# Developing a MACRO meta-model for Swedish drinking water abstraction zones

**Stefan Reichenberger [1], Mikaela Gönczi [2], Nils Kehrein [1], Sebastian Multsch [1], Nicholas Jarvis [2], Jenny Kreuger [2]**

[1] knoell Germany GmbH, 68163 Mannheim, Germany
(contact: sreichenberger@knoell.com)
[2] Swedish Agricultural University (SLU), Uppsala, Sweden
(contact: nicholas.jarvis@slu.se)

## Introduction

- In Sweden farmers are legally obliged to apply to local authorities for permits for pesticide use if their land lies within a designated water abstraction zone.
- A standalone modelling tool developed by SLU (MACRO-DB) is available to facilitate risk assessment and decision-making in water abstraction zones.
- The tool, which is used both by local authorities (who make the decisions) and farmers/landowners and consultants, is based on the well-established leaching model MACRO 5.2 (Larsbo and Jarvis, 2003; Larsbo et al., 2005).
- Our aim is to develop a robust meta-model of MACRO-DB, as a fast and easy-to-maintain web-based tool for these risk assessments.

## Objectives

The objectives of this study were to

1) create a large synthetic dataset of pesticide leaching with MACRO for a pilot region
2) implement, calibrate and validate a meta-model using the CART methodology (Breiman et al., 1984)

## MACRO simulations

- 18720 leaching simulations were performed with MACRO 5.2 for a pilot region in Southern Sweden (SW Skåne; cf. Fig. 1).
  - 39 soil scenarios (defined by geological substrate, hydrologic class, soil texture and organic matter content),
  - 1 climate (zone 1001 in Fig. 1),
  - 1 crop (spring cereals)
  - 3 application seasons
  - 160 dummy compounds (combinations of Kfoc, DegT50 and Freundlich exponent).
- Simulation period: 26 years (6 years warm-up + 20 years evaluation period).
- Target variable: mean leaching flux concentration over 20 years at 2 m depth (PECgw).
- Finally, simulations were grouped into classes according to predicted leaching concentrations (Table 1).

**Table 1: Grouping the 18720 MACRO simulation runs into leaching classes**

| broad classes | fine classes | mean leaching flux conc. (µg/L) min | max | nb runs (%) |
|---|---|---|---|---|
| A | cat_1 | 0 | < 0.001 | 48.2 |
| | cat_2 | 0.001 | < 0.01 | 4.58 |
| | cat_3 | 0.01 | < 0.1 | 7.83 |
| B | cat_4 | 0.1 | < 1.0 | 10.4 |
| | cat_5 | 1 | | 28.9 |

## Classification and Regression Trees (CART)

- CART (Breiman et al., 1984) is a group of decision tree learning methods.
  - Classification trees (CT): predicted outcome is a categorical variable
  - Regression trees (RT): predicted outcome is a numerical variable
  - CART decision trees are constructed top-down, by choosing a variable at each step that best splits the data. Finally, trees are "pruned" in an internal cross-validation step.
  - CT is not applicable to our problem (meta-model predictions need to be scalable with the application rate)
- Regression trees (RT)
  - Tree building is strictly based on variance: Groups (nodes) are split such that the variance between the daughter nodes is maximized. → very transparent method
  - Complexity parameter (cp): if any split does not increase the overall $R^2$ of the model by at least cp, then that split is considered as not worth pursuing and not made. The default value of cp = 0.01 has been reasonably successful at "pre-pruning" trees, but it sometimes over-prunes, particularly for large data sets (Therneau et al., 2019).
  - Predictions: The predicted value of the target variable is equal to the mean of the group in which a data point ends up after going through the decision tree

## Meta-model development

- A tool (rCART) for meta-model development with CART was implemented in R, making use of the R package rpart (Therneau et al., 2019).
- rCART splits the data randomly into a calibration and a validation dataset (parameter splitRatio specifies the fraction of data points to be used for calibration)
- rCART was run for different values of cp and splitRatio (0.25-0.999), different ways of data censoring/truncation and logarithmic vs. non-logarithmic leaching concentrations.
- Output for each CART run
  - figures: decision tree (cf. Fig. 2); scatterplot for prediction (cf. Fig. 3)
  - complete rpart results for the tree building
  - predictive performance measures: RMSE; $R^2$, $r^2$, PBIAS, fraction of correctly predicted leaching class (exact match of concentrations not necessary for decision-making tool)
  - list of outliers
  - relative importance of explanatory variables (cf. Table 2)

