Visualizing Gaia data

Kosti Koistinen, 518223

2 12 2021

In this exercise a random sample of GAIA space telescope data was gathered, processed and visualized. The purpose of the exercise was to practice big data handling and data visualizing. The report is divided into 5 parts: Introduction, data gathering, visualization, error estimation and conclusions. The code of the data analysis in R is in Rmd-file.

Introduction

Stellar distributions

'The distribution of stars in Milky Way has been a point of interest for a very long time. There are several sky mapping projects, which have been surveying the stellar distributions in Milky Way. In fact, the mapping of stars - statistical astronomy - is probably the oldest form of astronomy. Ptolemy introduced sky charts first over 2000 years ago, with information of stellar coordinates and their relative brightnesses, the apparent magnitudes.

One of the most recent surveys was a mission conducted by ESA's space telescope Gaia. It surveyed the skies for almost 12 years, detecting over 3 billion Milky way stars [**1**]. That is only few percent of the total amount of stars in Milky Way, but the results are significant. Using measurements of stellar parallaxes obtained from Gaia photometric data a lot of physical parameters could be determined, one of which are star effective temperature and distance. In this excercice was studied how well a sample of Gaia's data represents the real distribution of stars.

Different stellar types

Stars can be classified by many different parameters. Ptolemy classified stars by their brightness, and the method is still valid. By investigating the brightness and parallax of the star one can estimate the stellar surface temperature and it's total luminosity. Star's size and type can then be determined.

In the beginning of 20th century, astronomer Ejnar Hertzsprung noticed, that stars can be divided into different groups [**2**]. While studying a sample of stars, he noticed, that most stars lie near the main sequence. It can be visualized by plotting the luminosity of a star versus it's surface temperature. Because most stars are in main sequence, it can be deducted that the stars spend there most of their life span. As the stars grow older, they radius increases, and their luminosity increases while the surface temperature decreases: They exit the main sequence. For Sun-like stars, the stars go through a Red Giant phase, and after the star has "died", it falls into White dwarf group. The white dwarfs are very hot, but faint remnants of stars.

The Red Dwarf - type stars are the most common type of stars in the Universe [**3**]. Approximately two thirds of Milky way stars are red dwarfs. They are small,cool and faint stars, with effective temperature (surface temperature) ranging from 2700 K to 4000 K. The red dwarfs are the coolest stars in the main sequence. Due to low rate of fusion processes in red dwarfs, they live much longer than Sun-like stars. They can stay in main sequence for tens of billions of years.

The goal in this task is to study and visualize the sample of a star set similarly as Hertzsprung did over hundred years ago. The aim is to find out, whether the distribution of stars in their classes represent the reality. There should be a majority of red dwarf stars in the sample, because they are the most common stars in Milky Way. Also one would expect to find a lot of stars near by and fewer, hotter stars further away.

Data gathering

The 50000 star sample was generated from 1.3 billion target catalogue, which included the geometric distances [4]. The sample was made in the ARI website: *http://gaia.ari.uni-heidelberg.de/tap.html*. The Sample was created with following ADQL query:

select top 50000 source_id, ra, parallax,parallax_error,dec, phot_g_mean_mag,phot_g_mean_mag + (5 * log10(parallax)- 10) as g_mag_abs, r_est, r_lo, r_hi, teff_val
FROM gaiadr2_complements.geometric_distance
JOIN gaiadr2.gaia_source USING (source_id)
WHERE parallax>0
order by random_index

Note the absolute g-magnitude variable *M_mag* which is created from the parallax and apparent magnitude (eq.1), where M_{abs} is the absolute magnitude, m_{app} is the apparent magnitude and Px is the parallax of an object.

$$M_{\text{abs}} = m_{\text{app}} + 5 \cdot \log_{10}(Px) - 10 \tag{1}$$

The data was then downloaded as CSV and processed in excel. The dataframe was created with Rstudio readxl-package. Because the point of interest was to study the stellar types and distances of the stars, all the rows that had missing values in those columns was deleted. 7725 entries was left (Table 1).

