Implementation of Statistical Learning Methods to Classify and Predict Water Maser Phenomena

Ty Nunley[1], Dr. Nusrat Jahan[1], & Dr. Anca Constantin[2]

[1]Department of Mathematics and Statistics, James Madison University,

[2]Department of Physics and Astronomy, James Madison University

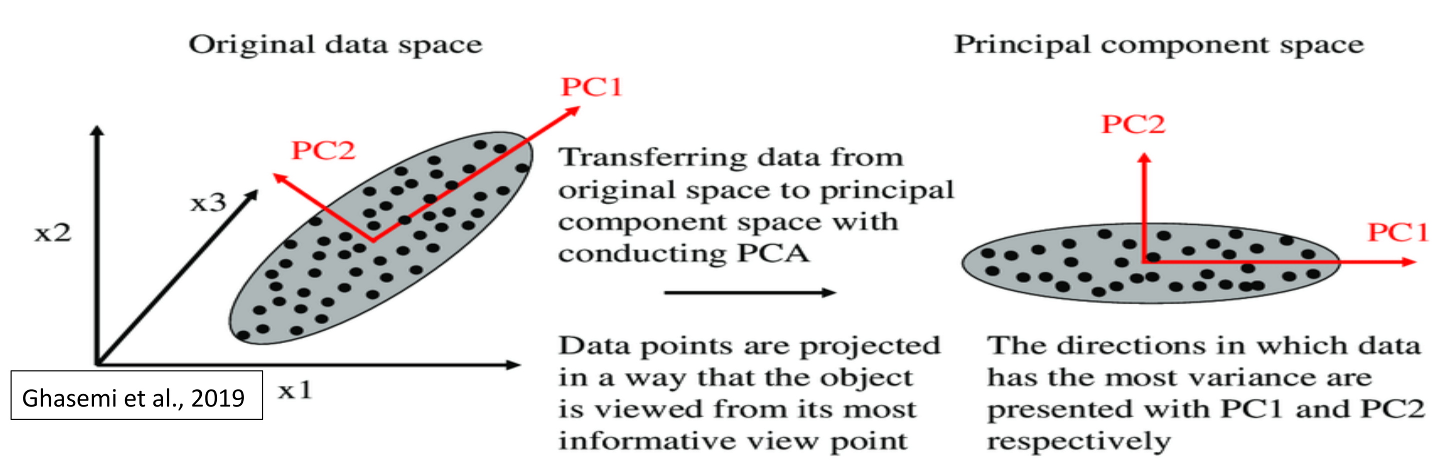
Abstract

We present here an investigation of the power of statistical learning techniques to classify and predict new Microwave Amplification by Stimulated Emission of Radiation (maser) emissions from galaxy centers. The maser phenomenon is important because when detected at levels that surpass millions of times the brightness of similar emissions in star forming regions of our galaxy (i.e., mega-masers), it can be used as a unique tool to constrain both masses of supermassive black holes and the current cosmological models (and therefore the fate of the universe). Unfortunately, mega-maser detections are extremely rare, accounting for ~3 -5% of all surveyed galaxies. We use supervised principal component analysis (SPCA) and random forests to develop a classification tool to distinguish between the non-maser and mega-maser galaxies based on optical data. The SPCA allows us to identify the most relevant optical properties for discriminating between mega-masers and non-masers, and the random forests allows us to make predictions for new mega-maser identifications in future galaxy surveys.

Tools used for Statistical Learning

Principal Component Analysis (PCA)

PCA is an unsupervised (exploratory) statistical learning tool used to reduce the dimensionality of a data set. Our goal is to describe the most amount of variability in the data using the least number of dimensions possible. Through PCA, we are creating linear combinations, or principal components (PCs, or combinations of eigenvectors) of the parameters that characterize a sample of objects (i.e., galaxies). Eigenvalues show how much variability in the data is explained by each PC.