## References

- Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software. Monterey, CA. ISBN 978-0-412-04841-8.
- Larsbo M, Jarvis N (2003). MACRO 5.0 - A model of water flow and solute transport in macroporous soil. Technical description. SLU, Dept. Soil Sci., Uppsala, 47 pp.
- Larsbo M et al. (2005). An Improved Dual-Permeability Model of Water Flow and Solute Transport in the Vadose Zone. Vadose Zone Journal 4, 398-406. DOI: 10.2136/vzj2004.0137
- Steffens K, Jarvis NJ, Lewan E, Lindström B, Kreuger J, Kjellström E, Moeys J. 2015. Direct and indirect effects of climate change on herbicide leaching – a regional scale assessment in Sweden. STOTEN 514, 239-249.
- Therneau TM, Atkinson EJ, Ripley B (2019). rpart: Recursive Partitioning and Regression Trees. https://cran.r-project.org/package=rpart

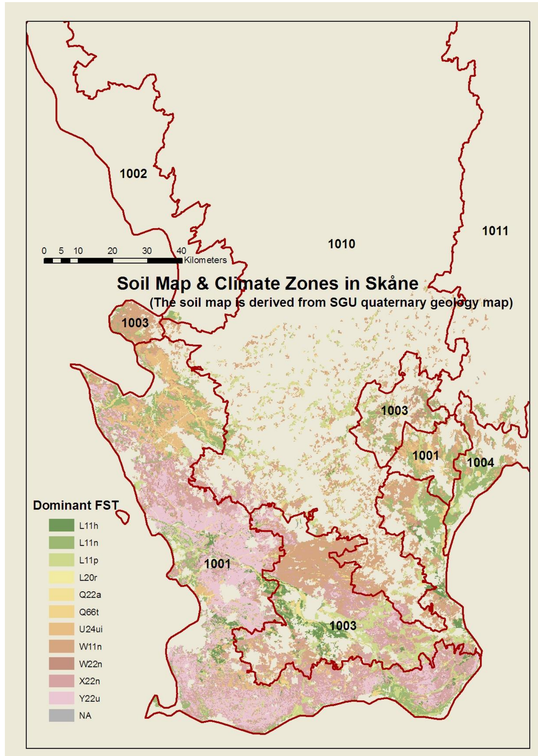



Fig. 1: Soil map and climate zones of Skåne (Steffens et al., 2015)

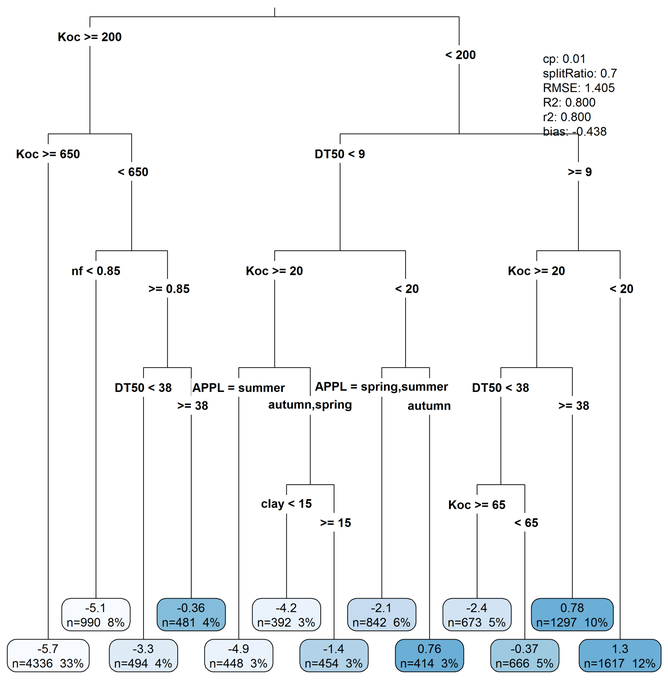



Fig. 2: Example regression tree. cp = 0.01; splitRatio = 0.7; logarithmic conc. censored at - 6