Table 1: Few rows of data and selected information of the objects

ID	Px	P-err	RA	DEC	m mag	M mag	Dist. [pc]	+	-	Eff. temp. [K]
4.205573e+18	0.34	0.26	285.00	-7.21	18.25	5.88	3249.43	1811.68	6386.40	NA
4.205566e+18	0.25	0.12	285.40	-7.15	17.28	4.23	3673.27	2472.06	6118.24	NA
4.053207e+18	1.05	0.90	275.63	-25.13	19.64	9.76	3396.41	1289.40	7019.17	NA
4.049399e+18	0.26	0.41	274.01	-30.53	19.24	6.31	4405.74	2109.81	8552.42	NA
5.889060e+18	0.23	0.07	234.27	-52.22	16.64	3.46	3679.61	2919.37	4873.74	4368.46
2.026241e+18	0.39	0.08	293.52	28.53	16.99	4.92	2494.96	2033.38	3213.57	5070.04
5.889051e+18	0.33	0.09	233.82	-52.26	16.84	4.40	2820.40	2204.55	3838.54	4361.00
4.049411e+18	0.94	0.70	274.10	-30.49	19.69	9.55	3723.32	1360.22	7904.17	NA
6.335235e+18	0.54	0.38	228.13	-4.90	19.16	7.82	1905.89	1121.19	3585.56	NA
1.870129e+18	0.57	0.06	310.98	34.49	16.42	5.19	1681.60	1526.27	1871.02	4919.00

Table 1 column specifications:
1. (ID): Target identification number in GAIA catalog.
2. (Px): Parallax of the object in arc sec.
3. (P-err): Standard error of the parallax in arc sec.
4. (RA): Rectascension of the object in minutes.
5. (DEC): Declination of the object in minutes.
6. (m m): Measured g mean magnitude.
7. (M mag): Calculated absolute magnitude from parallax and g mean magnitude.
8. (Dist.): Distance of an object in parsecs.
9. (+): Error of distance.
10. (-): Error of distance.
11. Eff. temp: Effective temperature of the object in Kelvins.

Visualization

Herzsprung-Russell Diagram

HR-diagram is a very useful plot to investigate different stellar populations. Hertzsprung-Russell diagram has luminosity or magnitude in y-axis, and effective temperature or spectral class in x-axis. The HR-diagram is plotted in figure 1.