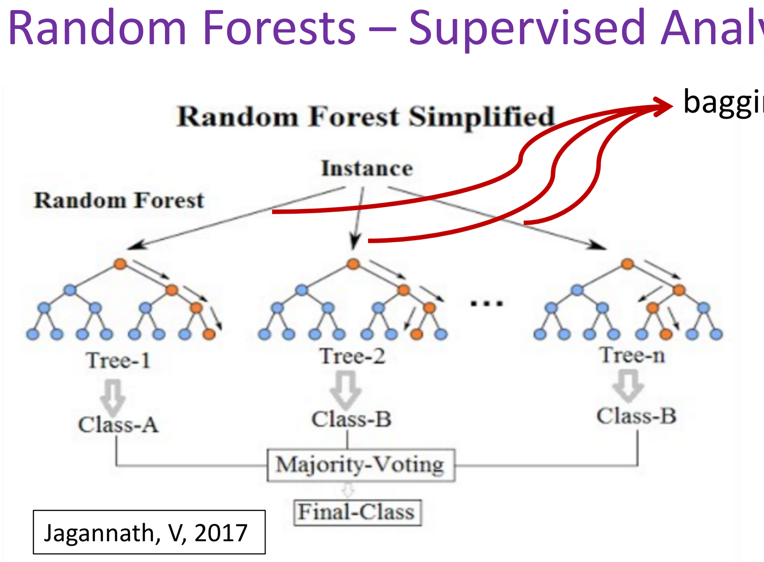


Ghasemi et al., 2019

Supervised Principal Component Analysis (SPCA)

Using the eigenvectors and eigenvalues from PCA, SPCA selects a subset of parameters that are most relevant to the response (i.e., the statistical decision on the mega-maser/non-maser classification). Relevant parameters have high correlations with the response and have high discriminatory power for classification, that ultimately provide a classifier.

Random Forests – Supervised Analysis



Jagannath, V, 2017

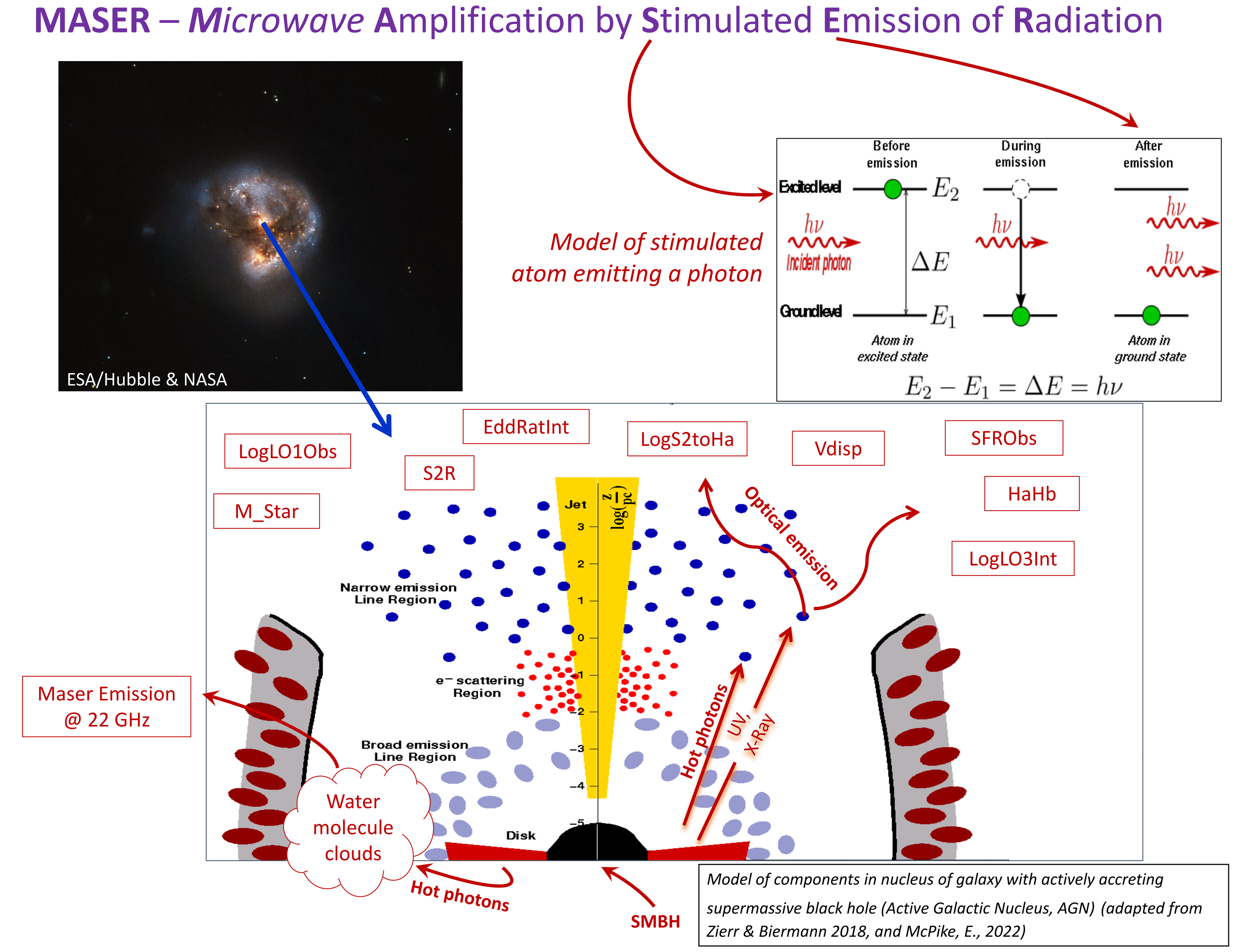
Random Forests (RF) are a supervised learning method used to classify a binary response through a "forest" of decision trees created through bagging data.

