chem-bla-ics

When to stop including QSAR model variables...

Egon Willighagen

Published November 8, 2005

Citation

Willighagen, E. (2005). When to stop including QSAR model variables. *Chem-bla-ics*. https://doi.org/10.59350/hxb0r-66s49

Keywords

Cheminf

Yesterday I reviewed an article which published a QSPR model which looked something like:

\y = 151 + 50p1 - 12p2 - 0.006p3\

with quite OK prediction results (R=0.9880). But I was not quite comfortable with the coefficient for the \p3\ variable. The article did not calculate significances for the coefficients, so it was not obvious from the article wether is was useful to include them. I then looked at the range for p3, which was 110-150; so, the maximal influence this variable can have is \150*0.006 = 0.9\. Now, the experimental values given in the article were rounded to integers, indicating that the maximal effect of the p3 variable is smaller than the experimental error! It's even worse when you consider the difference between the min and max value (40), then the influence would even be smaller (assuming that most model methods would put the mean temperature effect in the offset, 151 in this case).

Today, I reread an article with a similar issue. The model was something like:

\y = -0.81 + 0.03*p1 + 0.009*p2\

Here, \max(p2)-min(p2)\ is a smaller than 100, so the maximal effect of the variable would be in the order 0.9, which is of the same order of the root mean square error of prediction (RMSEP) for this model. Indeed, the article already states that the coefficient is only significant at the 95% level, and not at the 99% level. But, without having calculated the RMSEP for a model without the p4 variable, I would guess that leaving it out would give equally good prediction results.

Concluding, I would say the the p2 variable does not include relevant information.

Do you think it is reasonable to include the p2 variable in the second model?

References