Poster Reprint

ASMS 2019
TP298

Comprehensive machine learning prediction of GC/MS pesticide recovery based on the molecular fingerprinting for food QA/QC

Takeshi Serino* [1,2]; Sadao Nakamura[1];
Yoshizumi Takigawa[1]; Norton Kitagawa[3];
Shigehiko Kanaya [2]

[1] Agilent Technologies, Hachioji City, Japan

[2] Nara Institute of Science and Technology,
Ikoma City, Japan

[3] Agilent Technologies, Santa Clara, CA

Introduction

Challenges in pesticide analysis by GC/MS

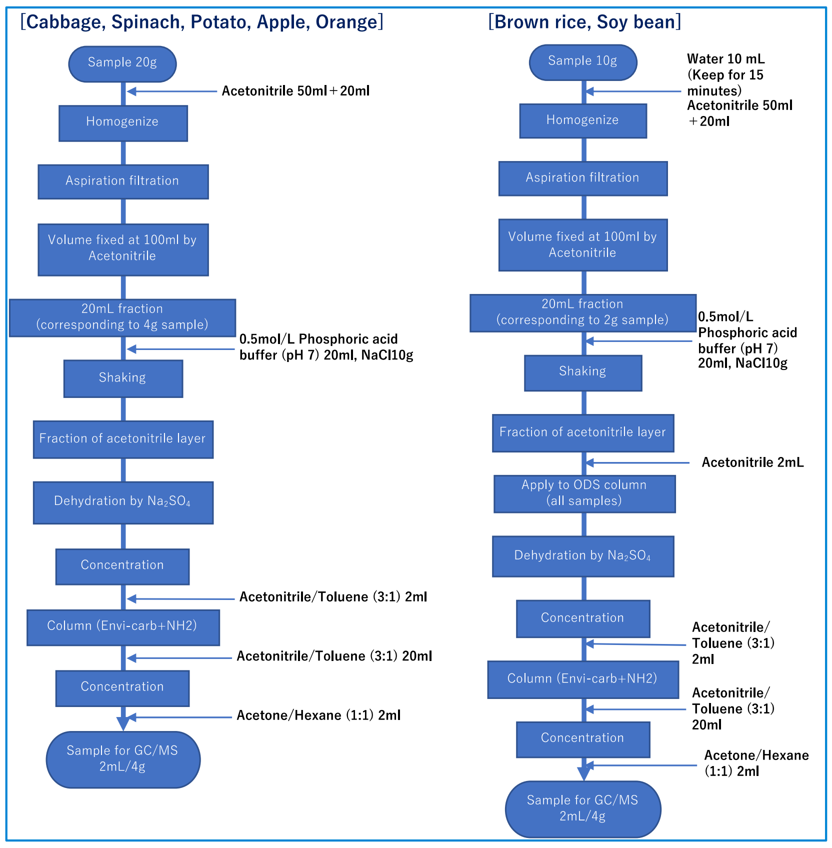
Pesticides are widely used for production of the crops. The minimum residual level of pesticides in foods is regulated by the governments worldwide to protect the consumers' health. GC/MS is widely used for detection of various residual pesticides in foods for safety; however, the recovery ratio can vary by various factors such as the sample matrix, the pesticides' chemical properties of the pesticides and the sample preparation treatment, etc.. The prediction of recovery rate is important for ensuring the food safety. In the present study, we demonstrate how to select the most suitable machine learning regression models to predict the pesticides recovery of crops of GC/MS using the molecular fingerprinting for the quality control, quality assurance and method development of food analysis.

Experimental

Data set of pesticide recovery of vegetables

Pesticide recovery rates were gathered from the literature [1], the samples (7 types of vegetables and fruits) were treated according to the procedure of Japan Positive List [2] as shown in Figure 1.The recovery rate was calculated with the solvent standard peak area, which was calculated by the calibration curve (CC) of 20 ppb, 50 ppb, 100 ppb, 200 ppb as expressed by the formula (1).

$$\text{Recovery rate (\%)} = \frac{\textit{Peak area of 50 ppb spiked in the sample}}{\textit{Peak area of 50 ppb in the solvent standrd CC}} \times 100 \quad (1)$$


[Cabbage, Spinach, Potato, Apple, Orange]

