



The main problem here is that the variables are not equally distributed among the batches and as a result, the data that we get can be 'contaminated' with patterns unrelated to the biological signal of our interest. The histogrammes show that different types of bladder cancers are separated quite well when it comes to batches (mTCC is mainly in the 1t batch, mTCC-CIS in the 2nd and mTCC+CIS in the 3rd) and so we can imagine a situation in which the five batches correspond to the days in which data was gathered or different rearchers resposible for gathering it. Such variables greatly influence the results of our analysis and, if left unaccouted for, can lead to false biological conclusions.