## Results and Discussion

- MACRO simulations yielded uneven distribution of leaching concentrations (cf. Table 1):
  - 77 % of the data points in very low or very high range;
  - few data points in the middle range (classes cat_2, cat_3 and cat_4).
- CART results
  - performance better for logarithmic concentrations; however, data need to be censored (e.g. at 1.0e-06 or 1.0e-04 µg/L)
  - For Koc = 10000 L/kg all PECgw < 1.0e-15 µg/L → values will be censored anyway. However, excluding the data points with Koc = 10000 L/kg did not increase overall performance.
  - Fraction of data points used for calibration: best predictions obtained for splitRatio = 0.67-0.75.
  - Complexity parameter cp: Decreasing cp improved the prediction, albeit asymptotically. With cp = 1.0e-05 the trees are visually very complex, but still not too deep (max. 13 split levels)
  - including simulated percolation volume at 2 m depth (WWW) as additional explanatory variable did not have significant effect
  - comparison of 8 variants with logarithmic concentrations yielded only minor differences (Table 3; Fig. 3)
    - importance of explanatory variables: Koc >> DT50 >> nf > others
    - prediction of leaching class very good for cat_1 and cat_5, but poor for cat_2 → related to number of data points in each class in the calibration dataset
- Discussion
  - Results too pessimistic because we could not use all 18720 simulations for tree building. → need independent test data set
  - Better choose random substance properties for meta-model development as opposed to regular grid (better exploration of parameter space)?

Table 3: Performance of different variants (logarithmic conc., cp = 1.0e-05, splitRatio = 0.7)

| variant | | | calibration | | | prediction | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | variable importance (%) | | | fraction of correctly predicted leaching class (%) | | | | | | |
| censoring value (lg µg/L) | use percolation | include Koc = 10000 | Koc | DT50 | nf | $R^2$ | cat_1 | cat_2 | cat_3 | cat_4 | cat_5 | all 5 |
| -6 | no | y | 46 | 27 | 7 | 0.973 | 95.9 | 42.4 | 70.2 | 72.6 | 95.5 | 88.7 |
| -6 | yes | y | 45 | 26 | 7 | 0.972 | 96.3 | 45.1 | 70.1 | 73.9 | 95.1 | 89.0 |
| -6 | no | n | 43 | 24 | 8 | 0.970 | 95.3 | 49.1 | 63.7 | 73.2 | 96.7 | 88.2 |
| -6 | yes | n | 42 | 23 | 8 | 0.970 | 95.2 | 49.8 | 64.0 | 74.7 | 96.4 | 88.4 |
| -4 | no | y | 44 | 30 | 5 | 0.976 | 95.6 | 48.5 | 68.7 | 73.9 | 96.2 | 89.3 |
| -4 | yes | y | 43 | 30 | 5 | 0.975 | 95.8 | 49.6 | 68.6 | 76.1 | 96.0 | 89.6 |
| -4 | no | n | 42 | 29 | 6 | 0.974 | 96.3 | 44.3 | 68.4 | 73.5 | 96.3 | 88.2 |
| -4 | yes | n | 41 | 28 | 6 | 0.973 | 96.0 | 45.4 | 69.3 | 74.2 | 96.5 | 88.5 |

## Preliminary conclusions

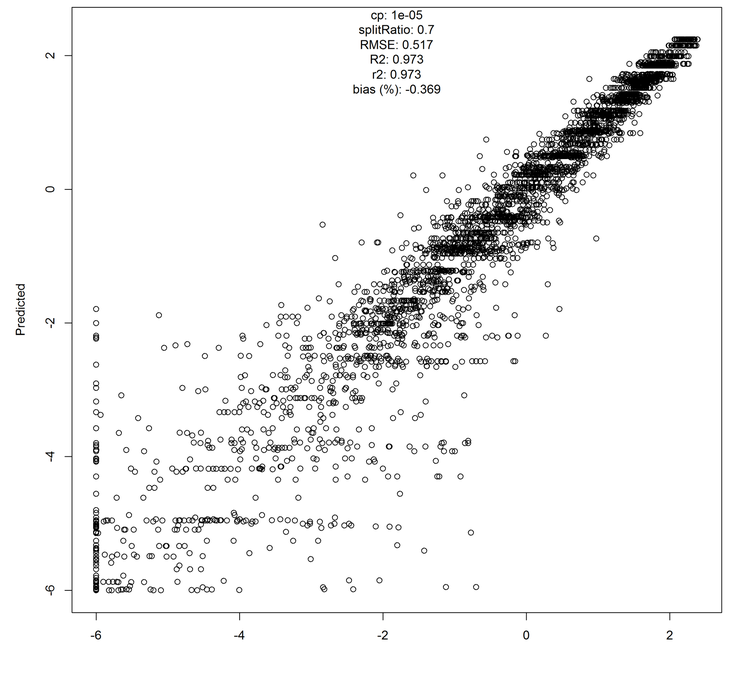
- The Regression Tree (RT) methodology, which is strictly variance-based, was not able to predict leaching concentrations well for the middle concentration range (0.001 – 0.1 µg/L).
- This is most probably due to the distribution of the target variable in the MACRO dataset, with predominantly very high or very low values.
- Possibly RT is not the most suitable approach either for the regular grid of substance properties we used → to be tested.