Sample 20g

Acetonitrile 50ml＋20ml

Homogenize

Aspiration filtration

Volume fixed at 100ml by Acetonitrile

20mL fraction (corresponding to 4g sample)

0.5mol/L Phosphoric acid buffer (pH 7) 20ml, NaCl10g

Shaking

Fraction of acetonitrile layer

Dehydration by Na_2SO_4

Concentration

Acetonitrile/Toluene (3:1) 2ml

Column (Envi-carb+NH2)

Acetonitrile/Toluene (3:1) 20ml

Concentration

Acetone/Hexane (1:1) 2ml

Sample for GC/MS 2mL/4g

[Brown rice, Soy bean]

Sample 10g

Water 10 mL (Keep for 15 minutes) Acetonitrile 50ml ＋20ml

Homogenize

Aspiration filtration

Volume fixed at 100ml by Acetonitrile

20mL fraction (corresponding to 2g sample)

0.5mol/L Phosphoric acid buffer (pH 7) 20ml, NaCl10g

Shaking

Fraction of acetonitrile layer

Acetonitrile 2mL

Apply to ODS column (all samples)

Dehydration by Na_2SO_4

Concentration

Acetonitrile/ Toluene (3:1) 2ml

Column (Envi-carb+NH2)

Acetonitrile/ Toluene (3:1) 20ml

Concentration

Acetone/Hexane (1:1) 2ml

Sample for GC/MS 2mL/4g

Figure 1 Sample preparation workflow for Japan Positive List

Experimental

Assignment of Molecular Descriptor (MD) using SMILES strings

SMILES strings (SMILES) from the PubChem website were added to the data set for 248 pesticides, which have unique SMILES and one chromatographic peak in GC/MS, i.e. we excluded the pesticides which have several chromatographic peaks for data consistency in present study. 224 Molecular Descriptors (MD) were added to the data set using the rcdk package of R program After deleting "N/A" descriptor values, the data set of 7 crops of 248 pesticides with 178 remaining MD (Table 1) were finally used for machine learning to build the prediction model of recovery rate.

Building and predicting the recovery rate by machine learning methods

89 machine learning methods of regression analysis (Table 2) were used for prediction of recovery rate of the pesticides, which includes 69 ordinary learning methods and 20 ensemble learning methods.

