Multivariate Statistical Analysis Project work

Severi Rissanen

March 2019

1 The Data Set

I chose as my data set a collection of forest fire observations made at the Montesinho national park in Portugal collected from January 2000 to December 2003. The data set originally appeared in the paper [1], where they used it for a regression model predicting forest fires and their sizes. The data consists of Canadian Fire Weather Index (FWI) values, weekdays, months, locations in a 9×9 grid, temperature, wind and rain values as well as the total burned areas for all recorded forest fires. The used FWI values were the Fine Fuel Moisture Code (FFMC), describing the moisture content of fine fuels (like grass, twigs and pine needles), the Duff Moisture Code (DMC), which measures "the moisture content of loosely compacted organic layers of moderate depth", the Drought Code (DC), rating "the moisture content of deep, compact organic layers" and the Initial Spread index (ISI), which quantifies the expected fire spread rate [2]. The FWI values are defined so that higher numbers mean more dryness and higher chances of forest fires.

2 Research questions

The aim of this project was to try to get some insight in to the conditions where forest fires start in different settings. Especially, the spatial and temporal distribution of the forest fires was of interest, and I decided to analyze and compare the conditions in different months further with multivariate analysis. In Sec. 3, univariate and bivariate analysis is done for the entire data set, while Sec. 4 gives the results of multiple correspondence analysis on certain parts of the data. Section 5 concludes the report and gives some critical evaluations.

3 Univariate and Bivariate Analysis

Table 1 summarizes some location and scatter estimates for the numerical variables. Out of the nine variables, two stand out somewhat. The 1st quartile, median and 3rd quartile of the amount of rain on days with forest fires are all zero, indicating that no rain was present on most of the days. Indeed, upon closer inspection, only 8 out of the 517 recorded forest fires had a non-zero amount of rain in the same day. Of course, this makes sense considering that rain probably stops forest fires from happening quite effectively, but it also means that the variable is not very useful, so I decided to discard it in further analysis. The first quartile of the burned area-variable is also zero, and while the median is only at 0,52, the mean is at 12,9. This and the fact that the standard deviation is also very high compared to the mean, tell us that the data must be quite heavily skewed and heavy-tailed. This is also seen from the histogram in Fig. 1. The 1st quartile being at zero is explained by the fact that apparently burned areas under a certain threshold were not measured accurately and instead were marked as zero in the data[1]. Since the distribution of the burned area is quite problematic, I decided to not use it in the analysis, and focused on the amount of fires in different conditions instead.

Table 1: The means, standard deviations, medians, and 1st and 3rd quartiles of the numerical variables.

	FFMC	DMC	DC	ISI	Temperature	RH	Wind	Rain	Area
Mean	90,6	111	548	9,02	18,9	44,3	4,02	0,0217	12,9
SD	5,5	64	248	4,56	5,81	16,3	1,79	0,296	63,7
Median	91,6	108	664	8,40	19,3	42,0	4,00	0,00	0,52
1st quartile	90,2	68,6	438	6,50	15,5	33,0	2,70	0,00	0,00
3rd quartile	92,9	142	713	10,8	22,8	53,0	4,90	0,00	6,57

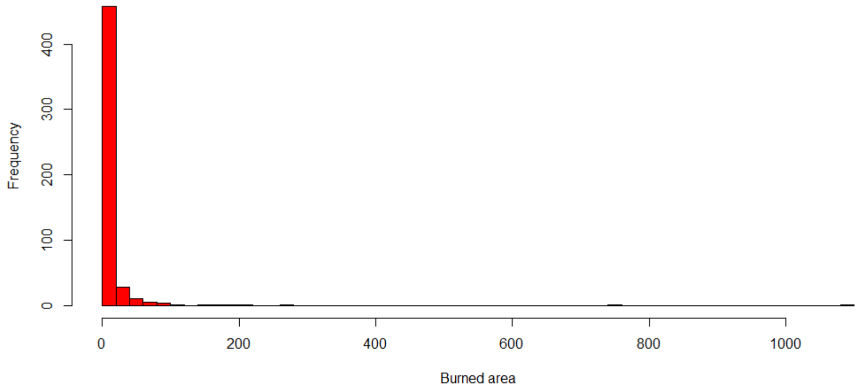

Figure 1: Histogram of the total burned area-variable.

