PHSX815_Project4:
Exoplanet Populations

Yoni Brande

May 1, 2021

1 Abstract

We use clustering algorithms to identify subgroups of exoplanet populations, and try to determine whether our clusters are physical or artifacts from observational gaps.

2 Introduction

Over 4000 exoplanets have been found since the first planets outside our solar system were discovered in the early 1990s **(Wolsczan, Mayor)**. Of the 4383 confirmed planets, 3354 have been discovered by the transit method, observing periodic characteristic dips in the brightness of a host star as the planet moves in front of the star's disk in our line of sight. These discoveries have mostly been driven by space-based transit surveys such as the wildly successful Kepler, K2, and TESS missions **(Borucki, Ricker)**, as well as ground-based surveys such as MEarth, KELT, and HAT (cite cite cite). Radial velocity surveys have discovered the next largest sample of exoplanets, and other methods have discovered the rest.

Now that we are in an era of statistically useful exoplanet samples, we are able to do population-level studies of exoplanet properties. For example, one of the first unexpected populations noticed were the Hot Jupiters, large gaseous exoplanets orbiting very close to their host stars. This challenged conventional wisdom on planet formation and evolution, and gave us a more complete understanding of the astrophysical processes taking place in planetary systems. Unfortunately, our observational capabilities are limited, and we can't observe every planet (or small body) that possibly exists. This leads to a fundamental conflict in our population studies: Are we observing a truly representative sample, and are the trends we exist physical? Or are we just looking at the effects of observational biases? Here we attempt to present exoplanet population data in ways that allow us to find real clusters of similar exoplanets.

3 Exoplanet Data

Different exoplanet discovery methods rely on different ways to observe planets either directly or indirectly. The transit method measures periodic dips in the brightness of a star as the planet transit's the star's disk, giving us a measurement of the relative areas of the planet's and star's disk. By knowing the size of a star, we easily find the size of the planet. The period of the signal gives us the size of the planet's orbit, and other orbital parameters can be determined by more subtle aspects of the transit signal. However, this gives us no information on the planet's mass, so we look to other detection methods, like the Radial Velocity method. Since a planet and its host star

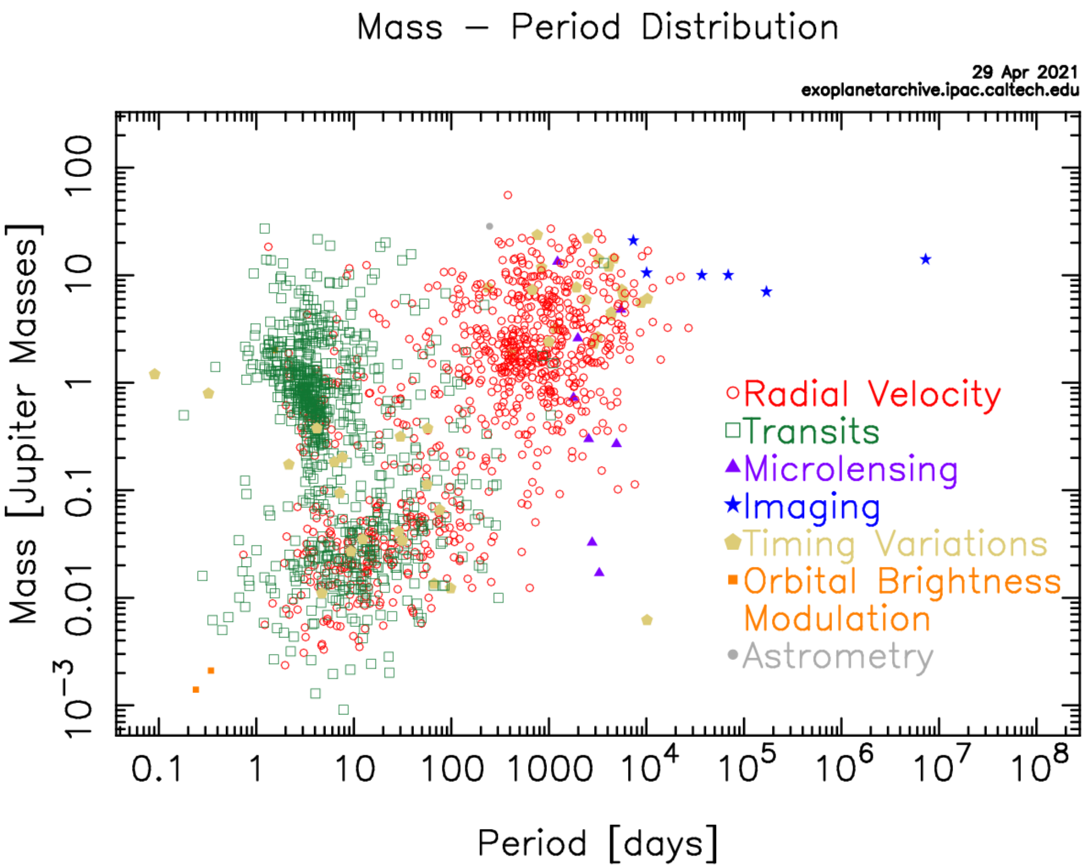

Figure 1: Exoplanet Mass Period Plots

both orbit a common barycenter, we can look for the effects of the motion of the star around that barycenter. As the star moves towards and away from us along our line of sight, we can observe the red- and blueshifting of the star's spectrum, and use the strength of the Doppler signal to determine the mass of the planet inducing it. Both transits and radial velocities are most sensitive to large planets on short orbits, and since transiting planets all have inclinations near 90°, many of them are also amenable to radial velocity observations, giving us both mass and radius (and thus density and bulk composition) for many observed exoplanets.