Table 1 178 molecular descriptors in present study

Descriptor Class	Descriptor (Description)
ALOGP Descriptor (2)	ALogP (Ghose-Crippen LogKow), ALogP2 (Square of ALogP)
APol Descriptor (1)	Apol (Sum of the atomic polarizabilities (including implicit hydrogens)
Aromatic Atoms Count Descriptor (1)	naAromAtom (Number of aromatic atoms)
Aromatic Bonds Count Descriptor (1)	nAromBond (Number of aromatic bonds)
Atom Count Descriptor (2)	nAtom (Number of atoms), nB (Number of boron atoms)
Autocorrelation Descriptor Charge (5)	ATSc1, ATSc2, ATSc3, ATSc4, ATSc5 (ATS autocorrelation descriptor, weighted by charges)
Autocorrelation Descriptor Mass (5)	ATSm1, ATSm2, ATSm3, ATSm4, ATSm5 (ATS autocorrelation descriptor, weighted by scaled atomic mass)
Autocorrelation Descriptor Polarizability (5)	ATSp1, ATSp2, ATSp3, ATSp4, ATSp5 (ATS autocorrelation descriptor, weighted by polarizability)
BCUT Descriptor (6)	BCUTw.1l (nhigh lowest atom weighted BCUTS), BCUTw.1h (nlow highest atom), BCUTc.1l (nhigh lowest partial charge), BCUTc.1h (nlow highest partial charge) BCUTp.1l (nhigh lowest polarizability), BCUTp.1h (nlow highest polarizability)
BPolDescriptor (1)	bpol (Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens))
Carbon Types Descriptor (9)	C1SP1 (Triply bound carbon bound to one other carbon), C2SP1 (Triply bound carbon bound to two other carbons), C1SP2 (Doubly hound carbon bound to one other carbon), C2SP2 (Doubly bound carbon bound to two other carbons), C3SP2 (Doubly bound carbon bound to three other carbons), C1SP3 (Singly bound carbon bound to one other carbon), C2SP3 (Singly bound carbon bound to two other carbons), C3SP3 (Singly bound carbon bound to three other carbons), C4SP3 (Singly bound carbon bound to four other carbons)
Chi Chain Descriptor (10)	SCH.3-7 (Simple chain, orders 3-7), VCH.3-7 (Valence chain, orders 3-7)
Chi Cluster Descriptor (8)	SC.3-6 (Simple cluster, orders 3-6) , VC.3-6 (Valence cluster, orders 3-6)
Chi Path Cluster Descriptor (6)	SPC.4-6 (Simple path cluster, orders 4 to 6), VPC.4-6 (Valence path cluster, orders 4-6)
Chi Path Descriptor (16)	SP.0-7 (Simple path, orders 0-7), VP.0-7Valence path, orders 0-7
Eccentric Connectivity Index Descriptor (38)	ECCEN (A topological descriptor combining distance and adjacency information), khs.sCH3 (Count of atom-type E-State: -CH3), khs.dCH2 (=CH2), khs.ssCH2 (-CH2-), khs.tCH (#CH), khs.dsCH (=CH-), khs.aaCH (:CH:), khs.sssCH (>CH-), khs.tsC (#C-), khs.dssC (=C<), khs.aasC (:C:-), khs.aaaC (::C:), khs.ssssC (>C<), khs.sNH2 (-NH2), khs.ssNH (-NH2-+), khs.aaNH (:NH:), khs.tN (#N), khs.sssNH (>NH-+),khs.dsN (=N-), khs.aaN (:N:), khs.sssN (>N-), khs.ddsN (-N<<), khs.aasN (:N:-), khs.sOH (-OH), khs.dO (=O), khs.ssO (-O-), khs.aaO (:O:), khs.sF (-F), khs.ssssSi (>Si<), khs.dsssP (->P=), khs.dS (=S), khs.ssS (-S-), khs.aaS (aSa), khs.dssS (>S=), khs.ddssS (>S==), khs.sCl (-Cl), khs.sBr (-Br)
Fragment Complexity Descriptor (1)	fragC (Complexity of a system)
H Bond Acceptor Count Descriptor (1)	nHBAcc (Number of hydrogen bond acceptors)
H Bond Donor Count Descriptor (1)	nHBDon (Number of hydrogen bond donors)
KappaShape Indices Descriptor (3)	Kier1-3 (First, Second, Third kappa (κ) shape indexes)
Largest Chain Descriptor (1)	nAtomLC (Number of atoms in the largest chain)
Longest Aliphatic Chain Descriptor (1)	nAtomLAC (Number of atoms in the longest aliphatic chain)
Mannhold LogP Descriptor (1)	MLogP (Mannhold LogP)
MDEDescriptor (19)	MDEC.11 (Molecular distance edge between all primary carbons), MDEC.12 (between all primary and secondary carbons), MDEC.13 (between all primary and tertiary carbons), MDEC.14 (between all primary and quaternary carbons), MDEC.22 (between all secondary carbons), MDEC.23 (between all secondary and tertiary carbons), MDEC.24 (between all secondary and quaternary carbons), MDEC.33 (between all tertiary carbons), MDEC.34 (between all tertiary and quaternary carbons), MDEC.44 (between all quaternary carbons), MDEO.11 (between all primary oxygens), MDEO.12 (between all primary and secondary oxygens), MDEO.22 (between all secondary oxygens), MDEN.11 (between all primary nitrogens), MDEN.12 (between all primary and secondary nitrogens), MDEN.13 (between all primary and tertiary niroqens), MDEN.22 (between all secondary nitroqens), MDEN.23 (between all secondary and tertiary nitrogens), MDEN.33 (between all tertiary nitrogens)
PetitjeanNumberDescriptor (1)	PetitjeanNumber (Petitjean number)
RotatableBondsCountDescriptor (1)	nRotB (Number of rotatable bonds, excluding terminal bonds)
RuleOfFiveDescriptor (1)	LipinskiFailures (Number failures of the Lipinski's Rule Of 5)
TPSADescriptor (19)	TopoPSA (Topological polar surface area)
VAdjMaDescriptor (1)	VAdjMat (Vertex adjacency information (magnitude))
WeightDescriptor (1)	MW (Molecular weight)
WeightedPathDescriptor (5)	WTPT.1 (Molecular ID), WTPT.2 (Molecular ID / number of atoms), WTPT.3 (Sum of path lengths starting from heteroatoms), WTPT.4 (Sum of path lengths starting from oxygens), WTPT.5 (Sum of path lengths starting from nitrogens)
WienerNumbersDescriptor (2)	WPATH (Weiner path number), WPOL (Weiner polarity number)
XLogPDescriptor (1)	XLogP (XLogP)
ZagrebIndexDescriptor (1)	Zagreb (Sum of the squares of atom degree over all heavy atoms i)
Petitjean Shape Index Descriptor (1)	topoShape (Petitjean topological shape index)
Others (17)	nAcid (Acidic group count descriptor), nBase (Basic group count descriptor), nSmallRings (the number of small rings from size 3 to 9), nAromRings (the number of aromatic rings), nRingBlocks (total number of distinct ring blocks), nAromBlocks (total number of "aromatically connected components"), nRings3, 5, 6, 7 (individual breakdown of small rings), tpsaEfficiency (Polar surface area expressed as a ratio to molecular size), VABC (Atomic and Bond Contributions of van der Waals volume), HybRatio (the ratio of heavy atoms in the framework to the total number of heavy atoms in the molecule.), tpsaEfficiency.1 (Polar surface area expressed as a ratio to molecular size), TopoPSA.1 (Topological polar surface area), topoShape.1(A measure of the anisotropy in a molecule)