Figure 2 shows the amount of forest fires started on each day of the week. Roughly speaking, the amount of fires on Tuesdays, Wednesdays and Thursdays is lower than on Fridays, Saturdays and Sundays, while Monday is in between.

This could be explained by visitors to the park starting fires on accident on the weekends. Fig. 3 shows the total amount of fires in each month as well as the average temperatures for the days of the months with forest fires. Clearly, August and September form some sort of a forest fire season, with most of the fires in the data set happening there. The other months have quite few fires with the exception of March with a moderate amount of 54 fires. While the average temperature in the forest fire situations was quite high in August and September, it clearly isn't a reliable predictor of amount of fires by itself. While the average temperature value in June is almost the same as in August and September, only 17 fires were recorded in that month.

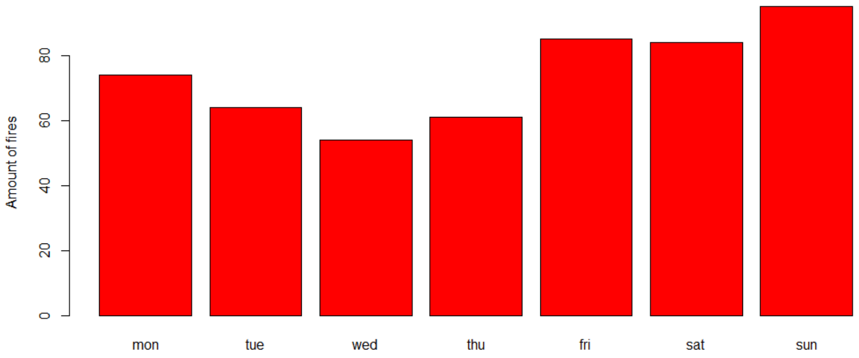

Figure 2: Total amount of fires started on different days of the week.

To get some idea of the final variables in the data set, the X and Y locations in the grid that the national park was divided in to, the amount of fires started are plotted in different grid areas in Fig. 4. The figure also includes the actual map overlaid on the plot to get a sense of the spatial distribution of the fires. It's clear that the distribution is not homogeneous at all even inside the park, with some areas having no fires while some areas being clear forest fire hot spots.

4 Multivariate Analysis

I decided to conduct multiple correspondence analysis on the data, since most of the interesting variables from the point of view of a non-scientist were categorical. To be more precise, the MCA was done on all of the remaining numerical variables and the month-variable. The idea is that this way we can get some better qualitative understanding of the circumstances that cause the different months to have different amounts of forest fires as shown in Fig. 3. The spatial and day variables were discarded, since it would have introduced a quite high number of extra modalities in the analysis. Before starting the multiple

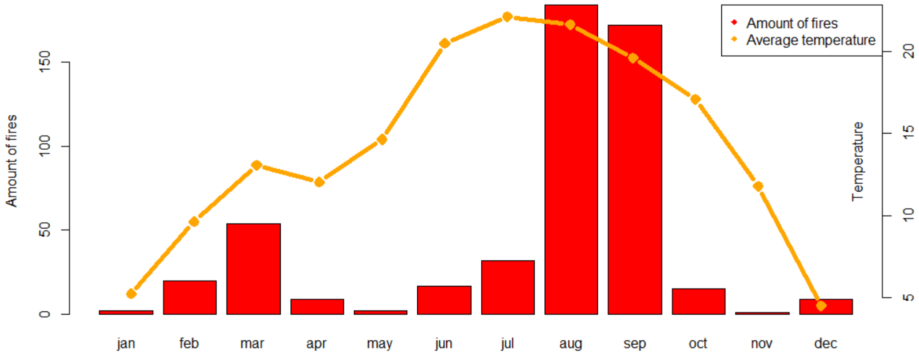

Figure 3: Total amount of fires started in different months and the average temperatures calculated from the days with forest fires.

