

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

Copyright

Copyright © Egon Willighagen 2005. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

chem-bla-ics

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

$$y = 151 + 50p_1 - 12p_2 - 0.006p_3$$

with quite OK prediction results ($R=0.9880$). But I was not quite comfortable with the coefficient for the p_3 variable. The article did not calculate significances for the coefficients, so it was not obvious from the article whether it was useful to include them. I then looked at the range for p_3 , which was 110-150; so, the maximal influence this variable can have is $150 \times 0.006 = 0.9$. Now, the experimental values given in the article were rounded to integers, indicating that the maximal effect of the p_3 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:

$$y = -0.81 + 0.03 \cdot p_1 + 0.009 \cdot p_2$$

Here, $\max(p_2) - \min(p_2)$ is 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 p_2 variable, I would guess that leaving it out would give equally good prediction results.

Concluding, I would say the p_2 variable does not include relevant information.

Do you think it is reasonable to include the p_2 variable in the second model?

References