Experimental

Table 2 Machine Learning methods for regression analysis used in present study

Algorithm	Methods in caret
(a) Ordinary learning methods	
Kernel (17)	gaussprRadial, gaussprPoly, krlsPoly, gaussprLinear, krlsRadial, rvmLinear, rvmRadial, rvmPoly, svmRadial, svmRadialCost, svmRadialSigma, svmLinear, svmLinear2, svmPoly, svmLinear3, kernelpls (PLS), widekernelpls (PLS)
Simple Linear (16)	lm, leapSeq, leapForward, leapBackward, lmStepAIC, bridge, bayesglm (GLM), glmStepAIC (GLM), icr (ICA), pcr (PCA), superpc (PCA), superpc (PCA), nnls (PLS), simpls (PLS), pls (PLS), plsRglm (PLS, GLM), glm (GLM)
Sparse modeling (11)	penalized, blassoAveraged, foba, ridge, relaxo, lasso, Blasso, lars, lars2, glmnet, enet
Neural Network (9)	rbfDDA, dnn, neuralnet, brnn, mlpML, mlp, mlpWeightDecay, msaenet, monmlp
Decision Tree (8)	rpart2, rpart1SE, ctree, ctree2, evtree, M5Rules, M5, WM
Centroid,kNN (3)	knn, kknn, SBC
Spline (2)	gcvEarth, earth
Others (3)	ppr, spikeslab, xyf (LVQ)
(b) Ensemble learning methods	
Decision Tree (14)	cforest, ranger, qrf, rf, parRF, extraTrees, Rborist, RRFglobal, RRF, treebag, bstTree, gbm, xgbTree, nodeHarvest
Simple Linear (3)	BstLm, glmboost (GLM), xgbLinear
Spline (3)	bagEarthGCV, bagEarth, xgbDART

Results

Correlation between recovery rate of pesticide and molecular descriptors.

Before building the prediction model for regression, the Pearson correlation coefficient between pesticide recovery rate and MD were investigated. The coefficients ranged between -0.254 and 0.523 as shown in the Figure 2, i.e. weak correlation between recovery rate and each single MD. Figure 3 and 4 shows examples of top 6 positive and negative MDs. The result showed that the recovery rate cannot be predicted by **any single MD**.