- Bagging = Bootstrapping and aggregating.

- Bootstrapping = the process of randomly sampling data multiple times, with replacement.
- Aggregating = tallying the data as it is run down the decision trees created by the random forest algorithm.

Our Data: Finding Maser Galaxies

MASER – ***Microwave*** **A**mplification by **S**timulated **E**mission of **R**adiation



Model of components in nucleus of galaxy with actively accreting supermassive black hole (Active Galactic Nucleus, AGN) (adapted from Zierr & Biermann 2018, and McPike, E., 2022)

Our Data: Parameter Information

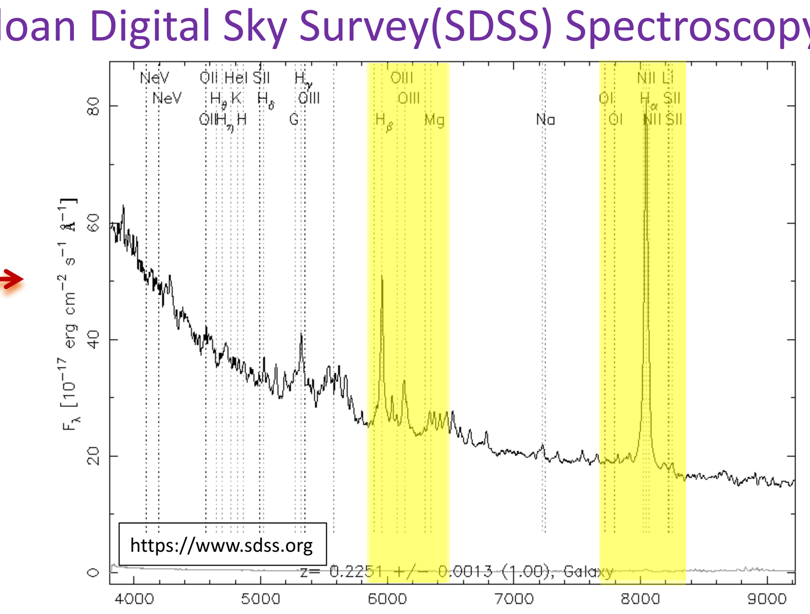
Maser Cosmology Project (MCP)

https://safe.nrao.edu/wiki/bin/view/Main/MegamaserCosmologyProject

- Largest catalog of surveyed galaxies in 22GHz for water maser emission.
- Masers: ~3% of surveyed galaxies.
- 80% of all masers surveyed are mega-masers ($L_{H_2O} > 10\ L_{Sun}$).
- 20% of mega-masers appear in a disk-like configuration.

Cross-matched

Sloan Digital Sky Survey(SDSS) Spectroscopy



Measurements of the host and nuclear optical emission of the SDSS galaxies are from the MPA/JHU catalogue.

Brinchmann et al. 2004, Kauffman et al, 2003, 2004

Parameter Information

z: Redshift; measure of how fast the galaxy is receding away from us.

d_L: Calculated luminosity distance from the redshift.

Vidsp: Velocity dispersion of stars in the host galaxy; measures the mass of the central black hole.

HaHb: The Balmer Ratio; A measure of the amount of obscuration along the line of sight within the galaxy.