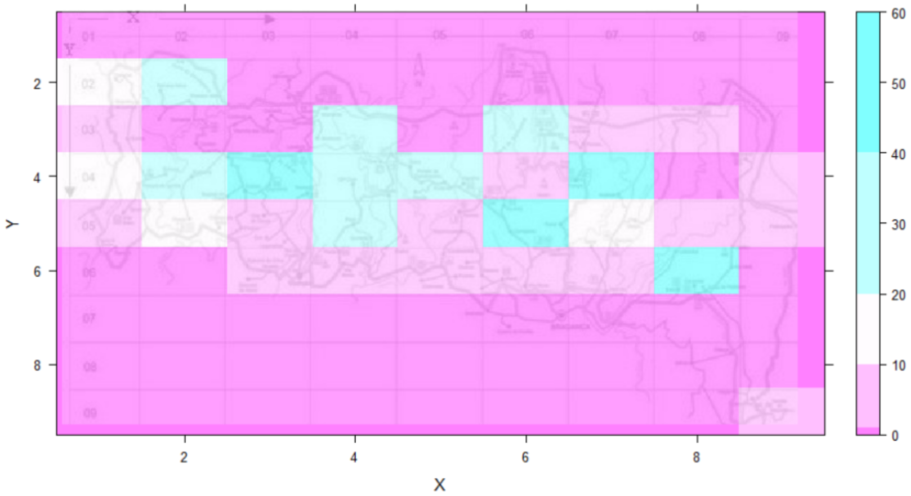

Figure 4: A heatmap of the amount of fires started in different areas of the national park overlaid on the actual map showing the grid. Map taken from [1]. Note that the map distorts the colors slightly.

correspondence analysis, the numerical variables naturally had to be categorized, which I chose to do by dividing the data with the 3-quantiles of each variable. The data under the first quantile were then categorized to "low", the groups between the first and second to "medium" and the remaining to "high" modalities of the corresponding variable. The quantiles are listed in table 2 and visualized in the appendix in Fig. 7. Since the amount of observations for some of the months were very low, I grouped some months to single modalities to avoid problems caused by the unrobustness of MCA. The grouped months are November, December and January, April and May, and also June and July.

Table 2: The two quantiles that divide the variables in to three modalities, as well as the minimum and maximum values of the variables.

Quantile	FFMC	DMC	DC	ISI	Temperature	RH	Wind
Min	18,7	1,10	7,90	0,00	2,20	15,0	0,4
1st	91,0	88,0	587	7,10	17,1	35,0	3,1
2nd	92,4	130	693	9,60	21,4	48,0	4,9
Max	96,2	291	861	56,1	33,3	100	9,4

Figure 5 shows the results of MCA with the first two principal components of the transformed column profiles. Even though the first two components explain only about 26% of the variance, we can try to gain some insight from the plot. We can immediately see that the modalities are divided roughly in to three groups, one surrounding August, one surrounding September and all of the other months. The forest fires in August seem to be associated with high temperature, high FFMC, high ISI, high DMC, low RH, medium DC and medium wind. The fires in September, on the other hand, are more associated with medium temperature, FFMC, ISI, and RH, low wind and high DC. It could be that in August the extreme weather conditions like low humidity and high temperature have enough time to dry the fine fuels and fuels of moderate depth (high FFMC and DMC), which in combination cause the large amount of forest fires. In then September the deeper parts of the soil have had time to dry as well (high DC) but the weather conditions are not as bad, which would be the main difference in the causes of forest fires. Another notable feature is that June and July are the closest of the other months to August, which could indicate that the weather conditions are similar, even though the amount of fires is typically not yet very high at that time of the year. June and July are also negatively associated with high DC and positively with low DC, which again makes sense from a physical point of view: In June and July the weather conditions may just be starting to dry the soil after winter and spring, which then progresses further in August and September.

As mentioned, the two first components don't explain that much of the variance, and including more in the analysis would likely make the results more accurate. The proportions of the explained variances by the different principal components are shown in Fig. 6.