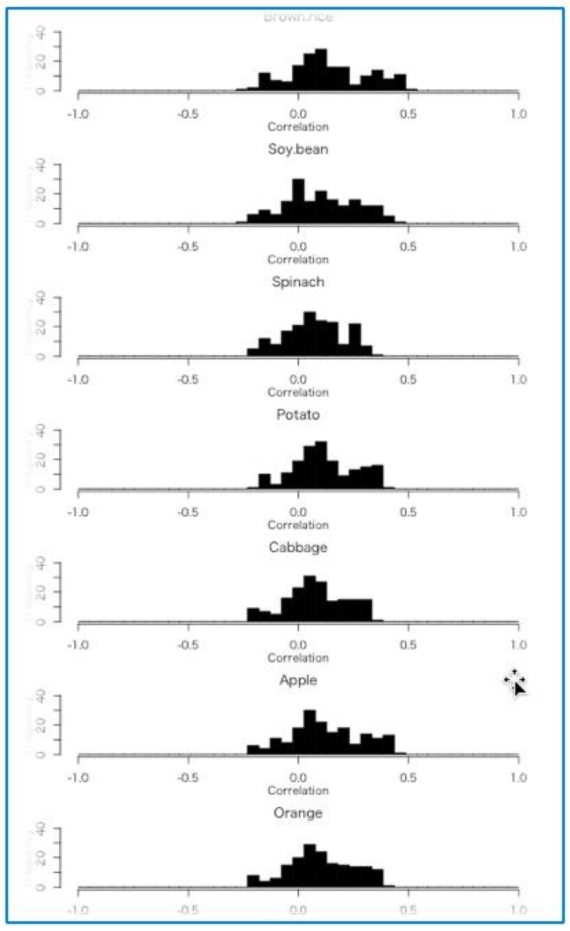


Figure 2 Histogram of recovery rate and MD for seven crops

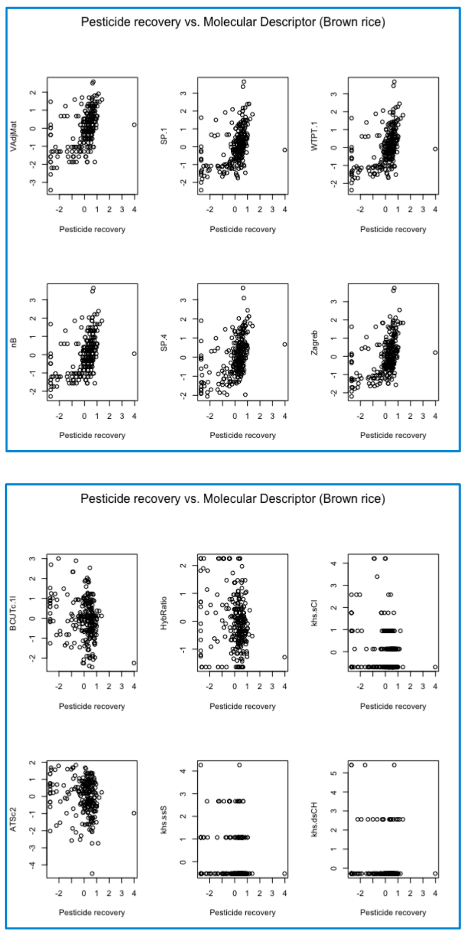


Figure 3 and 4 Scatter plots of top 6 positive and negative correlation coefficient MD for brown rice

Results

Building and evaluation of regression model for pesticide recovery rate using machine learning method

For the metrics of machine learning method performance, Prediction Error (PE) calculated by the formula (2) with the 10-fold cross validation, the time for building the model Execution Time (ET: in sec) and the generalization performance of prediction error PE$_k$ calculated by (3) were used.

Prediction Error (PE) $$PE_j = \frac{\sum_{i=1}^{N}\left(y_{obs}^{(ij)} - y_{pred}^{(ij)}\right)^2}{\sum_{i=1}^{N}\left(y_{obs}^{(ij)} - \bar{y}^{(j)}\right)^2} \qquad (2)$$

where, $y_{obs}^{(ij)}$ is the actual recovery rate, $y_{pred}^{(ij)}$ is the recovery rate from the prediction model, and $\bar{y}^{(j)}$ is the average of recovery ratio of that crop.

Generalization Performance Index (PE$_k$)

$$PE_k = \frac{\sum_{j=1}^{M} PE_j^{(k)}}{M} \qquad (3)$$

where $PE_j^{(k)}$ is the rank of the machine learning method in that crop, k is the method ID in the caret and M is the number of crop, 7 for the present study.