Generally, when presenting data for most exoplanets, we prefer to rely on radius, mass, and orbital period as the basic parameters. Mass-period and radius-period distributions are common presentations, and mass-radius distributions allow for curves of constant density to be plotted. By coloring each planet with its discovery method, we see some natural clusters appear.

In the mass-period plot (Fig. 1), we see that there are two distinct cluster of transit and radial velocity planets, plus possibly a third cluster of mixed detection methods. In the radius-period plot (Fig. 2), we mostly only have data for transiting planets, but we still see a main cluster near the bottom, and a secondary cluster spanning about 10 Earth radii near 5 days.

4 Clustering Algorithms

Given a multidimensional set of data, we can use clustering algorithms to find subsets of data points that are all more similar to each other than they are to data outside their clusters. Here, we adopt

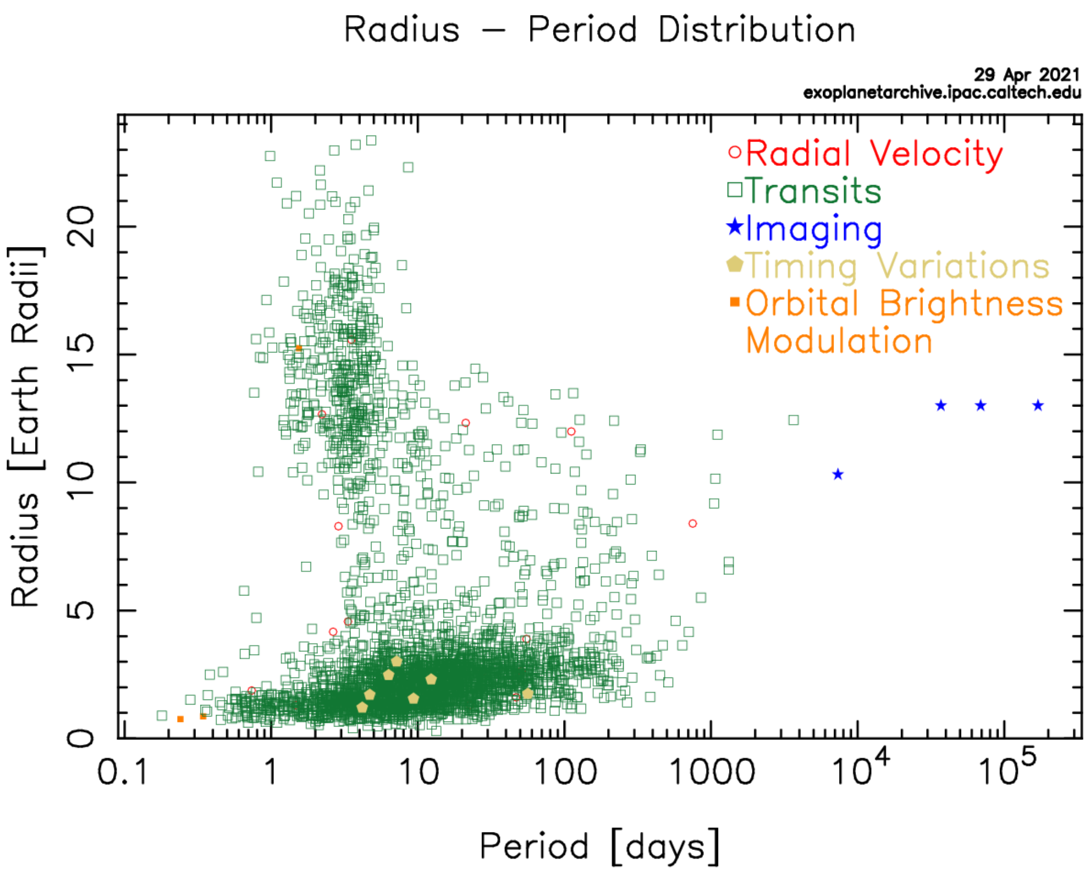

Figure 2: Exoplanet Radius Period Plots

the use of k-means clustering, a centroid-based method that assigns points to one of K possible clusters given K starting points, and iteratively reassigns points as the cluster centers move to some steady state. This minimizes the variance within each cluster, and when the cluster centers no longer move, the solution has converged to the minimum-variance solution. Of course, not every possible number of clusters will be useful, as for a dataset with N points, you might expect to be able to assign the dataset to any number of clusters in the range $[1, N]$. Some care will be needed to not overfit the data with too many possible clusters, but we must also be careful not to underfit the data, as too few clusters will also be uninformative.

(PUT MATH OF K-MEAN ALG HERE)

For data where it may be expected that the populations will have some overlap, we may also use expectation maximization to assign points to one or several overlapping Gaussian distributions. This Gaussian mixture method may be more versatile than K-means as it can account for points that somewhat fit both of two separate distributions, like a trough in a bimodal distribution, rather than placing a hard border and assigning them to one or the other cluster. Each cluster then becomes a Gaussian function with its centers and variances, and as the algorithm iterates, those centers and variances should converge to some maximum expectation value.

(PUT MATH OF EXPECTATION MAX. HERE)

We will compare the use of these two methods to sort exoplanets into varying populations.

5 Analysis

We ran the algorithms and here's what we found.

stuff

6 Conclusions

There are X populations of planets, in certain places. There are Y voids, possibly due to observational biases or something physical.