## Explanatory variables

**Table 2: Explanatory variables used in the CART procedure**

| Variable | description | type | values |
|---|---|---|---|
| QG | quaternary geology | categorical | eskers, moraines, sedimentary rocks |
| silt | silt content (%) | numeric | 5-70 % |
| sand | sand content (%) | numeric | 20-90 % |
| clay | clay content (%) | numeric | 5-30 % |
| TEXT | texture class | categorical | 1 (> 70 % sand), 2 (20-70 % sand) |
| HC | hydrological class | categorical | L, W, Y |
| SOM | organic matter class | categorical | h (high), n (normal), u (undeveloped) |
| APPL | application season | categorical | spring, autumn, summer |
| Koc | normalised Freundlich adsorption coefficient (L/kg) | numeric | 3-10000 L/kg |
| DT50 | degradation half-life in soil (d) at 20 °C and pF = 2 | numeric | 3-200 d |
| nf | Freundlich exponent | numeric | 0.7-1 |

## Predictive capability



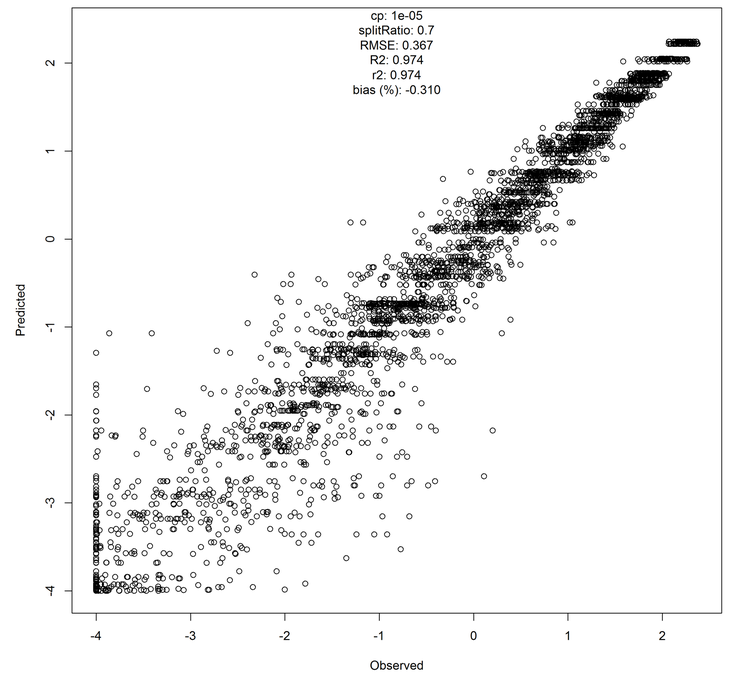




Fig. 3: Observed (i.e. simulated with MACRO) vs. predicted mean leaching concentrations. cp = 1.0e-05, splitRatio = 0.7, logarithmic concentrations. Top: censoring at -6. Bottom: censoring at -4 and exclusion of substances with Koc = 10000 L/kg.

## Next steps

- Create an independent test dataset by running new MACRO simulations with pseudo-random substance properties (e.g. drawn with Latin Hypercube Sampling).
- Run CART on whole calibration data set (18720 runs) and apply predictively to new test dataset
- Prepare lookup table from the 18720 runs and try different interpolation approaches (linear, log-linear, other) in a 3-dimensional space (Koc, DT50, nf). Apply best interpolation approach predictively to new simulations.
- Compare the predictive performances of CART and the interpolation approach.
- Potentially create new calibration dataset with different distributions of target variable and substance parameters.

## Outlook

- Once a working meta-model has been established for the test region of SW Skåne, the analysis will be extended to other climatic regions in Sweden.
- The meta-model will be integrated in a web-based tool for GW risk assessment in water abstraction zones.

## Acknowledgements

Financial support from the Swedish Agency for Marine and Water Management is gratefully acknowledged (Contract 1022-18).