Classifying Exoplanets with Machine Learning

Ana C. Barboza [1] — anacb@pm.me
S. Ulmer-Moll [1, 2] J. P. Faria [1, 2]

1 - Departamento de Fisica e Astronomia, Universidade do Porto, Portugal
2 - Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, Portugal

Main objectives

Over 4200 exoplanets have been detected so far (figure 1), and their diversity is remarkable. We aimed to determine the main types of exoplanets, develop a method that automatically associates exoplanets to their type, classifying them into labels with a machine learning algorithm.

Given the planetary mass and orbital period, we used the K-Means clustering method to classify three large groups of exoplanets: Hot Jupiters, Long Period Giants and Small Planets. In order to take into account more planetary and stellar parameters, we worked with the Uniform Manifold Approximation and Projection (UMAP) technique to visualize data on a 2D map, aiming to find structures within the high dimensional parameter space. We explored how different sets of input parameters impact the clustering of exoplanets and studied, in particular, the effect of stellar metallicity.

What we found

With UMAP, we were able to identify 5 different groups: **Hot Jupiters, Longer Period Giants, sub-Jupiters, sub-Neptunes** and **Rocky Planets**. We also analysed stellar metallicity and verified that, on average, giant planets orbit around higher metallicity stars than non giant planets.

The groups of giant exoplanets are **clearly identified** in the resulting UMAP 2D parameter space. For smaller planets, clusters were also visible but less separated. We also verified that the global structure is preserved, noticing, for example, the smaller planets ($< 8R_{\oplus}$) are grouped together and well separated from the Hot Jupiters. Adding more samples of well characterized small planets would certainly help their classification.

Quick access links (click on items)

- **Introduction & Background**: Why Machine Learning?
- **Methodology used:** K-Means, UMAP, parameter selection
- **Results:** K-Means and the 3 main groups, UMAP with 6 parameters
- **Discussion:** Discussing UMAP results, effect of stellar metallicity
- **Conclusions & References**

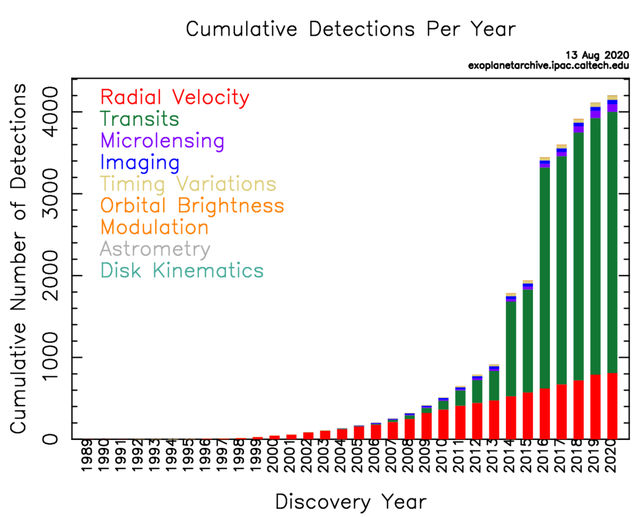

Figure 1. Exoplanet detections per year. Click to access source

.next

Introduction & Background

Exoplanetary diversity is remarkable!

- The orbital parameters of the known exoplanets span several orders of magnitude and are clearly distinct from those of the Solar system planets.
- A complete characterization is difficult: only few observables can be measured for a large number of planets.
- Recently, machine learning algorithms have been developed to estimate these parameters, namely by [Ulmer-Moll et al., 2019] for the planetary **radius** and [Tasker et al., 2020] for the planetary **mass**.
- Apart from this, new measurements will keep being performed. This means that, in the next years, more and more exoplanets will be candidates for classification.

