Comparison of Nonparametric Regression and Scaling Factor Methods for Difficult-to-Measure Radionuclides

Young Ho Roh [a], Yong Kyun Kim [a*]
[a]*Department of Nuclear Engineering, Hanyang University, Seoul, Korea*
[*]*Corresponding author: ykkim4@hanyang.ac.kr*

Keywords : Difficult-to-measure radionuclide, Scaling factor, Nonparametric regression, Radionuclide inventory, Radioactive Waste

1. Introduction

Disposal of low- and intermediate-level radioactive waste in Korea requires the determination of activity concentrations for 14 nuclides [1]. Scaling factor (SF) methods are widely used to evaluate radionuclide inventories in low- and intermediate-level radioactive waste. These methods indirectly estimate difficult-to-measure (DTM) radionuclide concentrations from the concentration of an easily measured key nuclide [1].

SF methods represent the concentration relationship between nuclides using a single log-linear regression equation or a constant concentration ratio [1]. Their predictive accuracy may therefore be limited when the relationship cannot be adequately described by a single straight line in log space.

This study compared SF with five nonparametric regression models to assess whether more flexible fitting of concentration relationships improves predictive performance.

2. Methods and Results

2.1 Dataset and Study Design

Radioactive waste data from pressurized water reactors were obtained from Electric Power Research Institute (EPRI) report NP-5077 [2]. Cs-137 was used as the common key nuclide to predict the concentrations of the fission products Sr-90, Tc-99, and I-129. The datasets comprised 445 samples for Sr-90, 448 for Tc-99, and 446 for I-129.

The SF method used log-linear regression when $r \geq 0.6$ (Eq. (1)) and the geometric mean of DTM-to-key concentration ratios when $r < 0.6$ (Eq. (2)) [1].

$$\log_{10}\hat{y} = a + b \cdot \log_{10} x_{Cs} \; (r \geq 0.6) \qquad (1)$$

$$\hat{y} = \left(\prod_i \frac{y_i}{x_i}\right)^{\frac{1}{n}} \cdot x_{Cs} \; (r < 0.6) \qquad (2)$$

where $\hat{y}$ is the predicted DTM nuclide concentration, x_{Cs} is the Cs-137 concentration, a and b are the regression intercept and slope, and r is the Pearson correlation coefficient between the $\log_{10}$ concentrations. In Eq. (2), y_i and x_i are the DTM nuclide and Cs-137 concentrations of sample i, and n is the number of samples used to fit the SF method.

Five models were used for comparison: random forest (RF), Gaussian process (GP), isotonic regression, locally weighted scatterplot smoothing (LOWESS), and k-nearest neighbors (kNN) [3–7]. LOWESS used a neighborhood fraction of 2/3, kNN used five neighbors, and RF used 300 trees with a minimum leaf size of five. GP used a radial basis function-based kernel.

2.2 Evaluation Methods

The performance of all prediction models was evaluated using leave-one-out cross-validation (LOOCV). In each iteration, the models were trained using all samples except one and then used to predict the concentration of the excluded sample [8].

Predictive performance was evaluated using the percentage of predictions within a factor of 10 of the measured value, satisfying Eq. (3), and the mean absolute $\log_{10}$ error (MALE), calculated using Eq. (4).

$$|log_{10}\hat{y}_i - log_{10}y_i| \leq 1 \qquad (3)$$

$$MALE = \frac{1}{N}\sum_{i=1}^{N} |log_{10}\hat{y}_i - log_{10}y_i| \qquad (4)$$

In Eqs. (3) and (4), $\hat{y}_i$ and y_i denote the predicted and measured DTM nuclide concentrations of sample *i, N* is the number of evaluated samples.

2.3 Predictive Performance

Table I presents the predictive performance for each DTM nuclide obtained through LOOCV. Performance differences between SF and nonparametric models varied with r between the $\log_{10}$ concentrations of the DTM nuclide and Cs-137. For Sr-90 (r = 0.78) and Tc-99 (r = 0.62), SF used log-linear regression because $r \geq 0.6$. Most models showed only minor differences from SF in the percentage of predictions within a factor of 10 and MALE (Fig. 1).

Table I: Predictive performance of SF and five nonparametric regression models under LOOCV

Method	Sr-90 (N = 445)		Tc-99 (N = 448)		I-129 (N = 446)	
	FO10 [%]	MALE	FO10 [%]	MALE	FO10 [%]	MALE
SF	62.5	0.913	69.4	0.799	40.4	1.519
RF	61.1	0.911	68.1	0.823	71.1	0.787
GP	62.5	0.901	69.9	0.786	75.6	0.751
Isotonic	62.0	0.895	69.2	0.785	74.7	0.753
LOWESS	62.5	0.900	70.8	0.784	75.3	0.752
kNN	55.5	1.035	63.6	0.921	65.2	0.893

FO10, percentage of predictions within a factor of 10 of the measured value. Differences were computed from unrounded rates.