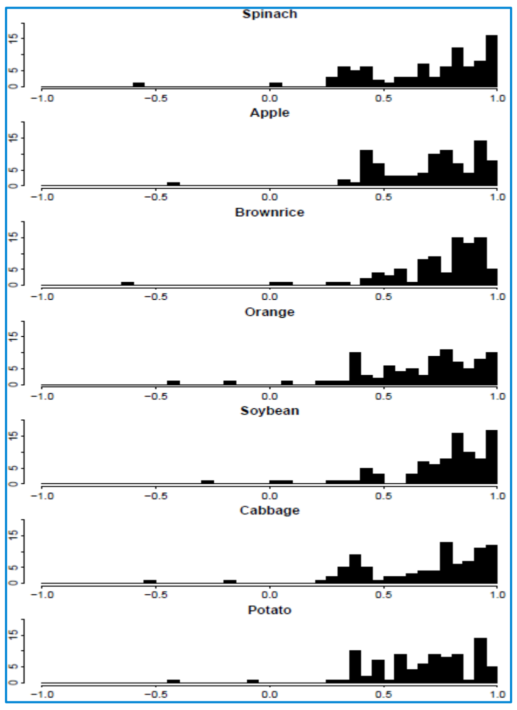


Figure 5 Histogram of correlation coefficient between actual and predicted pesticide recovery

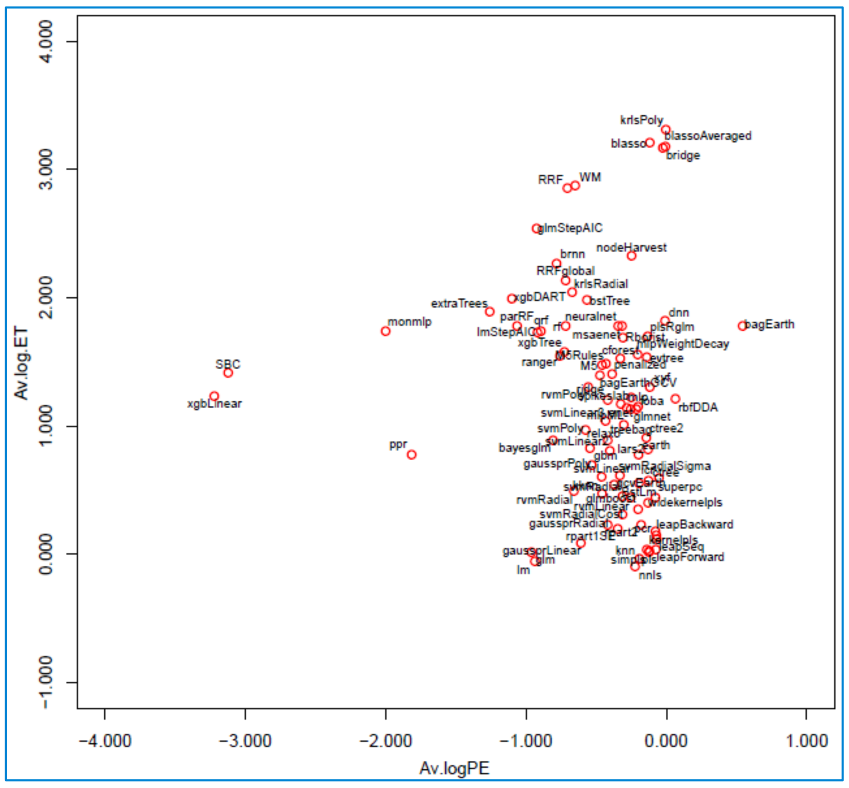


Figure 6 Prediction Error (PE) and Execution Time (ET) for 89 machine learning methods.

Discussion

PE and ET results: Best machine learning methods

Figure 5 shows that some machine learning methods have high correlation coefficients between the actual pesticide recovery rate and predicted recovery rate, i.e. good prediction performance. PE and ET of 7 crops average in log scale were shown in Figure 6. ET were ranged from 0.83 sec to 7,394 sec (approx. 2 hours).

The top 20 machine learning methods of PE in 7 crops average were shown in Figure 7. Four excellent methods (SBC, xgbLinear, monmlp, ppr) were listed as the candidates of best methods for the present study.

