chem-bla-ics

Looking at your statistical models...

Egon Willighagen

Published June 20, 2010

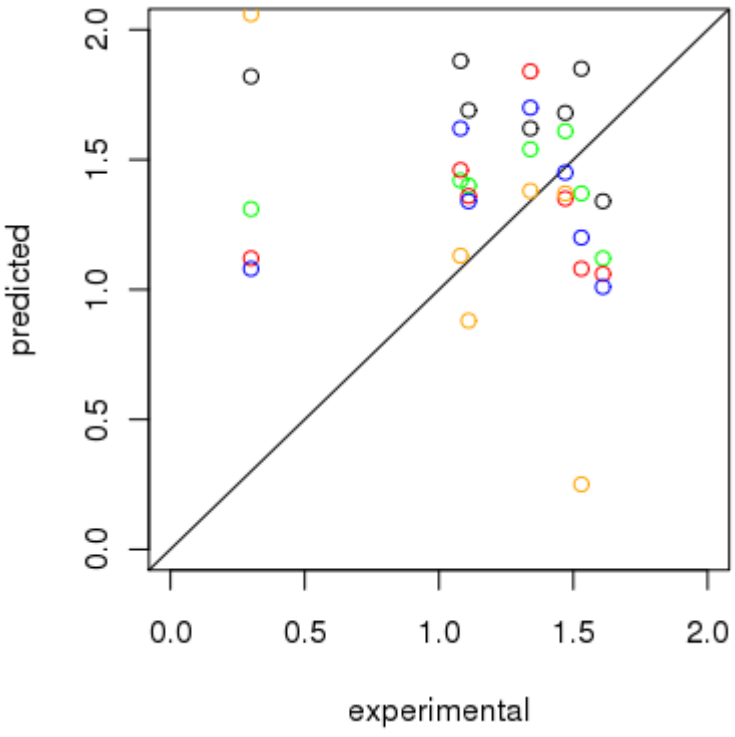
Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/6948f-pbe89

Keywords
Chemometrics, Qsar

I do not think I have ever blogged the paper that played an important role in my thesis (doi:10.1021/ci990038z); research of one of the papers in my thesis, started with the hypothesis proposed therein. The paper had a really good idea; but, unfortunately, it did not contain the data to support the hypothesis. That gets me to one important lesson I learned: **a QSAR data set of less than 100 molecules is not enough to make untargeted statistical models.**

The paper reads quite nicely, and the results are clear: by combining spectral types, the RMSEP goes down. Good! Lower prediction errors; that's what we all want. So, a M.Sc. student of mine set off, but after about half a year, he was still unable to make statistically good models. He used bootstrapping to 'prove' it was not his fault: there was not enough data for the method to learn the underlying patters. Hence my above lesson. My student went on with larger data sets, and laid out the foundation of what later became the paper on using NMR spectra in QSPR modeling (doi:10.1021/ci050282s). Now you understand why QSAR is missing.

So, if those results are so clear, then why does it not work? As said, the data set was too small for pattern recognition methods to see what was going on. The RMSEP numbers just came out nicely; however, if we had only made the below plot, if would have warned. But I failed to do that at the time. Lesson learned: **do not just look at the data, but also look at the model.** And look really means looking with your eyes at graphical representations of that model. The plot:



The numbers in this plot are hidden in tables in the paper. The RMSEP values earlier mentioned are calculated from those. From the plot, you can see that the test data consisted of 5 compounds; the training set contained 37 compounds; all are congenerics, and do not span a high diversity. Now, the plot shows five models: black is COMFA; orange is based on experimental IR spectra; red, green, and blue are models where two types of representations

are combined. From the RMSEP values it can be seen that combining representation improves the RMSEP values. That's what you want, and sort of makes sense.

Now, I did not make this plot until I started writing up the paper, and tried to figure out why the QSAR data set did not work. My eyes opened wide when I saw the orange dots! Anti-correlation! WTF?!?! I mean, we are looking at a plot visualizing the predicted versus the experimental activity... Actually, the others are not really convincing either, are they? Looking at the predictions for compounds with experimental values around 1.0-1.5 (if you really want to know the unit, read the article), the pattern is pretty much anti-correlated too. Thinking about it, it seems the RMSEP is mainly reflecting the error of the left most compound, the one with a experimental activity of about 0.3.

Clearly, the orange model is hopeless, but the others are not really better. Now, the paper actually makes statements comparing the various combinations of representations, but, in retrospect and looking at this plot, I wonder if the green model is really different from the blue or red models.

Since then, I always make these kind of plots, just to see what my model is like. Since then, I distrust papers that only show RMSEP, Q^2, or other quality statistics. Now, the tricky part is, you need those statistics if you want to automate model selection; the variance on those model quality statistics is actually so high (see also my other post today), that you must carefully validate that model selection too, visually of course.

I have been long thinking what to do with these observations. I did not dare publish them in my thesis; I did not dare write a letter to the editor. Perhaps I should. But even writing up this blog makes me feel uncomfortable. Besides the fact that I might be wrong, I also do not like to point out mistakes (IMHO); particularly, when those are published in a respectable journal. I was fooled by the statistics too (and was already well trained), so I cannot comment on the authors overlooking the issue. Or the reviewers! Or the community at large. Also, I do not know what the fate of this paper should be. The idea is quite interesting, even though the published results do not support it. Not shown here, but the bootstrapping results show that the apparent slight improvement is merely a numerical artifact, just happening by chance, based on luckily selecting the test compounds; the data is just insufficient in size to draw any conclusion.