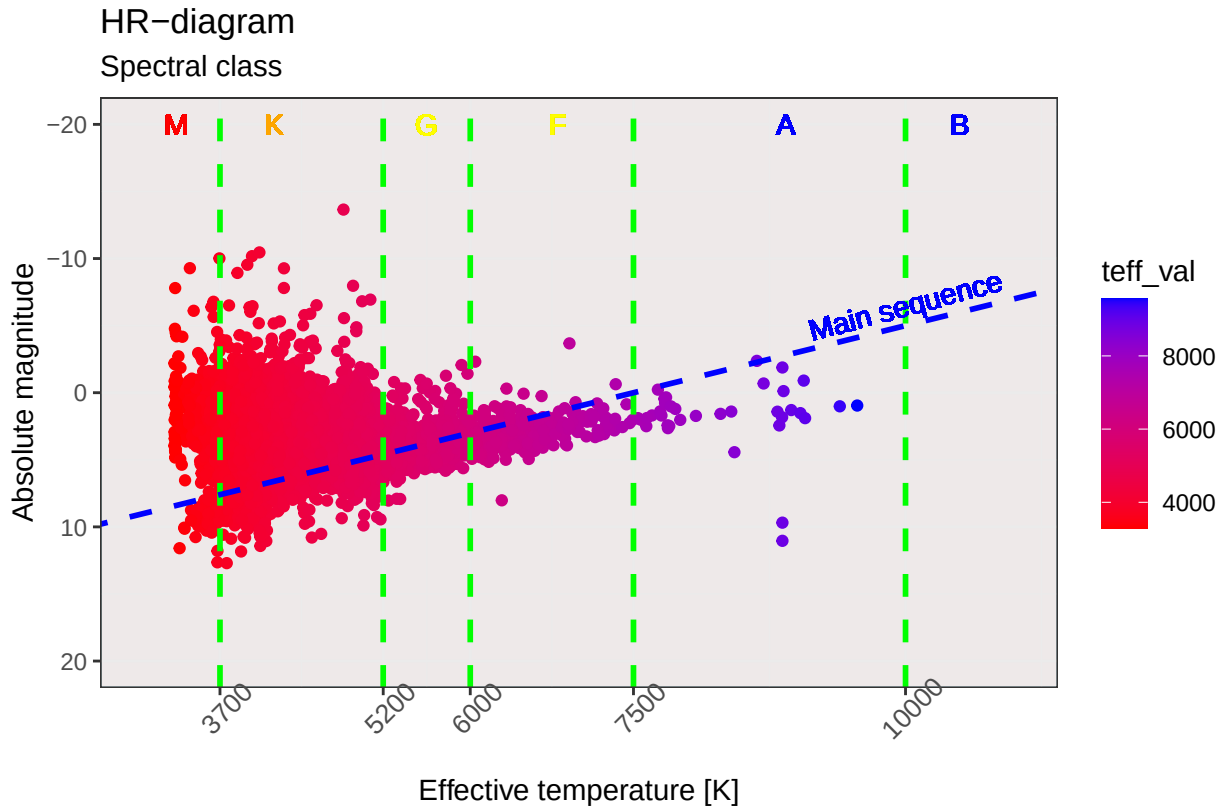


Fig.1 Hertzprung−Russell diagram. The main sequence is estimated with a dashed line.

The Main sequence of stars is clearly visible, most of the stars are distributed along the line. The line is synthetic and the approach is very heuristic. It is plotted just to illustrate data trend being around the main sequence, as expected. The areas of plot are divided by spectral classes O,B,A,F,G,K,M. They describe the peak of the spectrum of a star in visible wavelenghts. The red dwarfs are cooler, and therefore, by Wien's law, they emit light in lower wavelengths than brighter, hotter stars [**5**]. By comparison, Sun's effective temperature is 5800 K, and it belongs to spectral class G with absolute magnitude of 4.8. Thus the Sun belongs to the main sequence.

Bottom left is the area of coolest and smallest stars, the Red Dwarfs. They are in main sequence (a linear plot does not represent the data well in the borders of spectral classes.) Top left are the red Giants. Because size of the stars remain unknown, it is impossible to draw an exact line on the types of stars. The Absolute magnitude is an indicator. The large red giants are very luminous compared to dwarf stars.

What else can be learned from HR-diagram? One can notice two dots bottom right. They are probably remnants of the stars, White dwarfs. It is a phase of which all medium sized K-A type stars end up to. There are also objects such as bright giant stars, very luminous and short life hot stars, but unfortunately, the sample does not contain such objects. That could be due to the occurrence of such objects: they are rare in Milky Way [**3**].

Already it was very clear that most of the stars are not distributed as expected: There are only few Red Dwarfs compared to other, larger and hotter stars (figure 2). Why? Either the sample is very bad, or the

resolution of GAIA's camera is not efficient enough to detect dim stars. This can be checked with parameter distance. If there is correlation between the magnitude and distance of the star, it suggests, that only larger stars are detected with larger distances. Also it can be stated, that because most of the stars seem to group near the main sequence, the data is reliable. In reality, most of the stars are in main sequence, because in the life span of a star they spend most of their time there.