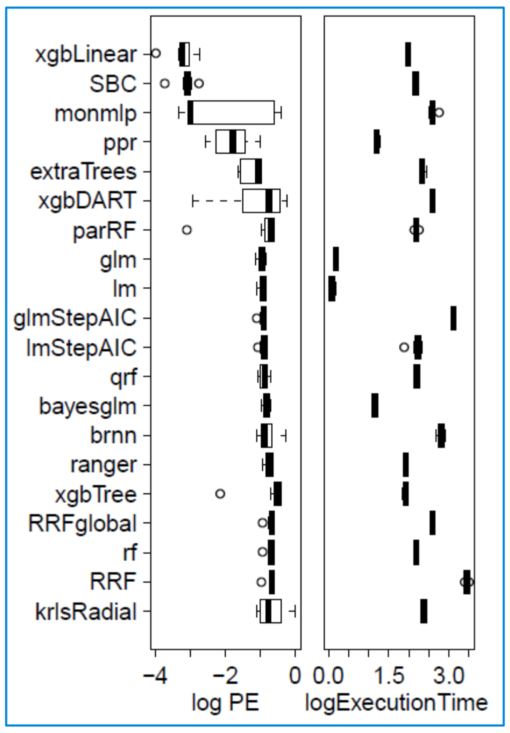


Figure 7 Top 20 PE machine learning methods and their ET

Figure 7 shows some machine learning methods have variance by the crops.

The generalization performance index and normalized PE are shown in Figure 8. Please note that 6 machine learning methods (rbfDDA, bridge, blassoAveraged, bagEarth, lars, lasso) with PE > 1.0 were excluded, since these models do not yield a meaningful prediction by definition.

Figure 8 indicates that **SBC** and **xgbLinear** are excellent performance in both PE and PE_k.

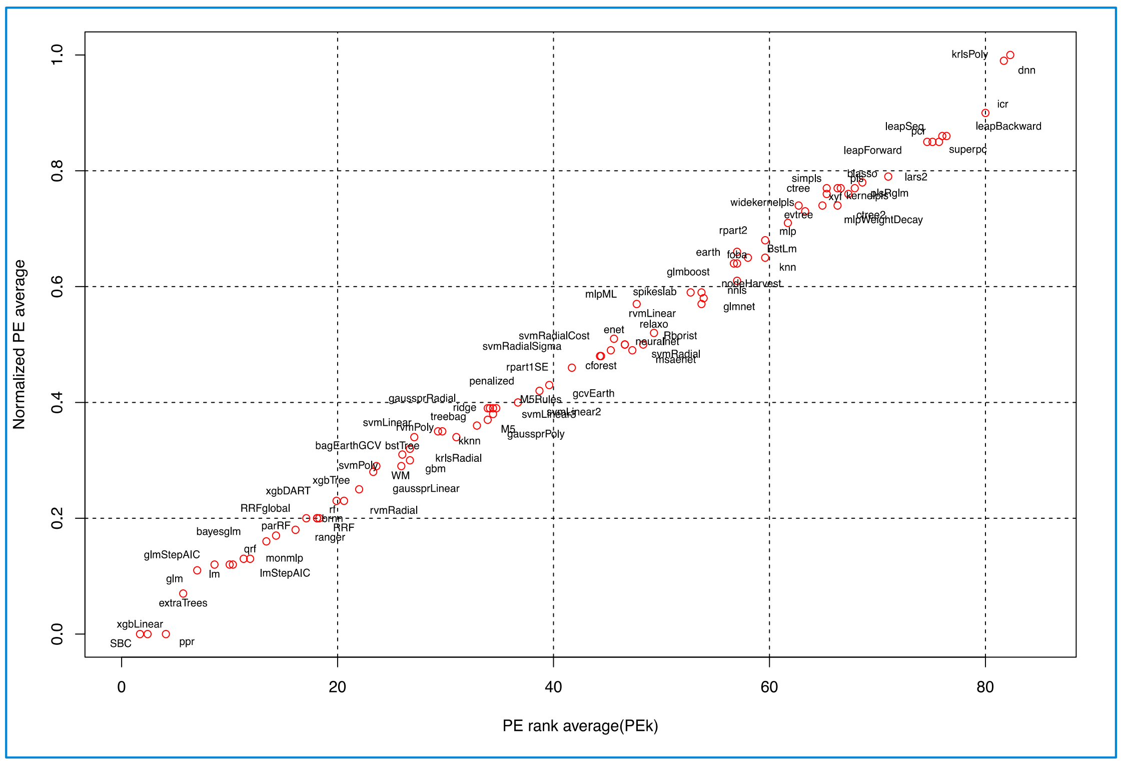


Figure 8 Prediction Error and generalization performance index (PE_k)

PE and ET in Machine learning category

The 89 machine learning methods were classified in 8 categories of ordinary machine learning methods and ensemble machine learning methods.

