Problemset 1 AMAS

Asger Thormann

February 11, 2026

1 Kenpom 2009

1.1 Extracting data

When loading the dataset I find that the indexing doesn't match for the value DRtg. I have to choose the right index to get the value. I ran into the problem of the blocking by the kenpom website, so i've used the static webpage provided. I use a package called Beautiful Soup to extract the data.

1.2 Histogram

I've chosen to show the intervals on the x-axis and plotting the counts as a bar plot within this interval. I found that many lines where lying on top of each other when plotting a regular histogram so I find this to be more clear.

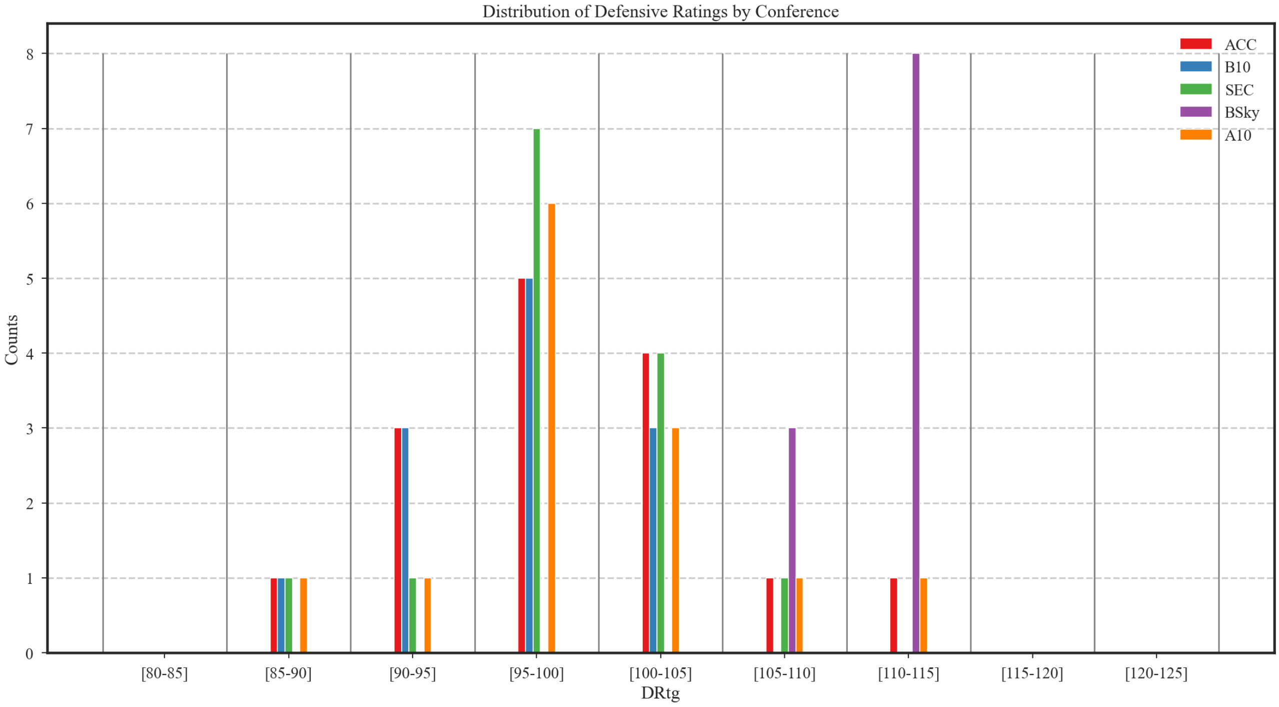

Figure 1: DRtg distribution for each conference

2 Kenpom 2009 and 2014

2.1 Difference in ORtg

I extract the dataset as before and match the team names within each conference. I find that the some teams change conference between 2009 and 2014 as seen in table 1. To create the plot I choose to only show the teams that are in the same conference. The label 'Other' defines teams that are not in any of the 5 conferences. I also find that some teams are only in the 2009 or the 2014 dataset respectively, these are also neglected. The scatterplot is shown in fig. 2

