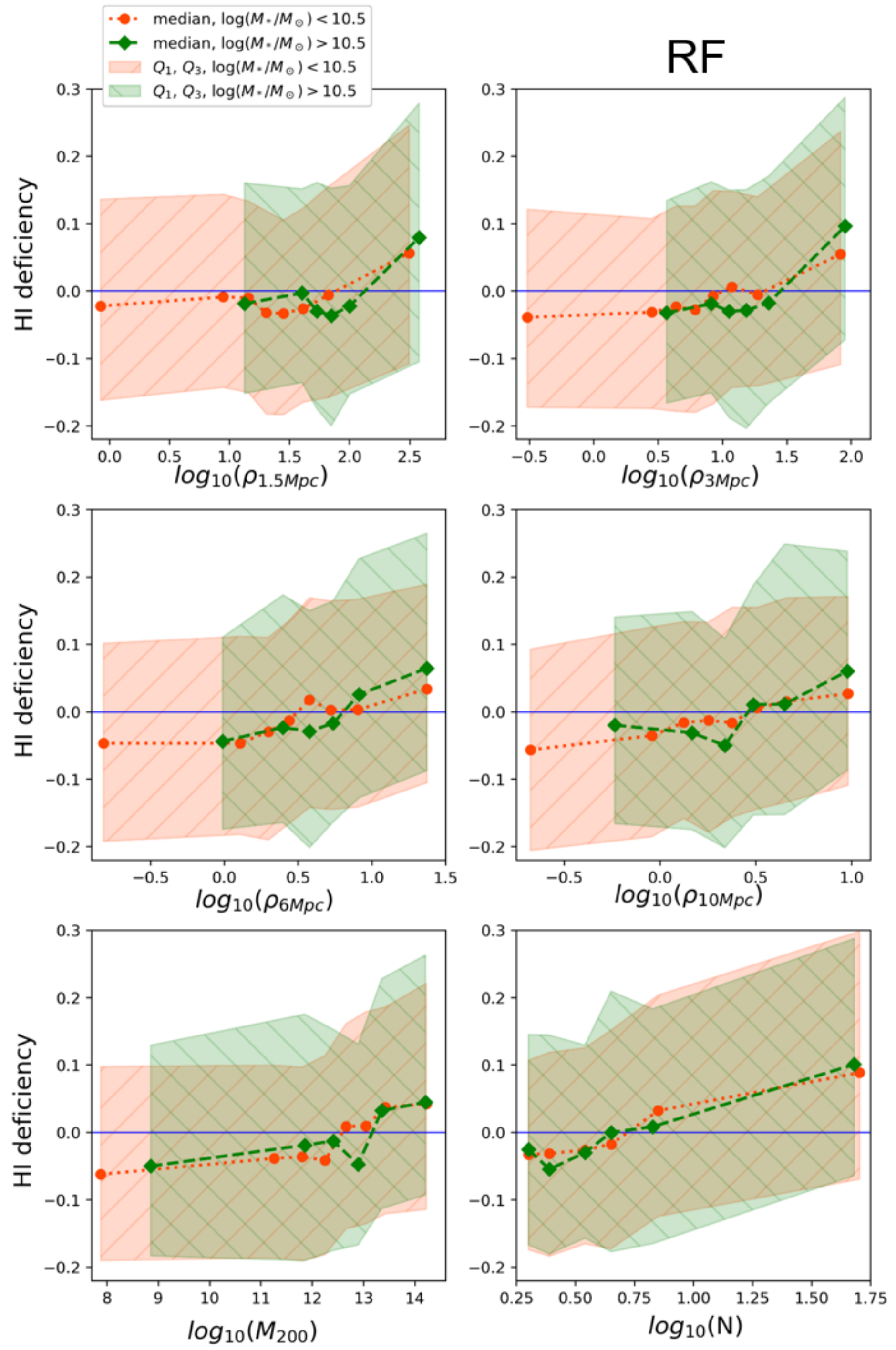
HI deficiency of ALFALFA galaxies

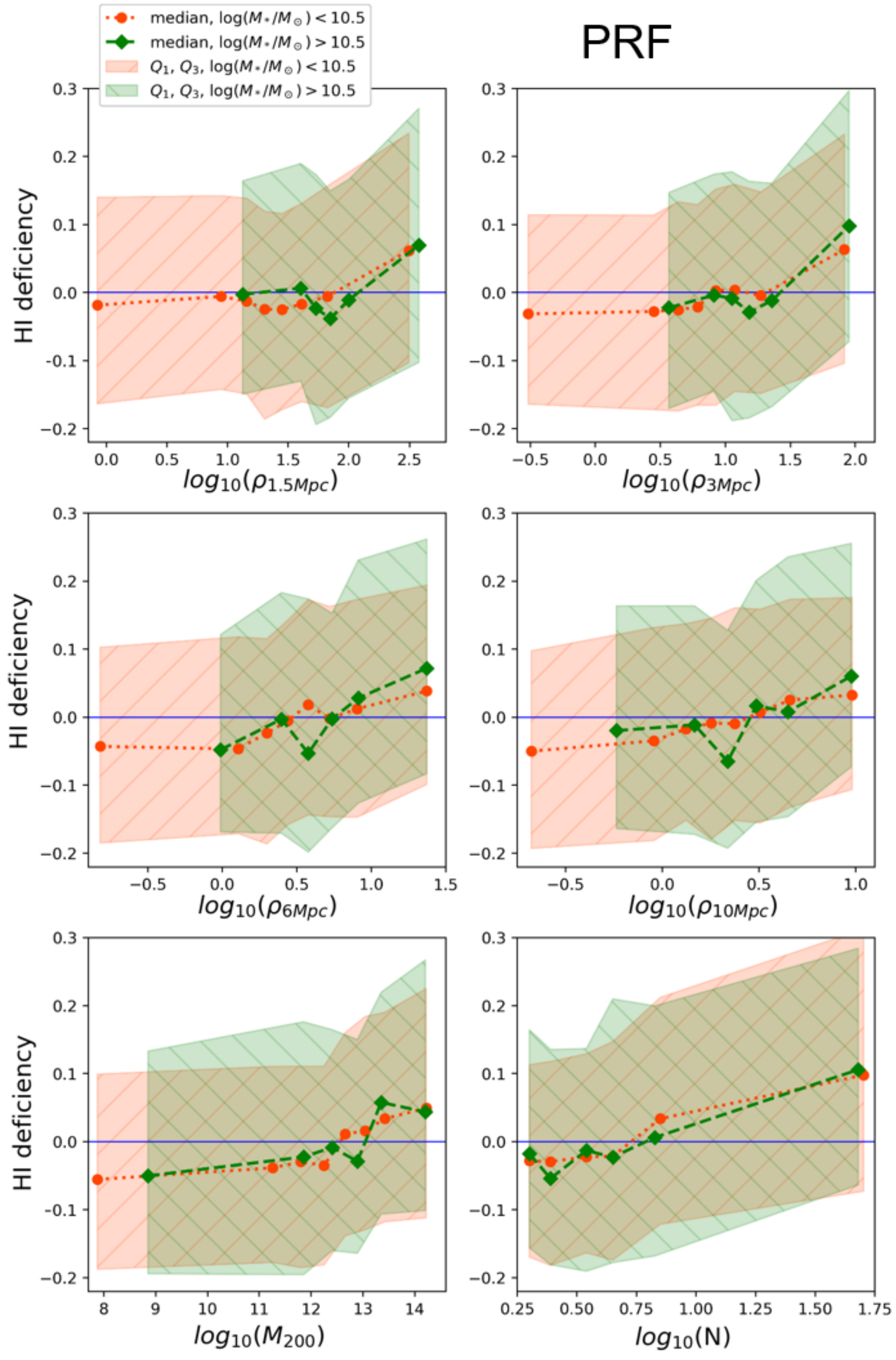
- Used data:
  - ALFALFA (Haynes et al. 2018)
  - ALFALFA-SDSS cross-match (Durbala et al. 2020)
  - Galactic environment (Tempel et al. 2017)
  - SDSS DR15 (Blanton et al. 2017)

- Predictive model:
  - Isolated ALFALFA galaxies
  - Predictors (features) – SDSS photometry
  - Target variable – unaltered (expected) HI mass

- HI deficiency:
  - HI deficiency = expected HI – observed HI
  - Predictive model → expected HI for ALFALFA galaxies with environment

- HI deficiency increases with galactic environment


RF
median, log($M_*/M_\odot$) < 10.5
median, log($M_*/M_\odot$) > 10.5
Q_1, Q_3, log($M_*/M_\odot$) < 10.5
Q_1, Q_3, log($M_*/M_\odot$) > 10.5
HI deficiency
$log_{10}(\rho_{1.5Mpc})$
$log_{10}(\rho_{3Mpc})$
$log_{10}(\rho_{6Mpc})$
$log_{10}(\rho_{10Mpc})$
$log_{10}(M_{200})$
log_{10}(N)


PRF
median, log($M_*/M_\odot$) < 10.5
median, log($M_*/M_\odot$) > 10.5
Q_1, Q_3, log($M_*/M_\odot$) < 10.5
Q_1, Q_3, log($M_*/M_\odot$) > 10.5
HI deficiency
$log_{10}(\rho_{1.5Mpc})$
$log_{10}(\rho_{3Mpc})$
$log_{10}(\rho_{6Mpc})$
$log_{10}(\rho_{10Mpc})$
$log_{10}(M_{200})$
log_{10}(N)

Unaltered HI mass of SDSS galaxies

- Predictors (features) – SDSS photometry and spectroscopy
  - SDSS DR17 spectroscopy – around $3 \cdot 10^6$ galaxies
- Work in progress

References

- Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
- Durbala, A., Finn, R. A., Crone Odekon, M., et al. 2020, AJ, 160, 271
- Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, ApJ, 861, 49
- Reis, I., Baron, D. & Shahaf, S. 2019, AJ, 157, 16
- Tempel, E., Tuvikene, T., Kipper, R., & Libeskind, N. I. 2017, A&A, 602, A100