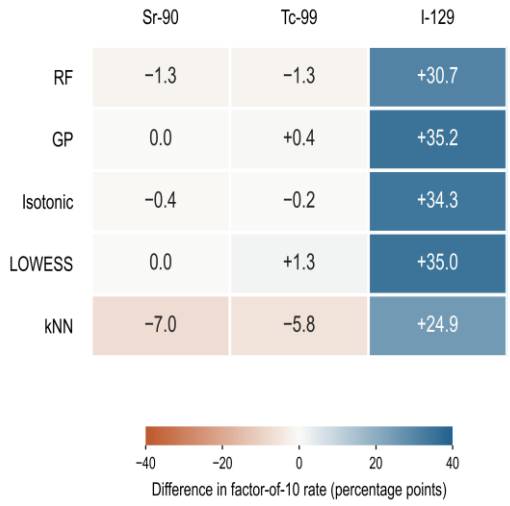

Fig. 1. Difference in the percentage of predictions within a factor of 10 relative to SF. Positive values indicate higher rates than SF; differences are expressed in percentage points.

For I-129 (r = 0.55), SF used the geometric mean concentration ratio because $r < 0.6$. All five nonparametric models improved predictive performance. The percentage of predictions within a factor of 10 increased by 24.9 to 35.2 percentage points relative to SF, and MALE also decreased (Table I; Fig. 1). Fig. 2 compares the predicted and measured concentrations of Sr-90, Tc-99, and I-129 for SF and LOWESS.

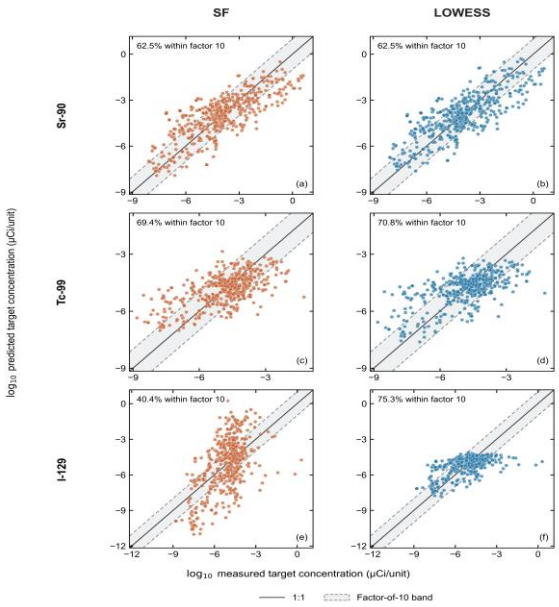

Fig. 2. Predicted versus measured DTM nuclide concentrations for SF and LOWESS.

3. Conclusions

SF and five nonparametric regression models were compared using LOOCV to predict Sr-90, Tc-99, and I-129 concentrations. For Sr-90 and Tc-99, most models showed only minor differences in predictive performance relative to SF. For I-129, SF used the geometric mean concentration ratio, and nonparametric regression improved both the percentage of predictions within a factor of 10 and MALE.

Nonparametric regression can therefore improve predictive performance relative to SF when a constant concentration ratio does not adequately represent the relationship between nuclides.

REFERENCES

[1] T.-H. Kim et al., Statistical Methodologies for Scaling Factor Implementation: Part 1. Overview of Current Scaling Factor Method for Radioactive Waste Characterization, Journal of Nuclear Fuel Cycle and Waste Technology 18 (2020) 517–536. https://doi.org/10.7733/jnfcwt.2020.18.4.517

[2] W. T. Best and A. D. Miller, Updated Scaling Factors in Low-Level Radwaste, EPRI NP-5077, Final Report, 1987.

[3] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324

[4] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006. https://gaussianprocess.org/gpml/

[5] R. E. Barlow and H. D. Brunk, The Isotonic Regression Problem and its Dual, Journal of the American Statistical Association 67 (1972) 140–147. https://doi.org/10.1080/01621459.1972.10481216

[6] W. S. Cleveland, Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association 74 (1979) 829–836. https://doi.org/10.1080/01621459.1979.10481038

[7] C. J. Stone, Consistent Nonparametric Regression, The Annals of Statistics 5 (1977) 595–620. https://doi.org/10.1214/aos/1176343886

[8] S. Arlot and A. Celisse, A Survey of Cross-Validation Procedures for Model Selection, Statistics Surveys 4 (2010) 40–79. https://doi.org/10.1214/09-SS054