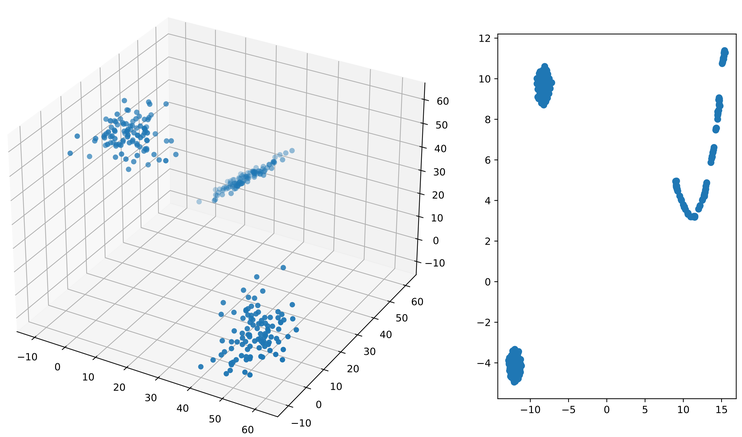

Figure 2. Example of a UMAP projection of a 3D space. Both local and global distance properties are preserved by the projection performed by the algorithm.

Machine learning algorithms are an efficient way of analysing data

- Clustering methods *automatically* identify groups of similar points in a dataset.
- For example, **K-Means** [MacQueen, 1967] aims to partition data into K groups if they are *visibly distinct*.

$\Longrightarrow$ As one can see in **figure 1** of [Laughlin, 2018], three distinct groups can be identified plotting the planetary mass as a function of orbital period, on a logarithmic scale. This calls for a clustering algorithm to perform automatic classification!

Exoplanet groups in high-dimensional space

Why stick to 2D maps? Why not look for relationships between planetary types and more than two parameters?

Data dimensionality reduction techniques allow for the visualization of high-dimensional spaces. Namely, Uniform Manifold Approximation and Projection (**UMAP**) [McInnes, Leland, 2018].

UMAP was chosen because it is based on a mathematical foundation that allows for the *preservation of **global** and **local** structure*, as one can observe in figure 2, so that:

1. distance between points is a **measure of similarity**.
2. it enables the search of **correlations** between groups and features.

.previous .next

Methods

Using orbital period and planetary mass for classification

The **K-Means** clustering algorithm is a method used to *automatically* label groups within data, if they are visibly distinct from each other. As one may see in figure 3, plotting planetary mass as a function of the orbital period on logarithmic scale, three groups are easily distinguishable.

K-Means was used in an attempt to find the following groups, as defined by [Laughlin, 2018]:

- **Hot Jupiters:** $M \sim 1 M_J$, $P \sim 3$ days.
- **Long Period Giants:** $M \sim 1 M_J$, $P > 100$ days.
- **Non-Giants:** all the other exoplanets.

Data used was taken from the Data and Analysis Center for Exoplanets (DACE).

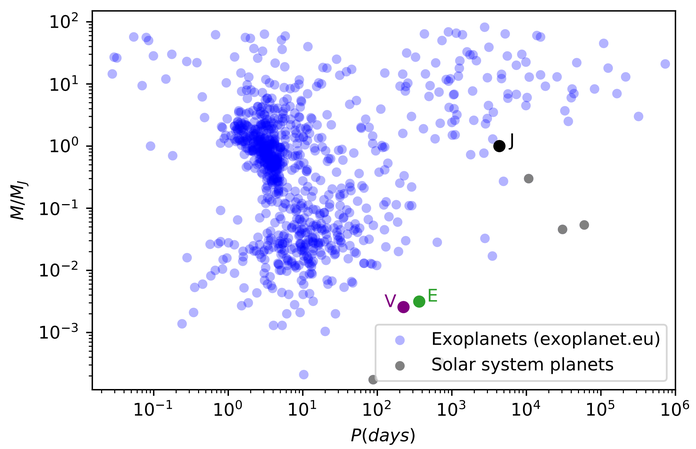

Figure 3. Exoplanetary mass as a function of orbital period. A logarithmic scale is used. Labeled points: V - Venus, E - Earth, J - Jupiter.

Adding more parameters: both stellar and planetary

UMAP allows for the visualization of high dimensional parameter spaces, conserving its key structure and projecting it on a 2D map.

- There is no guarantee that features lie on similar scales! For distances to have meaning, equally comparable values among the variables are required.
- Apart from a logarithmic scaling, all features were individually normalized between 0 and 1.

Exploring the Non-Giants, new groups can be defined following [Mousis et al., 2020]:

- **Rocky Planets** to **Super-Earths**.
- **sub-Neptunes**.
- **Neptunes** to **sub-Jupiters**.