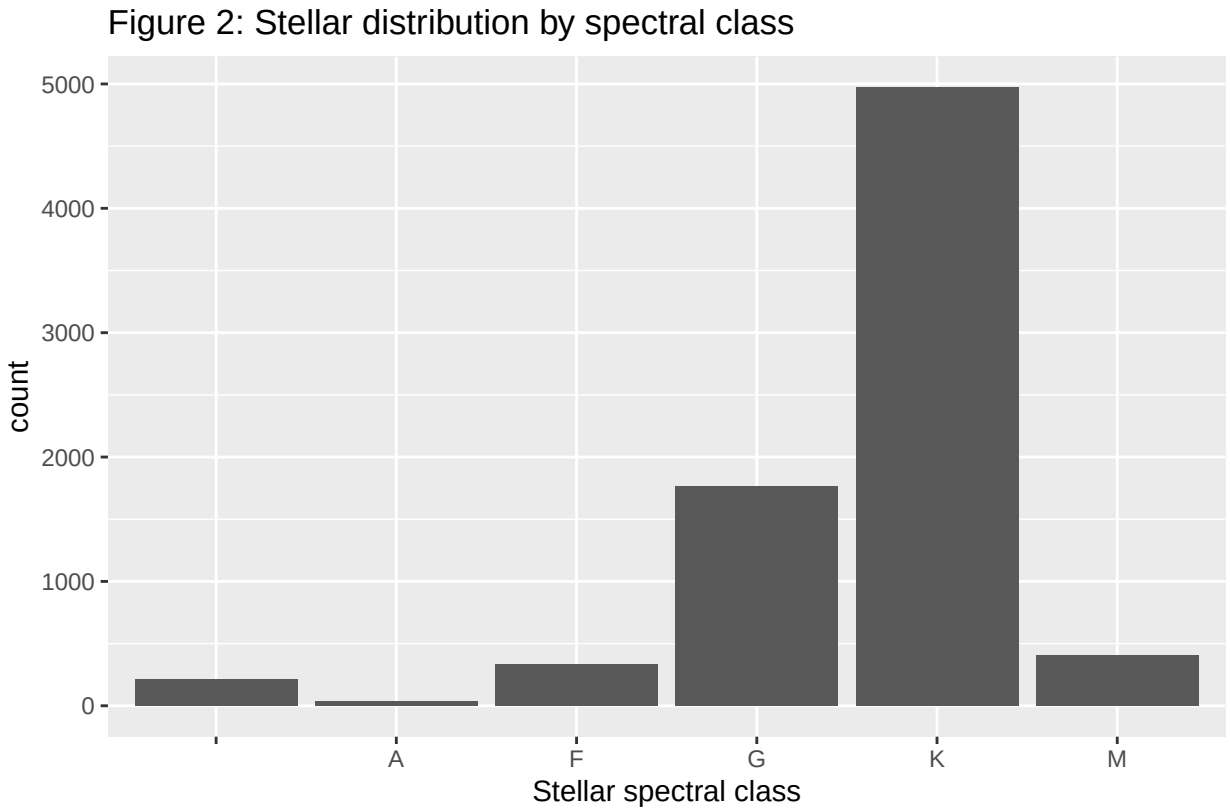


Fig 2: The G and K−type stars dominate the sample.

In figure 3, the frequency of different spectral classes as a function of distance is plotted. Most of the stars are within few thousand parsecs, but as expected, the brightest stars dominate the distribution in larger distances. Where are the red dwarfs then?

