**Regression model**

To better assess how well variables of our interests can be estimated given different combinations of measurements (and thus to prioritize efforts to measure different model inputs), two separate linear regression models are construct in the following pattern:

 (1)

where Y represents two log-transformed response variables, which are 1. individual animal's tissue concentration divided by average soil concentration (logBCF), and 2. application factor (logApp). α is the estimated coefficients, β is the constant, and ε is the error term. The independent variables Xs come from two categories a. pesticide properties: 1. log-transformed pesticide soil partition coefficient (logkoc), 2. log-transformed pesticide molecular mass (logMol), 3. log-transformed pesticide solubility in water at 20°C (logSol), and 4. log-transformed pesticide density (logDen). The second source is based on frog properties, 5. habitat factor (HabFac).

**Selection criteria**

Figure 1 summarizes the selection criteria into a flow chart with two parts. The first part aims to avoid model over-fitting, since the performance of these regression models (such as R2) naturally improves as additional model inputs are included. Thus Bayesian information criterion (BIC) is employed to provide insights into the extent of benefit of the additional information, since BIC resolves this problem by placing a penalty term for the number of parameters in the model. When data are sufficient, the best regression model is usually associated with the lowest BIC value. However, data from this study are only collected from five pesticides, which is a small portion of all the market available pesticide products (approximately over 10,000?). Under this circumstance, the BIC recommended regression model would not always be the best candidate when it is applied to new pesticides. As a result, a second step has to be introduced aiming to identify regression model structures, which reasonably explain current dataset and maintain the generality when they are applied to predict based on new dataset.

This process is done by applying cross-validation method. Six sub-steps are list as follows:

2.1. Divide the full dataset into a testing chemical dataset only contains on one pesticide (Yj), and a training dataset (Y-j) with the other four pesticides. By doing this for all five pesticides, we will have five pairs of testing and training dataset.

2.2. For each set of Yj and Y-j, fit a linear regression model against the training dataset and calculate its residuals (REStraining, -j). Applying the same linear model to the testing dataset and estimate its residuals (REStesting, j). Repeat this process for all the possible combinations of explanatory variables.

2.3. Repeat Step 2.2 for the other four pair of datasets.

2.4. Pool all the training residuals and testing residuals and estimate their root mean square errors (RMSEtraining, RMSEtesting) based on Equation 2[1]:

 (2)

2.5. Estimate the weighted RMSE based on the percentage of training () and testing () dataset. In this study, the testing data are about 20% (=20%) of the whole dataset:

 (3)

2.6. Sort the RMSEoverall in ascending order, and select models whose RMSEoverall is within 10% of its previous one. Then recommend regression models whose BIC values are with 10 units of the premium one[2]. Tables 1a and b summarizes the recommended regression models (bolded).

1. Deakin, R.E. and D.G. Kildea, *A Note on Standard Deviation and RMS.* Australian Surveyor, 1999. **44**(1): p. 74-79.

2. Burnham, K.P. and D.R. Anderson, Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, 2004. 33(2): p. 261-304.

1: Build regression models for all combinations of independent variables, and record BICs.

Step. 1

2.3. Repeat Step.2.2 for the other four pair of datasets

2.2. For one training dataset, fit all possible combinations of linear regression models, and applying the fitted models to the corresponding testing dataset. Then record residuals.

Step. 2

2.1 Divide the whole data into five pairs of testing dataset (Yj) and the corresponding training dataset (Y-j).

2.5. Calculate the weighted RMSEoverall

2.4. Estimate RMSE for all pooled residuals.

2.6. Sort the RMSEoverall, and recommend models whose BIC scores are within 10 units of the best.

Figure 1 Flow Chart

Table 1a. BCF

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **HabFac** | **logkoc** | **logMol** | **logSol** | **logDen** | **R2** | **BIC** | **Weight\_RMSE** |
| **1** | **HabFac** | **logkoc** |  | **logSol** |  | **0.69** | **318.12** | **0.80** |
| **2** |  | **logkoc** |  | **logSol** |  | **0.65** | **323.46** | **0.83** |
| **3** | **HabFac** | **logkoc** |  | **logSol** | **logDen** | **0.69** | **322.50** | **0.87** |
| **4** |  | **logkoc** |  | **logSol** | **logDen** | **0.65** | **328.07** | **0.90** |
| 5 | HabFac | logkoc |  |  |  | 0.56 | 358.22 | 0.95 |
| 6 |  | logkoc |  |  |  | 0.52 | 357.63 | 0.97 |
| 7 | HabFac | logkoc |  |  | logDen | 0.56 | 363.03 | 1.01 |
| 8 |  | logkoc |  |  | logDen | 0.52 | 362.49 | 1.03 |

Table 1b. App

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ID** | **HabFac** | **logkoc** | **logMol** | **logSol** | **logDen** | **R2** | **BIC** | **Weight\_RMSE** |
| **1** | **HabFac** | **logkoc** |  | **logSol** |  | **0.69** | **362.33** | **1.10** |
| **2** |  | **logkoc** |  | **logSol** |  | **0.64** | **368.97** | **1.16** |
| **3** | **HabFac** | **logkoc** |  | **logSol** | **logDen** | **0.69** | **366.90** | **1.26** |
| 4 | HabFac | logkoc |  |  |  | 0.47 | 425.85 | 1.28 |
| **5** |  | **logkoc** |  | **logSol** | **logDen** | **0.65** | **372.57** | **1.32** |