After performing data dimensionality reduction and visualization, the 5 groups were easily distinguishable. Hand labeling was performed such that most points in each cluster were covered.

Choosing input parameters

We chose the least amount of planetary features that allow for a fuller classification, so as to *maximize* the amount of data points:

- **Planetary:** mass, radius, orbital period.
- **Stellar:** mass, radius, effective temperature, *metallicity*.

While the planetary equilibrium temperature is a good indicator of the planetary type, its calculation requires further orbital parameters, namely the semi major axis a and eccentricity e (calculating as in [Seager, 2010]). Thus reducing the amount of points for this study.

.previous .next

Exoplanet groups with K-Means and UMAP

K-Means: the three groups

Out of 905 planets with valid mass and period values, the algorithm classified: 304 Non-Giants, 509 Hot Jupiters, 92 Long Period Giants, as shown in figure 4.

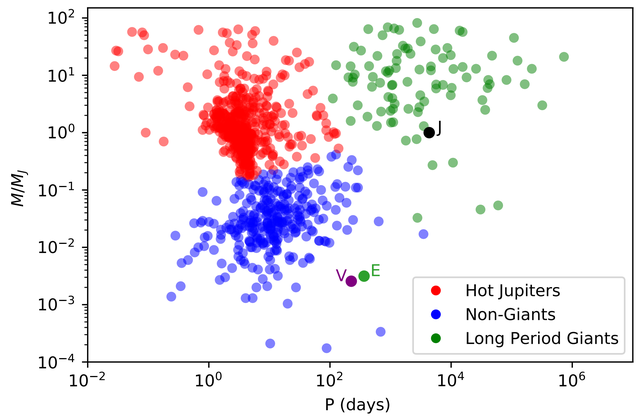

Figure 4. K-Means clustering results.'V' - Venus, 'E' - Earth, 'J' - Jupiter.

Some planets seem to be on a greyer area, since the borders of our clusters bear planets for which classification is not straightforward. For example:

- **Kepler-117 c**: M_p = 1.84 M_J, P = 50 days, T_{eq} = 710K. This planet's period is too long and its temperature too low for the Hot Jupiter group.
- **Uranus**: M_P = 0.045 M_J, P = 30571 days, T_{eq} = 63.6 K. A planet like Uranus suggests the existence of a class of exoplanets with long periods, cold temperatures and intermediate mass.

K-means was re-implemented taking only the group of Non-Giants, but no meaningful result was obtained, suggesting more parameters must be explored on the look for extra planetary types. This is where we look at UMAP!

UMAP using $\mathrm{M_*, R_*, T_*, M_p, R_p, P}$

Out of 769, 726 planets were labeled: 24 Long Period Giants, 419 Hot Jupiters, 41 sub-Jupiters, 138 sub-Neptunes and 104 Rocky Planets, grouped as in figure 5.

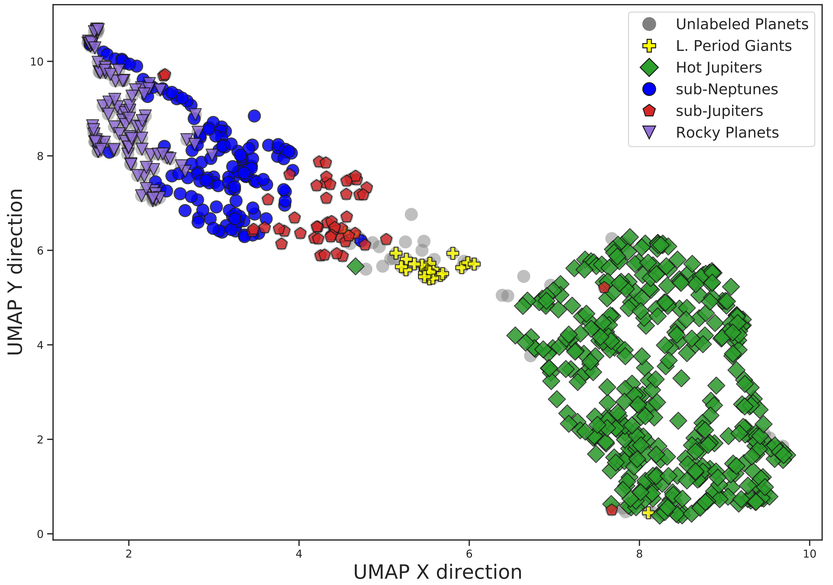

Figure 5. UMAP 2d resulting map.