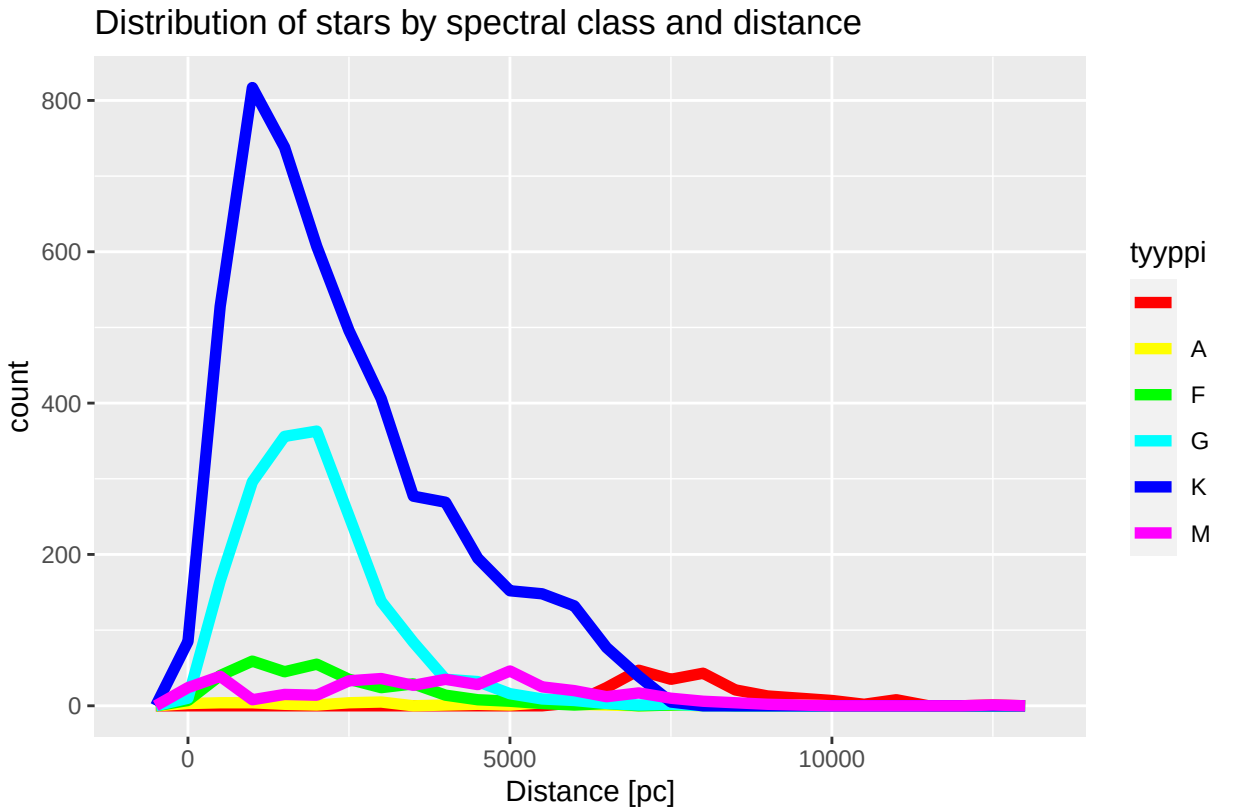
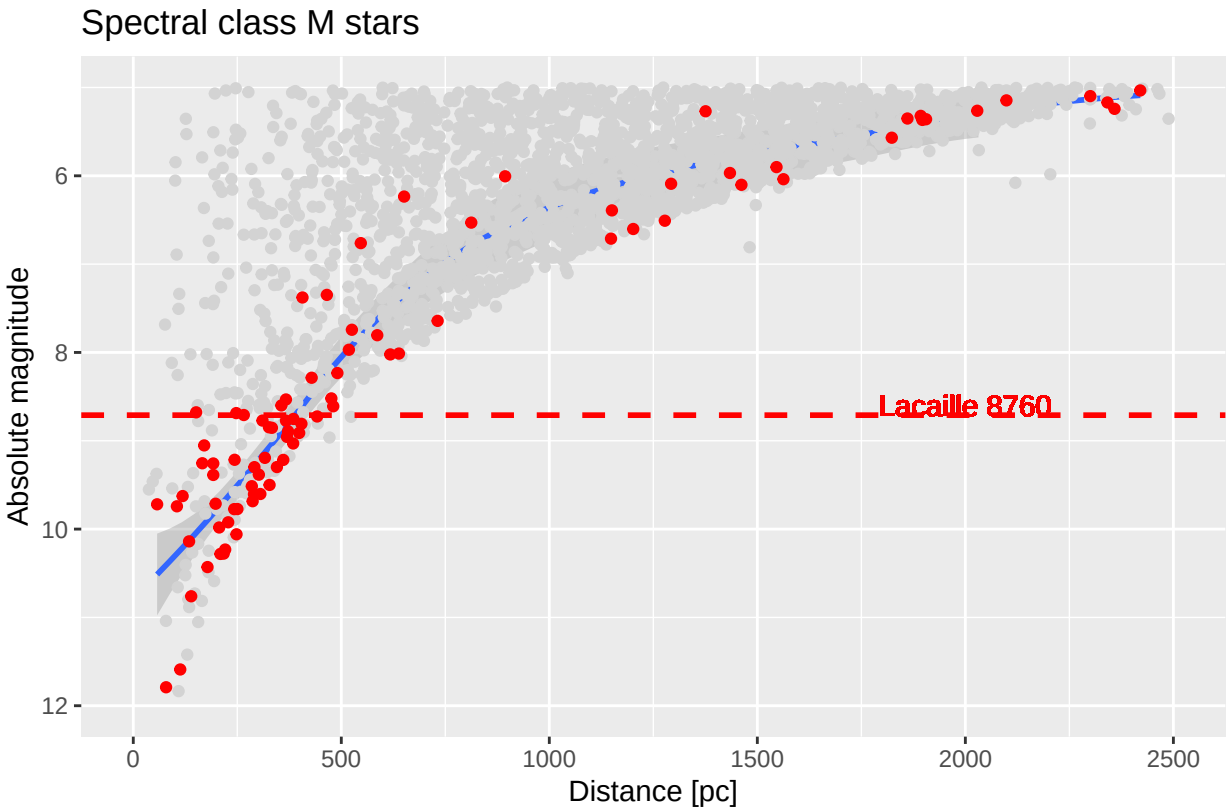


