ANALYSIS OF THE HI CONTENT OF GALAXIES WITH MACHINE LEARNING

Filip Janák, Comenius University, Bratislava
Boris Deshev, Astronomical Institute of the Czech Academy of Sciences, Prague

ABSTRACT

HI content of galaxies can serve as an important tracer that enables us to probe the galaxy environment and evolution. More precisely, we focused on the HI deficiency which is established as the difference between expected unaltered and observed HI content of a galaxy. Besides internal effects such as the presence or absence of an active galactic nucleus or the star formation rate, total HI deficiency of the system is closely related to environmental effects in large groups and clusters.

In this work we created a machine learning random forest model based on the isolated ALFALFA sources, which can predict an expected HI content of a galaxy with given optical quantities. Our model performs significantly better (RMSE ≈ 0.22 dex and R^2 ≈ 0.8) than the popular approach where HI mass is linearly related with the optical size (both in logarithmic scale), resulting in RMSE ≈ 0.33 dex and R^2 ≈ 0.56.

Finally, we trained the model on the subsample of isolated galaxies and predicted the expected HI content for ALFALFA galaxies with environment, which enabled us to compute the HI deficiency. Although the scatter is very high (difference between 1st and 3rd quartile around 0.3 dex), results show prevailing growing dependency of HI deficiency on the environment (increase of up to 0.15 dex).

INTRODUCTION

The neutral hydrogen is the most important element supplying the star formation, which preordains the galaxy evolution. Besides internal factors (e. g. presence or absence of an active galactic nucleus (AGN), star formation rate, total mass of the system) the HI content of a galaxy can be dramatically altered by effects of an environment. Those effects involve mechanisms such as tidal interactions among galaxies or between galaxies and the cluster, galaxy harassment, ram pressure stripping, viscous stripping, thermal evaporation or starvation (see Boselli & Gavazzi (2006) for more details).

Due to the heavy and complex effects of the environment on the HI content of a galaxy, it is much more instructive to first study HI content of galaxies which do not experience any effects of an environment (isolated galaxies). Considering only isolated galaxies (IGs), we can establish an expected (unaltered) HI content and investigate how much the effects of an environment alter the HI content. This, in return, enables us to study the galactic environment.

DATA

HI DATA

We used the Arecibo Legacy Fast ALFA (ALFALFA) extragalactic HI catalog (Haynes et al. 2018) as a source of the neutral hydrogen information (around 31 500 galaxies up to the redshift z<0.06).

OPTICAL DATA

Durbala et al. (2020) released a catalog (ALFALFA-SDSS) which contains a cross-match of the ALFALFA with the SDSS DR15 (Blanton et al. 2017) photometric catalog, which served as a source of the input (optical) data for our predictive model. We selected petrosian magnitudes, petrosian radii, petrosian radii containing 50% of the total flux and petrosian radii containing 90% of the total flux (each of them in g, r and i bands). We also included calculated concentration and color indices.

GALACTIC ENVIRONMENT

Generally, there exist many ways, how to quantify the galactic environment. Also, different authors may use different approaches, algorithms and definitions. We used a spectroscopic catalog by Tempel et al. (2017), which has the best cross-match with the ALFALFA-SDSS sample providing an environmental information to approximately one half of ALFALFA-SDSS objects. We used six environmental tracers from this catalog: the number of galaxies N_{gal} within a group (N_{gal} = 1 means, that the galaxy is isolated), the estimated mass of the group a galaxy belongs to M_{200} (mass located within the sphere where the mean density is 200 times larger than the mean density of the Universe) and the normalized environmental density of the galaxy for smoothing scales 1.5, 3, 6 and 10 Mpc.

SAMPLE OF ISOLATED GALAXIES

To create a predictive model which will be able to predict the expected (unaltered) HI mass, we selected only isolated galaxies (N_{gal} = 1) to obtain the training dataset. Pearson correlation coefficients between optical features (model predictors) and the total HI mass (the target variable) are given in Fig. 1.

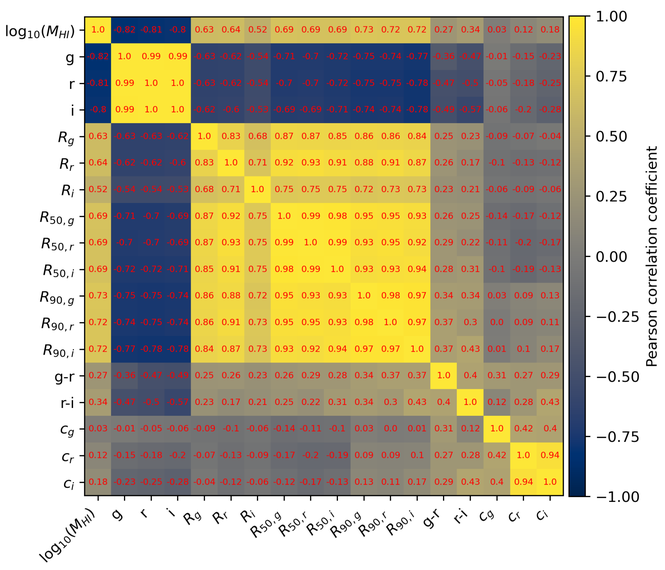

Fig. 1: The Pearson correlation matrix for features and the target variable for our sample of IGs.