	ACC	SEC	B10	Bsky	A10	Other
Number of teams (2009)	12	12	11	9	14	270
Number of teams (2014)	15	14	12	11	13	276

Table 1: Number of teams in the 5 conferences in the two datasets

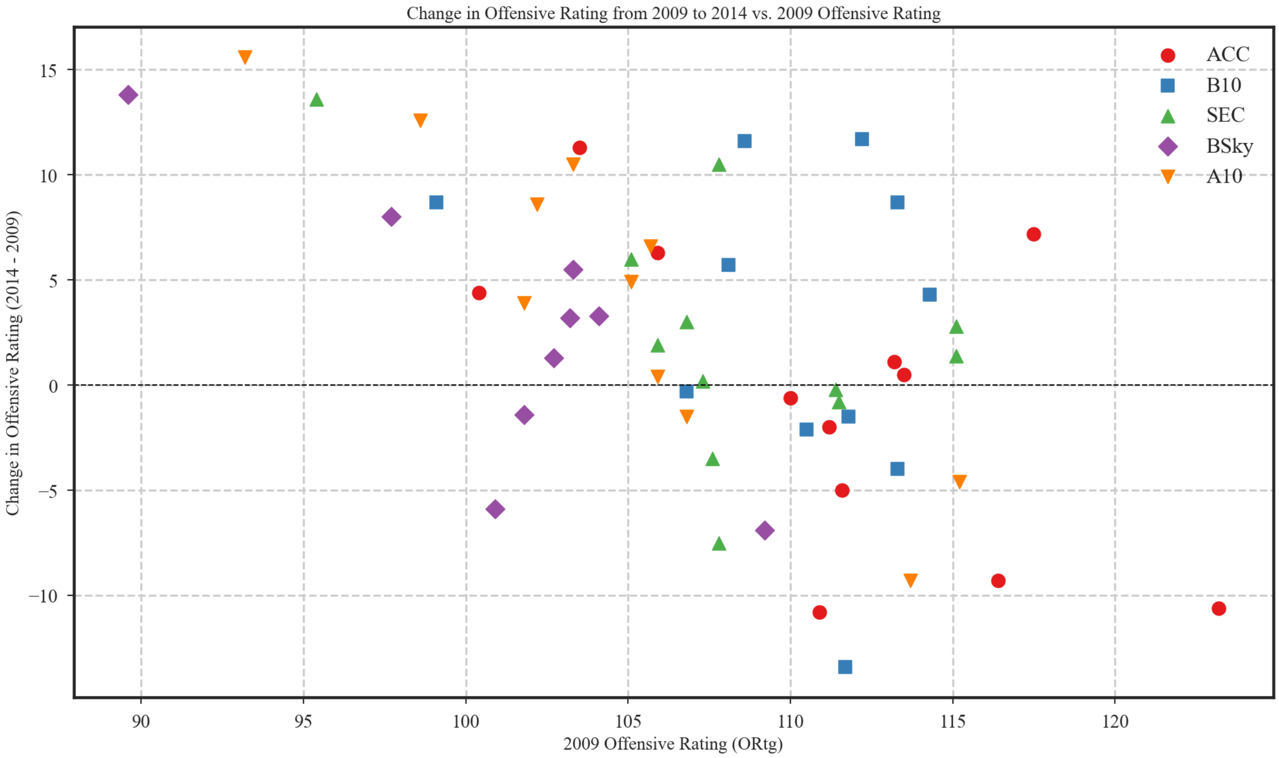

Figure 2: Difference in ORtg for each team in the same conference.

2.2 Mean and median

When finding the mean and median of the rest I find that some teams are only in 2009 and some are only in 2014 I have chosen not to include these in the calculation of mean and median. Further more as seen above some teams change conference, I've chosen to still compare these ORtg values. In summary: For the teams in the 5 conferences I find the mean and median if they are in the same conference in both data sets. I find the mean and median of the rest regardless even if they are not in the same conference. Thus teams that are in any of the 5 conferences in ONE of the datasets are neglected. The mean and median values are shown in table 2.