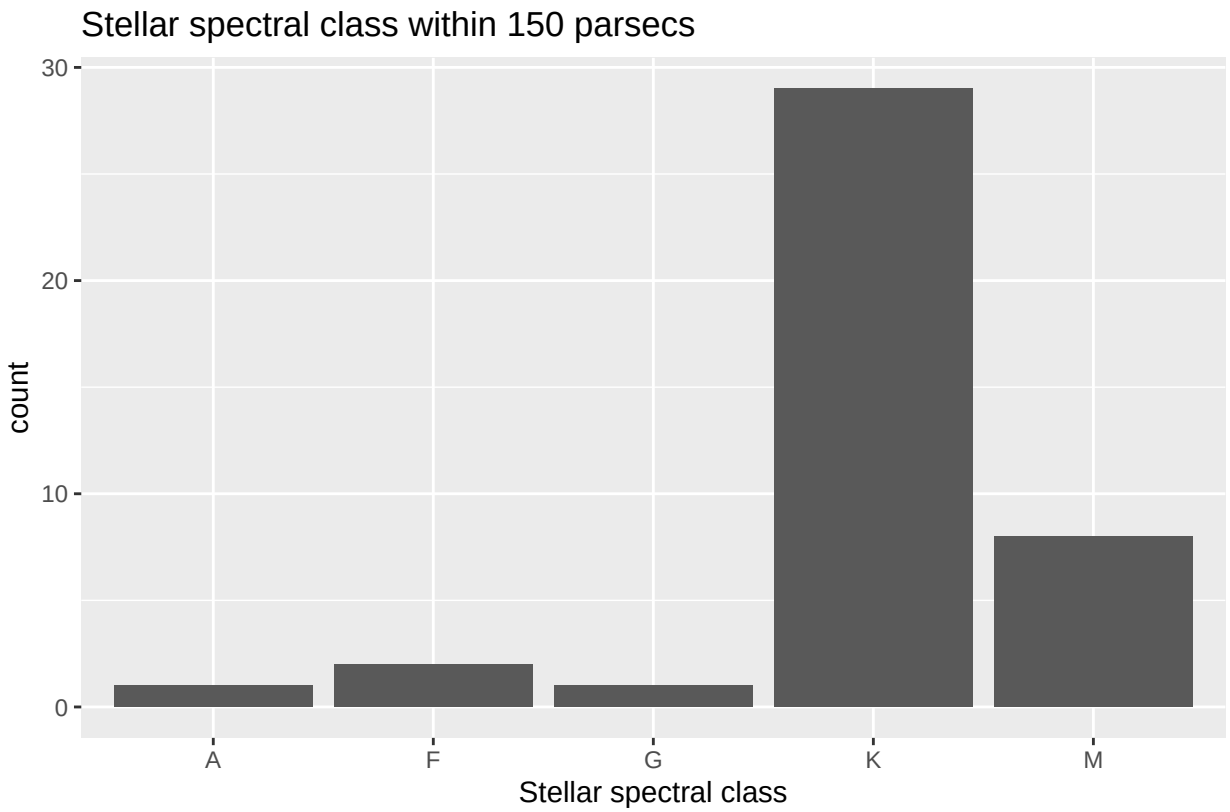
Fig.3 The G and K type stars dominate the sample within all distances.