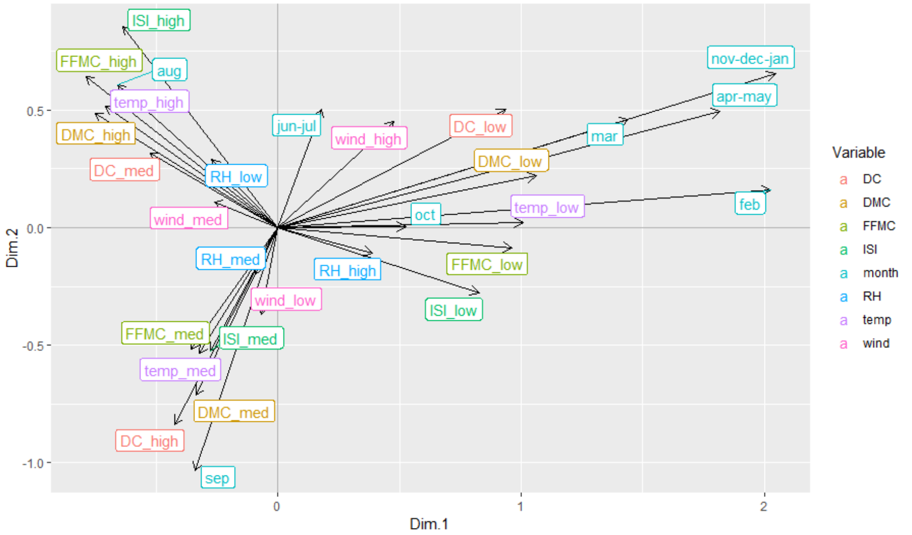

Figure 5: The modalities' projections on the first two modality principal components. The first component explains 16,5% of the variance, while the second explains 9,8%.

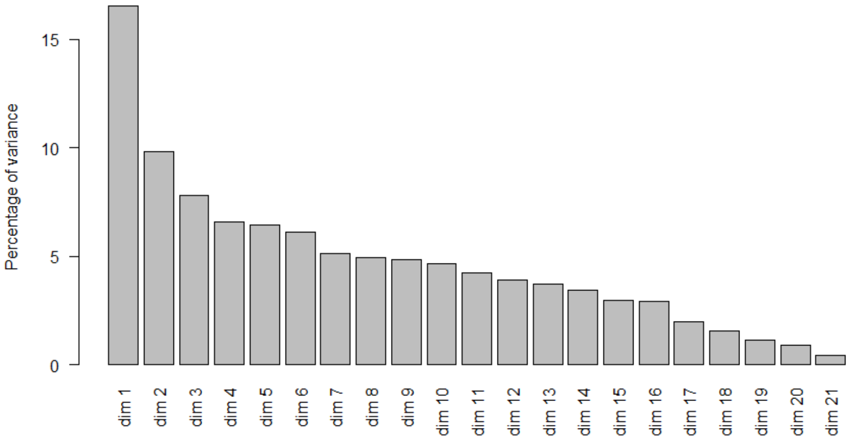

Figure 6: The percentages of data variance that each of the column principal components in MCA explain.

5 Conclusion and critical evaluation

In this work, forest fire data from a national park in Portugal was analyzed with the goal of finding out the conditions where fires tend to start and how.

It was found that fires were more frequent in weekends than weekdays, and certain parts of the park had much more fires than others, although the reason is not clear. Most of the fires happened in August or September, making them a sort of forest fire season. Using MCA, indications were found that the fires in August and September tend to start in somewhat different ways, with the fires in August being due to mostly high temperature, humidity and dryness of the soil up to moderate depth, while the fires in September started more in settings with the deeper levels of the soil being dry as well. The other months, possibly aside from June and July, were associated with low FWI index numbers and low temperatures. The first two column principal components in MCA only managed to explain 26% of the variance of the data, however, so the results possibly left out some important information.

In the MCA analysis it was found that the modalities split roughly in to three groups. This may be in part due to the fact that the numerical variables were divided in to three modalities each, since obviously they are not correlated between each other. The splitting does, however, tell us that for example low DC, low DMC and low temperature tend to go hand in hand, and all of the numerical variables tend to be attracted in the same ways to the different months and to each other. In further analysis, the numerical variables could be divided for example with the 3- of 4-quantiles, although this would also increase the dimensionality of the problem.

It is quite interesting that March has the highest amount of fires after August and September, and this is something that the analysis didn't shed much light on, as March got grouped together with, for example, December and January in MCA results. It may be that the first two dimensions in MCA just weren't adequate to describe March, or that the peak in forest fires is caused by something else entirely. It would be interesting to apply bivariate correspondence analysis or MCA with the variables being some combination of weekdays, spatial sectors and months to see if the fires in some months tend to start more in weekends than in others, for example.

References

[1] P. Cortez and A. Morais., "A data mining approach to predict forest fires using meteorological data.," 2007.

[2] `http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi`. Accessed 25.3.2019.