S2R: A line ratio of two emission lines from ionized S; a measure of the density of the emitting gas.

Intrinsic luminosities are obtained from the observed luminosities and are corrected for dust obscuration using the HaHb parameter.

LogLHaObs & **LogLHaInt**: The H_α observed/intrinsic luminosity.

SFRObs & **SFRInt**: Star formation rate, obtained from the H_α emission (in luminosity).

LogLO3Obs & **LogLO3Int**: Observed/intrinsic luminosity of the doubly ionized O.

LogLO1Obs & **LogLO1Int**: Observed/Intrinsic Luminosity of the neutral O.

EddRatObs & **EddRatInt**: Eddington ratio; a measure of efficiency of black hole acceleration, based on observed/intrinsic luminosity.

M_Star: Mass of the star in entire host galaxy (in solar masses).

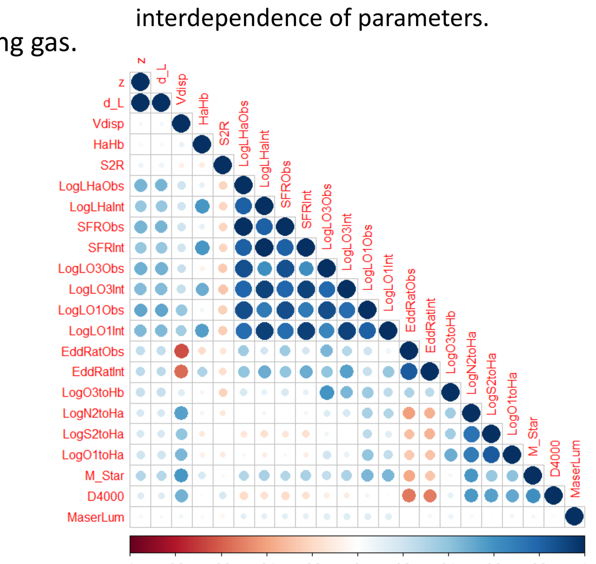
D4000: Measure of the age of the stellar population in the galaxy (based on strength of absorption features in optical spectrum).

Line-flux ratios used as diagnostics for identification of excitation by black hole acceleration:

LogO3toHb: $\frac{O(ionized)}{H}$ **LogN2toHa**: $\frac{N(ionized)}{H}$

LogS2toHa: $\frac{S(ionized)}{H}$ **LogO1toHa**: $\frac{O(neutral)}{H}$

An example correlation matrix of the dataset that reflects the interdependence of parameters.

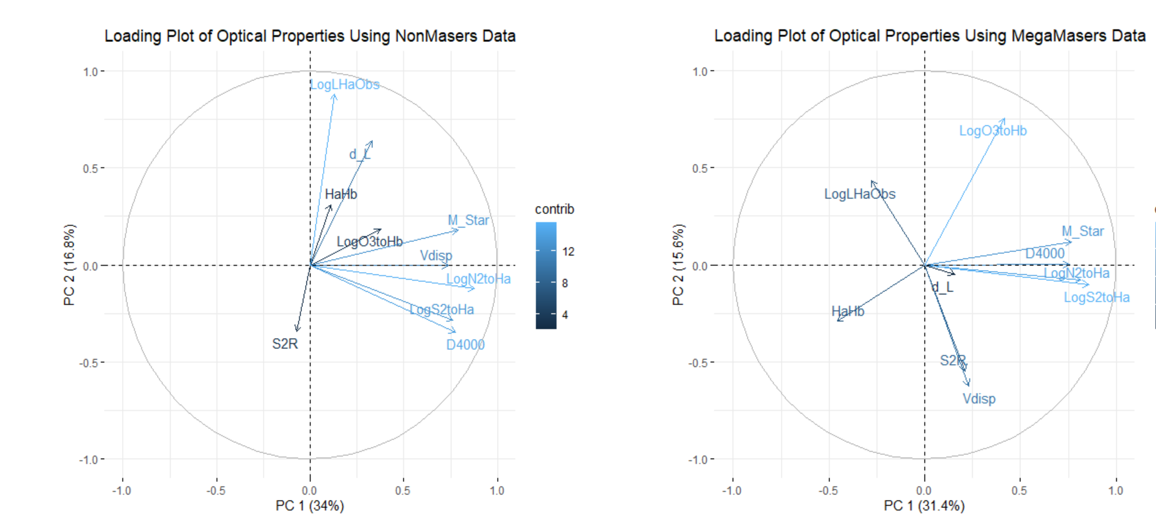


Blue indicates a strong positive correlation; Red indicates a strong negative correlation; White indicates weak/no correlation.