LINEAR MODEL

The common approach in literature connects the HI mass with the optical diameter (both on the logarithmic scale) with the linear relation - see e. g. Haynes & Giovanelli (1984). We created a linear fit for our sample of isolated galaxies (see Fig. 2) and compared our results with Haynes & Giovanelli (1984) (paper inspecting properties of 324 isolated galaxies) and Wang et al. (2016). However, Wang et al. (2016) used the HI diameter and not the optical diameter. Authors investigated around 500 nearby galaxies, without restriction to isolated galaxies only. Our linear model has RMSE ≈ 0.33 dex and R^2 ≈ 0.56.

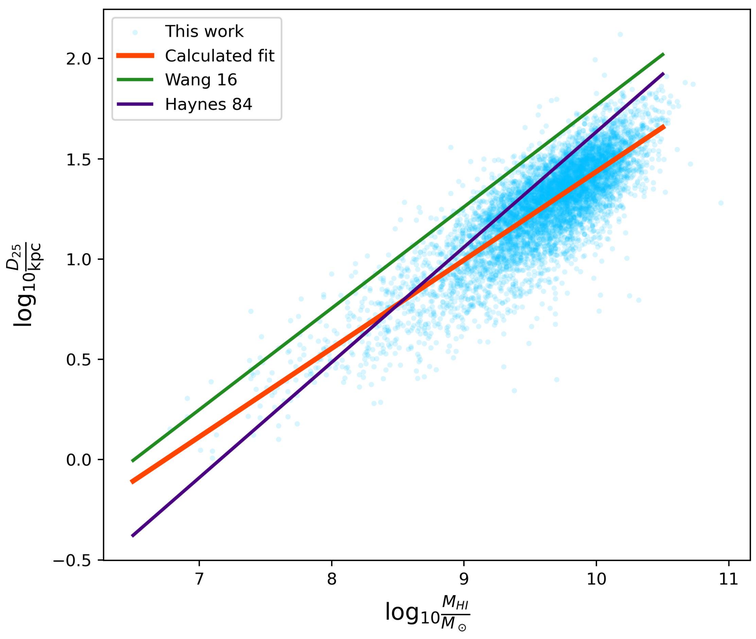

Fig. 2: HI mass from our sample of IGs versus the optical diameter fitted with the red line, compared with Wang et al. (2016) (green line) and Haynes & Giovanelli (1984) (purple line).

RANDOM FOREST MODEL

To create a more accurate predictive model, we used the random forest algorithm (Breiman 2001). We chose the random forest as it provides the high accuracy for the type of problem which we want to solve while delivering good control and transparency of the created model. We used 17 optical features as predictors (see Fig. 1) of the total expected unaltered HI mass of isolated galaxies. With RMSE ≈ 0.22 dex and R^2 ≈ 0.8, the random forest model performs much better than the simple linear fit. The relative feature importances of the random forest model are given in Table 1.

Impurity based		Permutation based	
g	28.5%	$R_{90,g}$	42.2%
$R_{90,g}$	24.3%	g	30.6%
$R_{90,i}$	12.3%	r	8.2%
r	12.1%	$R_{90,r}$	5.2%
$R_{90,r}$	5.6%	g - r color	3.4%
i	2.4%	$R_{90,i}$	2.8%
g - r color	2.3%		

Permutation and impurity based relative feature importances of our random forest model (only >2% are listed).

The strongest predictors became the absolute magnitude g and the petrosian radius $R_{90,g}$. Overall, petrosian radii containing 90% of the flux in all bands are significantly more important for the model than true petrosian radii or the 50% flux petrosian radii. The absolute magnitude in r band with petrosian radius $R_{90,r}$ are also important model predictors (relative importances above ~5% in both columns of Table 1). Generally, relative feature importances of the model are in very good agreement with correlations between features and the target variable.

HI DEFICIENCY

We used our random forest model trained on the whole sample of IGs to predict the expected HI mass of ALFALFA galaxies with environment and calculated the HI deficiency as the difference between predicted and observed HI mass. Besides environmental effects the HI content of a galaxy can be modified by internal factors. As we aim to study effects of the galactic environment, we want to minimize the influence of internal factors. Therefore we removed AGNs from our analysis and split the dataset into low-mass and high-mass galaxies, based on the stellar mass. Galaxies with decadic logarithm of the stellar mass in solar units smaller than 10.5 act qualitatively different than higher stellar mass galaxies: the molecular hydrogen fraction decreases significantly with the neutral hydrogen for higher stellar masses while remaining approximately constant for low-mass galaxies (Saintonge et al. 2017).