- One must bear in mind that, the closer the points are, more simillar they are in terms of *all* features. It can be seen that, in order of simillarity, the groups are: **Rocky Planets, sub-Neptunes, sub-Jupiters, Longer Period Giants, Hot Jupiters**.
- Some outliers are evident, and can be further studied. For example, **Kepler-56 b** is the sub-Jupiter outlier on the upper Hot Jupiter group. This may be because its host star is relatively luminous (8.8 $L_\odot$), so the algorithm recognized this difference.

.previous .next

Correlations between features and stellar metallicity

Exploring correlations between groups and features

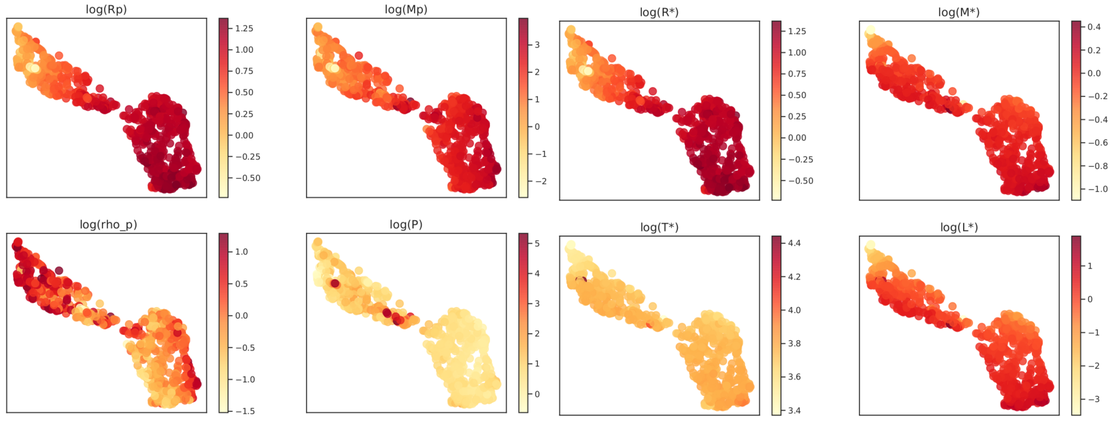

Figure 6. Values on a logarithmic scale. The map is colored as a function of several planetary and stellar parameters. From left to right, top to bottom: radius, mass, density, period. Density and Luminosity were computed but not set as an input feature.

- **Rocky Planets** show the largest density of all groups (density shown in the fifth panel of figure 6). In average $1 gcm^{-3}$ higher than the one of Earth, suggesting a higher iron composition. They seem to orbit around the least massive and luminous stars.
- **sub-Neptunes**' density centers around 2.84 gcm^{-1} which goes according to what is established in [Mousis et al., 2020].
- **sub-Jupiters**' density suggests they likely are gaseous planets. Neptune-like planets whose periods are less than 10 days are often referred to as "Hot Neptunes", This subtype did have its individual cluster, indicating that planets who fall within this radius range orbit similar stars, and are of similar nature.
- The lower orbital period that best fit the **Longer period Giants** cluster was of around 25 days. The central value of the distribution, however, was of 113 days. This reflects the lack of measurements for planets with longer orbital periods.
- **Hot Jupiters**, apart from being the group with the lowest central planetary density, seem to orbit around the largest, more luminous and massive stars.

Effect of stellar metallicity

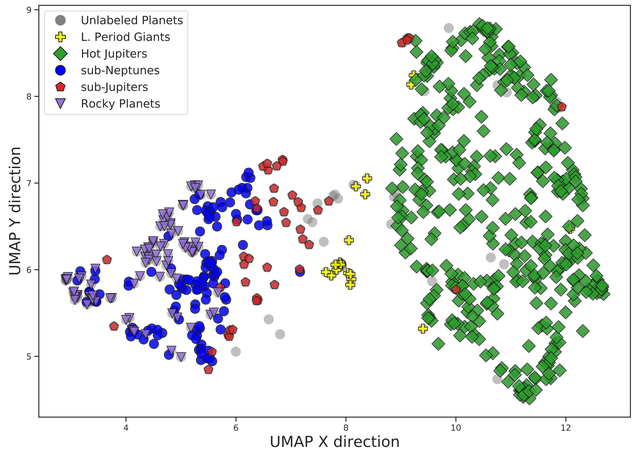

Figure 7. Metallicity was then used as an extra input parameter.