Conference	Mean	Median
ACC	-0.63	-0.05
B10	2.67	4.30
SEC	2.28	1.65
BSky	2.32	3.20
A10	4.34	4.90
Other	2.60	1.90

Table 2: Change of offensive rating for each conference

3 Including BE

In this section I include the conference BE.

3.1 Histogram

See figure 3.

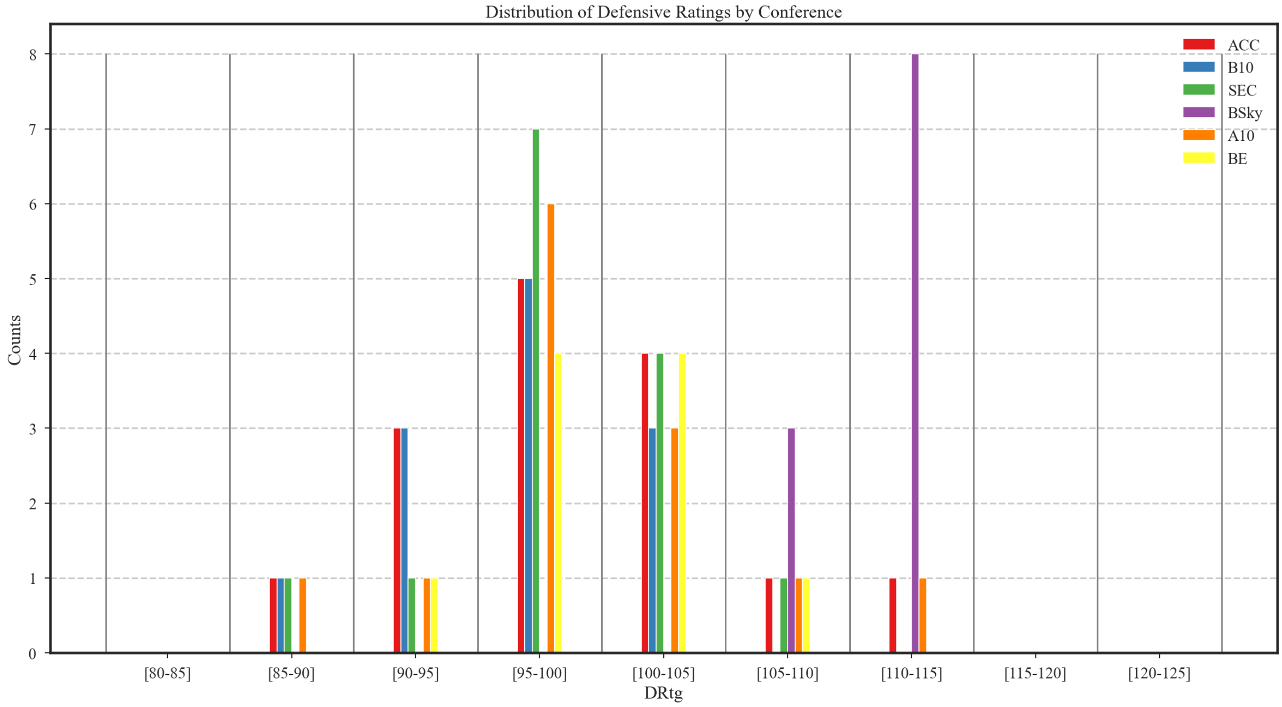

Figure 3: DRtg with the conference BE

3.2 Difference in ORtg

See figure 4.

3.3 Mean and median

I recalculate the mean and median for other teams as BE is now not included. 3.

Conference	Mean	Median
ACC	-0.63	-0.05
B10	2.67	4.30
SEC	2.28	1.65
BSky	2.32	3.20
A10	4.34	4.90
BE	1.14	1.90
Other	2.62	1.85

Table 3: Change of offensive rating for each conference

4 Extra credit

I read the PDF using a package and account for accents using and accent map to normalize these characters.