Appendices

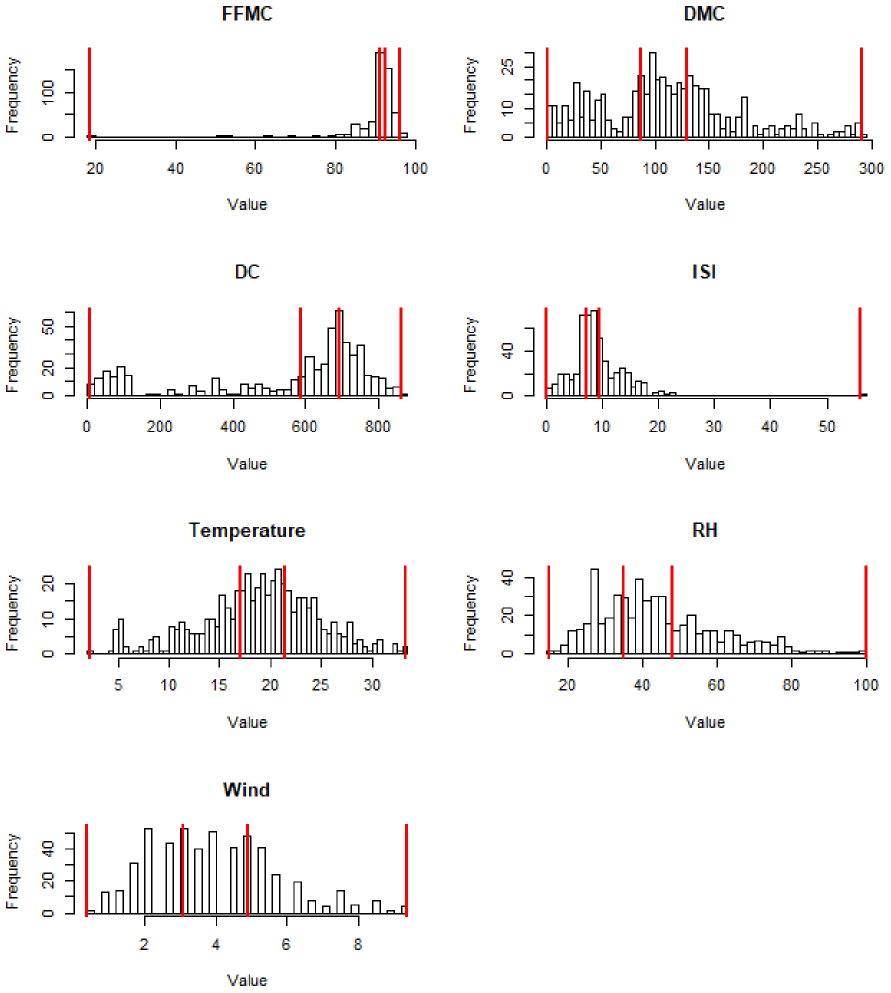

Figure 7: Histograms of the numerical variables used in MCA as well as vertical lines corresponding to the quantiles and minimum and maximum values that divided the variables to modalities.

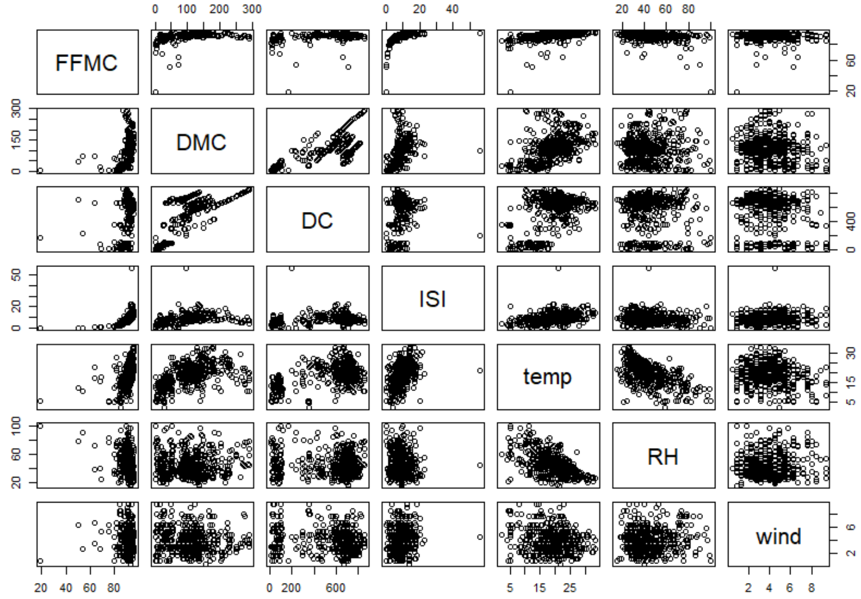

Figure 8: Pairwise scatterplots of the numerical variables used in MCA.