- After adding stellar metallicity as an extra input parameter, the groups of Non-Giant planets had their structure harmed (figure 7). This suggests a weaker correlation between this feature and these groups of planets. In figure 8, this map colored as a function of metallicity.

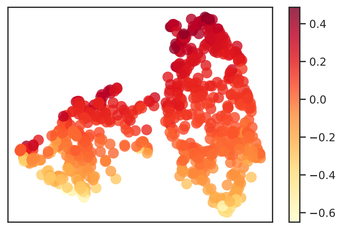

Figure 8. Resulting map colored as a function of metallicity

As a validation method, we calculated the average metallicity for giants and non giants, hoping to find the known giant-metallicity relationship [Santos, et al., 2004], resulting in:

- Giants: $[Fe/H] = 0.0789$
- Non-Giants: $[Fe/H] = -0.0007$

So this correlation is verified.

.previous .next

Conclusion & References

Limitations

- Our study was limited by the completeness of the exoplanet database. On the UMAP study, initially a group of “dense planets” was identified, which we later found to be outdated upper mass limits.
- Another limitation is that it mixes different measurement techniques, both on stellar and planetary parameters.

Conclusion

- Two machine learning algorithms (K-means and UMAP) were used to study exoplanet parameters.
- Using orbital period and planetary mass, the K-Means algorithm leads to the classification in three different groups: Hot Jupiters, Long Period Giants and Non-Giants.
- To study the impact of other parameters, the UMAP dimensionality reduction method was used. This resulted in the classification in five different groups. In order of similarity: Hot Jupiters, Longer Period Giants, sub-Jupiters, sub-Neptunes and Rocky Planets.
- The giant planet-metallicity correlation was explored, and was verified, since Long Period Giants and Hot Jupiters showed an average metallicity higher than the non giant planets.

Further work

- New groups could be identified by adding extra parameters. A good example could include Rocky Planets around M-Dwarves (upper left corner of the first map)
- A clustering algorithm could also be performed for automatic classification on the output of UMAP, however, the groups do not have the same amount of samples, limiting this task.

References

[Laughlin, 2018] Laughlin, G. (2018).
Mass-Radius Relations of Giant Planets: The Radius Anomaly and Interior Models.
In Deeg, H. J. and Belmonte, J. A., editors, *Handbook of Exoplanets*, pages 1–17. Springer International Publishing, Cham.

[MacQueen, 1967] MacQueen, J. (1967).
Some methods for classification and analysis of multivariate observations.
The Regents of the University of California.
ISSN: 0097-0433.

[McInnes, Leland, 2018] McInnes, Leland (2018).
How UMAP Works — umap 0.4 documentation.

[Mousis et al., 2020] Mousis, O., Deleuil, M., Aguichine, A., Marcq, E., Naar, J., Aguirre, L. A., Brugger, B., and Goncalves, T. (2020).
Irradiated ocean planets bridge super-Earth and sub-Neptune populations.
arXiv:2002.05243 [astro-ph].
arXiv: 2002.05243.

[Seager, 2010] Seager, S. (2010).
Exoplanet Atmospheres: Physical Processes.
Princeton University Press.

[Tasker et al., 2020] Tasker, E. J., Laneuville, M., and Guttenberg, N. (2020).
Estimating Planetary Mass with Deep Learning.
The Astronomical Journal, 159:41.

[Ulmer-Moll et al., 2019] Ulmer-Moll, S., Santos, N. C., Figueira, P., Brinchmann, J., and Faria, J. P. (2019).
Beyond the exoplanet mass-radius relation.
Astronomy & Astrophysics, 630:A135.
arXiv: 1909.07392.

More information and scripts used are available at the project’s **github repository**.

EUROPLANET SCIENCE CONGRESS OSPA

This poster participates in the

OSPA

Outstanding Student Poster Award Contest

previous