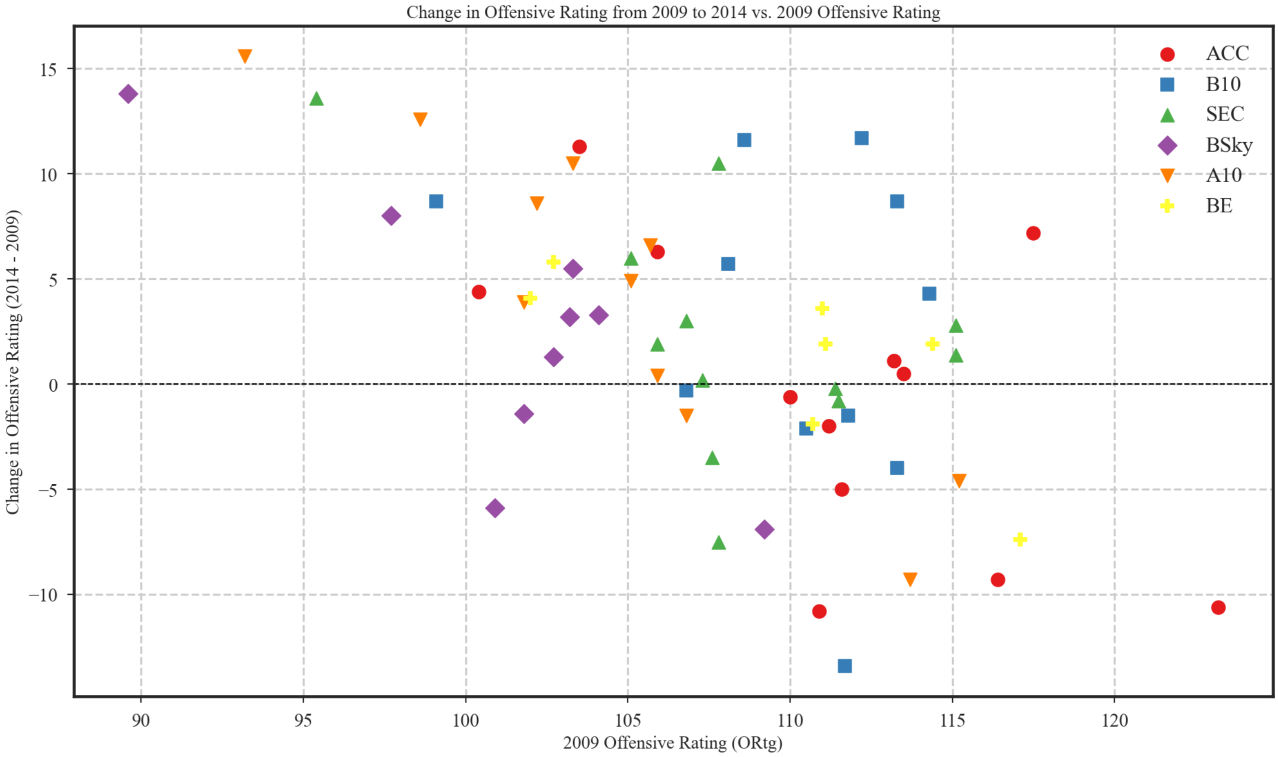

Figure 4: Difference in ORtg

I remove the papers cited on the first page, and the description of collaborations on the last page.

I remove the header on each page an the page number.

I use the parenthesis to search for collaboration titles and remove text within these.

I remove all instances of newline commands.

I split the names by "," and remove the numbers that the determine the association and white spaces. This gives me 3614 names in the list. I rearrange the names splitting by the last instance of "." and adding a white space. All names should now follow the format "Jacobs B." I check the names for spaces and "." characters as I assume all names have at least one initial and a last name. I get multiple names that has the form "BobJacobs". I take these names and split by uppercase letters leaving the format "Jacobs Bob" which I can sort alphabetically. I also find two cases that I have to rewrite manually: ('M.ConstancioJr.', 'R.E.RyanJr.'). I couldn't find a rule to filter these automatically.

In this I have assumed that names following the format "A. J. van der Horst" and so forth has a last name that starts with "v" and not "H". As stated in the problem last names with an accent is at the end of the list. Furthermore I've neglected that Chinese names have a different structure where the family name is the stated first and the the given name. I assume that some names in this list are stated this way but I have not filtered these.

The authors at the middle location (index 1806 and 1807) is Lippert M. and Lipunov V.M.