Results

PCA

We run PCA separately on non-maser and mega-maser galaxy samples. We present here loading plots that reflect the contributions of individual parameters (optical measurements) to the two main PCs (the x- and y-axis). Brighter colors correspond to higher percentages contributed by the parameters to the PCs. We find that the mega-masers and non-masers are governed by different types of contributions to the main PCs that account for 31-34% and 15-17% variation within the samples, respectively.



Vdisp, HaHb, S2R, LogLHaObs, and LogO3toHb arrows point in significantly different directions within the mega-maser and non-maser samples, suggesting that these parameters carry more significant information about what distinguishes between the two types of galaxies.

Random Forests

With a 60%/40% split into training and testing sets, the RF algorithm allows us to create non-maser and mega-maser classifiers based on the optical data, concentrating on the five parameters that PCA revealed as most relevant.

Initial approach

Since the fraction of mega-masers is significantly smaller than that of non-masers (~5%), the properties of the non-masers overwhelm the data in the random forest prediction; i.e., the sample is unbalanced. The accuracy of the prediction when random chance is introduced is small (Kappa = 0.16)

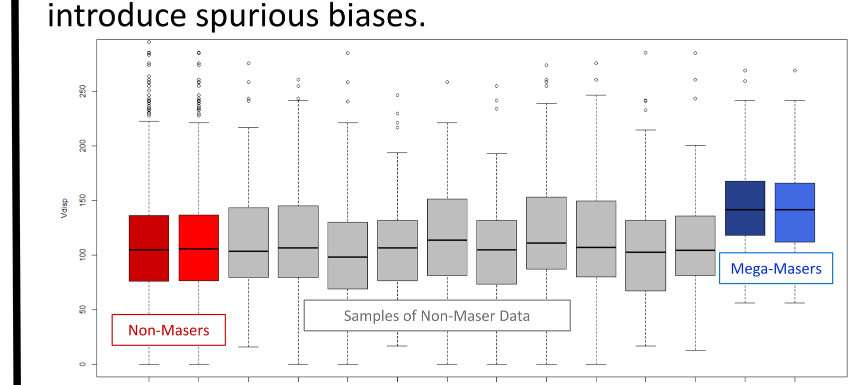
Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Landis, J. Richard, & Koch, Gary G., 1977

	Non-masers	Mega-masers
Total n (Parent Data)	1330	70
Total n (Training Data)	798	42
Total n (Test Data)	532	28
Correct Classifications within test data (%)	99%	14%

Ad-hoc approach

To address the unbalanced aspect of the RF analysis, we matched the size of the non-maser training set to that of the mega-masers. We show boxplots of matching training sets to show (for one parameter, vdisp) that the samples remain representative of the differences in mega-maser and non-maser parameters in the parent data, and thus do not introduce spurious biases.



	Non-masers	Mega-masers
Total n (Parent Data)	1330	70
Total n (Training Data)	84	42
Total n (Test Data)	532	28
Correct Classifications within test data (%)	94%	54%

The Kappa value has increased to 0.53, which indicates a significant improvement at classifying/identifying potential new mega-maser sources.

Future Directions

- Implement a SPCA approach to build a stronger classifier of mega-maser and non-maser galaxies.
- Investigate the factors that can improve the ad-hoc approach for RF to obtain a higher Kappa value.
- Design and develop a web tool with a user-friendly interface that provides mega-maser/non-maser classifications of various probability levels for any input feature set involving galactic parameters tested apriori with (published) observations to correlate with water maser emission of various morphologies.

References

Braatz, J., et al, 2009, ApJ, 695, 287; **Braatz, J**, et al, 2018; **Brinchmann, J.** et al, 2004, **Jagannath, V**, 2017, **Kauffman, G.**, et al, 2003, **Kauffman G.**, et al, 2004; **Landis, J. R.**, & Koch, G. G., 1977, **McPike**, E., 2022; **Ullah, I.**, et al, 2020, **Zierr, C.** & Biermann, P., 2018, A&A, 69, 1