Among the ordinary machine learning methods, the “Centroid kNN” category gives a better PE than others. The “Simple linear” and “Neural Network” are simple and common methods, however their PE were not excellent in the present study.

Among the ensemble learning methods, DT (Decision Tree) and Spline required longer ET with worse PE than the Spline Linear category.

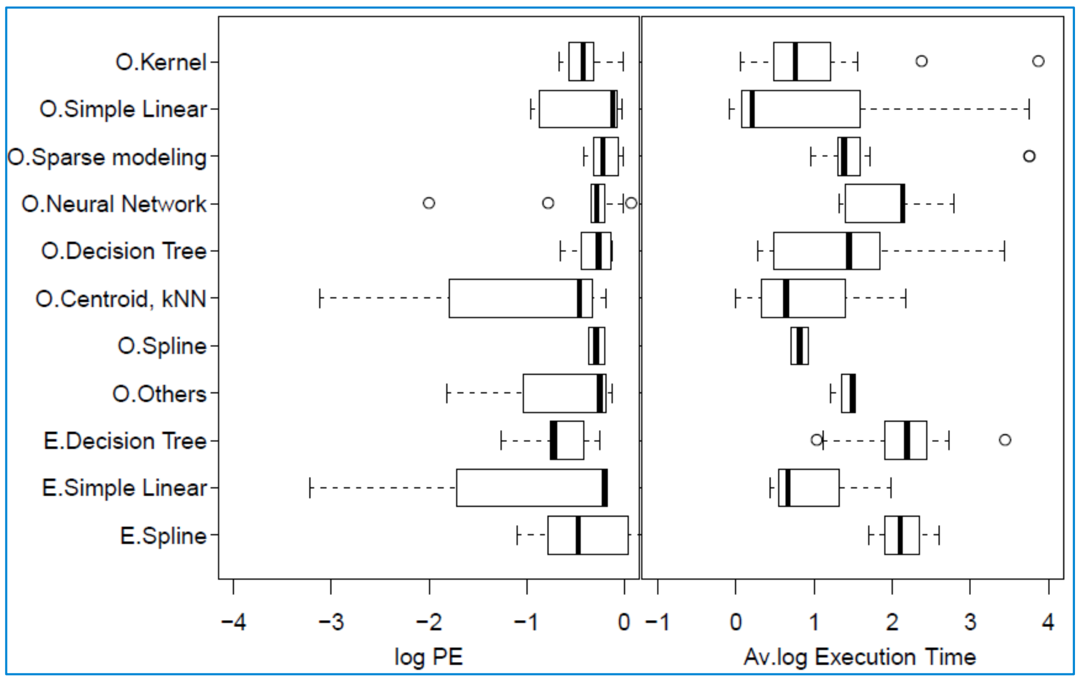


Figure 9 PE and ET of machine learning category

Conclusions

We developed the method for prediction of pesticides recovery using the machine learning. Various machine learning methods have been developed and available online, but the performance of prediction and execution time vary. The PE, ET and PE_k are the metrics to evaluate the optimum machine learning methods for prediction.

The **SBC** (<u>S</u>u<u>b</u>tractive <u>C</u>lustering and Fuzzy c-Means Rules) of Centroid kNN category and **xgbLinear** (e<u>X</u>treme <u>G</u>radient <u>B</u>oosting <u>Linear</u>) of Ensemble Spline Linear category are the optimum machine learning methods for predicting the pesticide recovery rate of present study.

References

[1] Sadao NAKAMURA, Takashi YAMAGAMI, Yukiko ONO, Kenichi TOUBOU and Shigeki DAISHIMA, Multi-residue Analysis of Pesticides in Agricultural Products by GC/MS Using Synchronous SIM/Scan Acquisition, BUNSEKI KAGAKU 62, 229-241(2013)

[2] https://www.mhlw.go.jp/english/topics/foodsafety/positivelist060228/dl/02-01.pdf

This information is subject to change without notice.

Published in USA, June 2, 2019