In figure 4, the "near space" distribution of stars is plotted in order to find the missing red dwarfs. The M-type stars within 3000 pc are plotted. The tick line represents the absolute magnitude of the brightest Red Dwarf known, The Lacaille 8760 [**6**]. Therefore, it can be estimated, that probably most of the stars under this threshold are Red dwarfs. All the rest are Red giant stars (well above the main sequence.) This can be even checked – See figure 6. The red giants are definitely off the main sequence.



M−type stars (red) separated from sample.

The red dwarfs are still missing: while searching them with various visualizations, they still remain hidden. One would expect that the dwarfs become abundant within even smaller distances than in previous figure. Indeed, figure 5 suggests a rise in abundance of red dwarfs within smaller distances. The brighter stars still dominate the sample, but it's clear that red dwarfs become abundant within small distances. Therefore, it can be guessed that red dwarfs are abundant in Milky Way, but are not visible in Gaia data. This hypothesis could be checked by downloading all the Gaia data from nearby stars (~500 pc:S), and calculate the ratio (maybe in the next excercise).



The red dwarfs become more abundant in small distances

Errors

A very short consideration of errors was then conducted. One must first consider the goodness of dataset. It was truly a random sample, but 84% of the data was discarded due to lack of information about effective temperature. This is a key issue: The effective temperature can only be calculated, if the distance of the star can be calculated. GAIA measurements are based on parallax. If the measurent of a parallax is not good enough, the estimate of distance nor the effective temperature can be established. One could ask: Are the faint and active red dwarfs exactly such objects? They are faint and highly variable. Maybe GAIA can't measure their parallax very accurately. This could explain the lack of red dwarfs in data set: maybe they were discarded already because of missing effective temperature values!

The errors of few objects distances and magnitudes are plotted in figure 7. Only a small sample was used to demonstrate how absolute errors contribute to distances and magnitudes. Within small distances, the errors are rather small. Thus it it can be said that red dwarfs truly are red dwarfs. As the errors grow larger, the type of a star becomes unclear. The absolute magnitude depends on the logarithm base 10 parallax(~distance) of the object. Therefore, both errors distance and absolute magnitude are related. The error is calculated with equation 2. Error of apparent magnitude was not considered this time, only error of parallax was used. The P_{err} is the error of parallax.

$$\Delta M_{\text{abs}} = m_{\text{app}} + 5 \cdot \log_{10}(Px \pm P_{\text{err}}) - 10 \tag{2}$$

Figure 7: The errors of the distances (and magnitudes)