Fig. 3 shows a calculated median of the HI deficiency for the subsample of low-mass galaxies (green) and high-mass galaxies (orange) with respect to the six environmental tracers. Galaxies are binned into six or eight equally populated bins. The x-axis position of each bin is determined as the middle value between the bin edges. The shaded area represents the 1st and 3rd quartiles for each distribution. Since we used the equal number of objects per bin, they all have the same statistical significance.

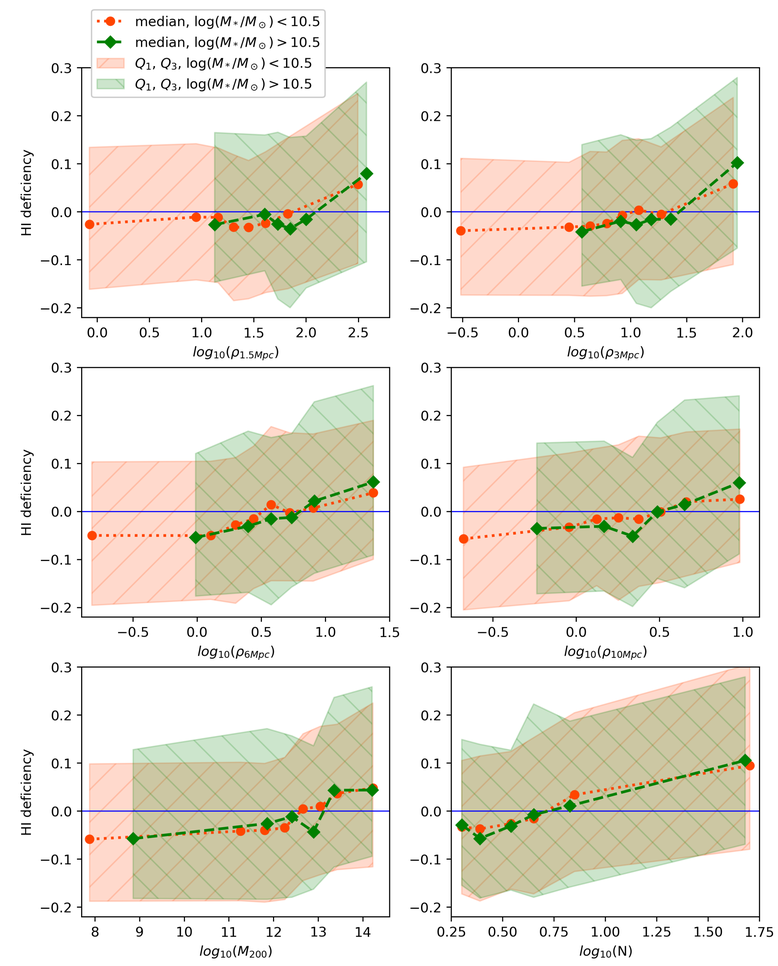

Fig. 3: HI deficiency with respect to different environmental tracers; low-mass galaxies (orange, eight or six bins with equal number of objects) and high-mass galaxies (green, six bins with equal number of objects). The shaded area represents the 1st and 3rd quartiles of the distribution.

The HI deficiency is related with the galactic environment, increasing from around -0.5 dex at sparse environments up to the 0.1 dex at the densest environments. The sparse environment edges of high-mass curves start below or around the low-mass curves, but finish above them (except for the $\log_{10}(M_{200})$ environmental tracer, see the right-most bins in Fig. 3). We can notice, that the scatter is huge (difference between 1st and 3rd quartile around 0.3 dex). Indeed, random errors in measurements as well as in the model predictions make the distributions broader, but should not offset them. Therefore, the difference between low-mass and high-mass median lines, although very small, may indicate real (internal) effects that the stellar mass has on the HI deficiency. Although the HI deficiency with respect to various tracers expresses moderately different profiles, the prevailing growing tendency with an environment is evident. This confirms, that the environment depletes the neutral hydrogen gas from galaxies.

ACKNOWLEDGEMENTS

•This work was supported by the Visegrad Fund grant No. 22210105.
•This research was supported by a grant from the European Astronomical Society thanks to the generous support of the MERAC Foundation and Springer Verlag.
•This work was supported by the VEGA - the Slovak Grant Agency for Science, grant No. 1/0761/21.

REFERENCES

•Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
•Boselli, A. & Gavazzi, G. 2006, PASP, 118, 517
•Breiman, L. 2001, Machine Learning, 45, 5
•Durbala, A., Finn, R. A., Crone Odekon, M., et al. 2020, AJ, 160, 271
•Haynes, M. P. & Giovanelli, R. 1984, AJ, 89, 758
•Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, ApJ, 861, 49
•Saintonge, A., Catinella, B., Tacconi, L. J., et al. 2017, ApJS, 233, 22
•Tempel, E., Tuvikene, T., Kipper, R., & Libeskind, N. I. 2017, A&A, 602, A100
•Wang, J., Koribalski, B. S., Serra, P., et al. 2016, MNRAS, 460, 2143