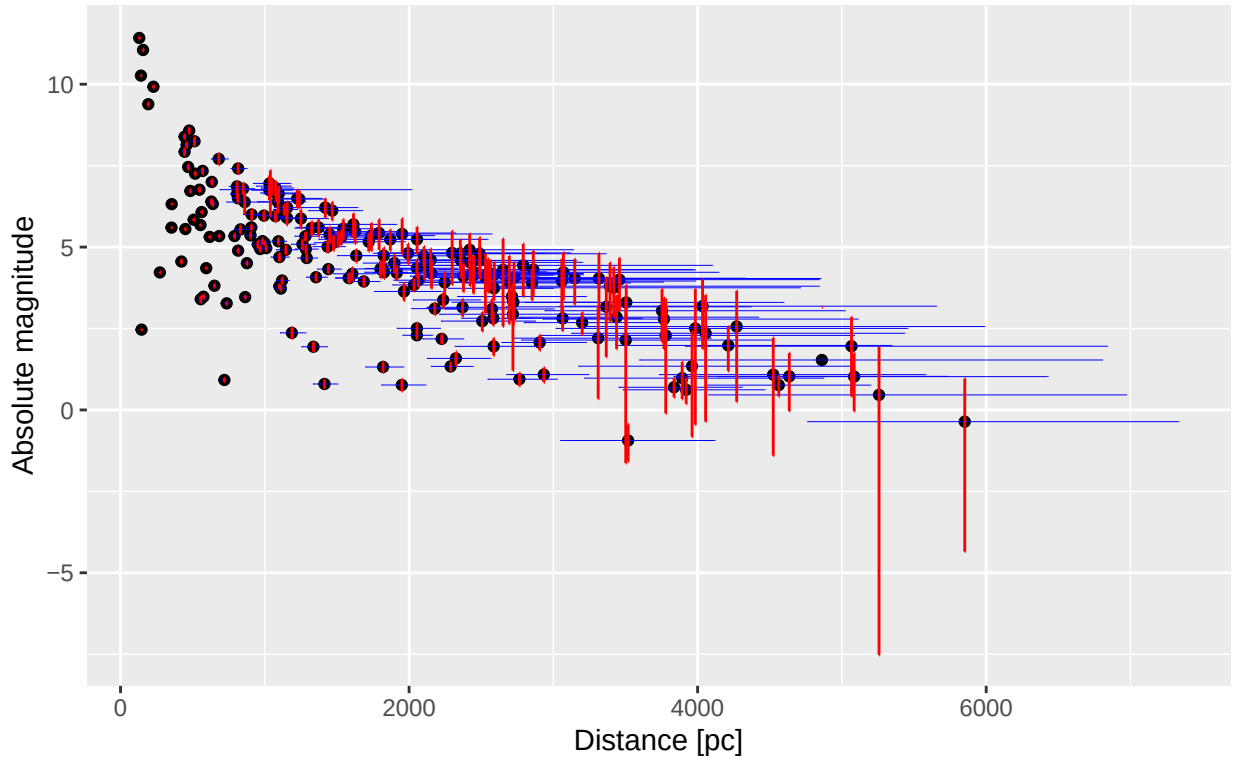


Fig 7: Errors of observations grow in large distances

Conclusions

1. The K type stars are most abundant in data not because they are abundant in Milky Way, but as they are most abundant when compared to total luminosity.

2. Only near by Red dwarfs are discovered. The distribution doesn't represent reality. Luminous objects are more likely to be discovered.

3. There are also a lot of red dwarfs discovered, but their effective temperature is harder to define because errors in parallaxes are probably large. This could not be checked because a lot of data had missing values.

4. Most of the detected stars are in main sequence as predicted. Few bright giants are seen, but most of the giants are red giants, as expected.

5. Only two white dwarfs are in data set. However, they are very faint, hot objects, so only near by white dwarfs are detected. In Milky Way, 10 billion White dwarfs. That is ~10% of all stars in Milky Way. However, they are very faint, and that's why they are not detected.

References:

[**1**] European Space Agency, GAIA mission, <*https://sci.esa.int/web/gaia*>
[**2**] Hertzprung, Ejnar (1908). "Über die Sterne der Unterabteilung c und ac nach der Spektralklassifikation von Antonia C. Maury". *Astronomische Nachrichten.* **179** (24): 373–380. doi:10.1002/asna.19081792402.
[**3**] Ledrew, Glenn (February 2001). "The Real Starry Sky". *Journal of the Royal Astronomical Society of Canada.* **95**: 32. Bibcode:2001JRASC..95...32L
[**4**] Bailer-Jones, C.A.L, Rybizki,J et al.: "Estimating distances from parallaxes IV: Distances to 1.33 billion stars in Gaia Data Release 2"*AJ* **158**, 58 (2018)
[**5**] Fritzsche, Hellmut. "Wien's law". Encyclopedia Britannica, <*https://www.britannica.com/science/Wiens-law*>.
[**6**] Lacaille 8760, Universeguide, <*https://www.universeguide.com/star/105090/lacaille8760*>