ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202412.00064

Prediction of radionuclide diffusion enabled by missing data imputation and ensemble machine learning

Authors: Tian, Mr. Junlei, Feng, Mr. Jiaxing, Shen, Jiacong, Yao, Mr. Lei, Wang, Jingyan, WU, Dr. Tao, Zhao, Dr. Yaolin 赵耀林, Tian, Junlei, Jiaxing Feng, Shen, Mr. Jiacong, Yao, Mr. Lei, Wang, Mr. Jingyan, Dr. Yaolin Zhao

Date: 2024-12-03T00:00:00+00:00

Abstract

Missing values in radionuclide diffusion datasets can undermine the predictive accuracy and robustness of machine learning (ML) models. A regression-based missing data imputation method using light gradient boosting machine (LGBM) algorithm was employed to impute over 60% of the missing data, establishing a radionuclide diffusion dataset containing 16 input features and 813 instances. The effective diffusion coefficient (De) was predicted using ten ML models. The predictive accuracy of ensemble meta-models, namely LGBM-extreme gradient boosting (XGB) and LGBM-categorical boosting (CatB), surpassed the other ML models, with R2 values of 0.94. The models were applied in predicting the De values of EuEDTA- and HCrO4- in saturated compacted bentonites at compaction ranged from 1200 kg/m3 to 1800 kg/m3, which was measured using a through-diffusion method. The generalization ability of LGBM-XGB model surpassed that of LGB-CatB in predicting the De of HCrO4-. Shapley additive explanations identified the total porosity as the most significant influencing factor. In addition, the partial dependence plot analysis technique showed clearer results for univariate correlation analysis. This study provides a regression imputation technique to refine radionuclide diffusion datasets, offering a deeper insight into analyzing the diffusion mechanism of radionuclide and supporting the safety assessment of the geological disposal of high-level radioactive waste.

Full Text

Supporting Information

Prediction of Radionuclide Diffusion Enabled by Missing Data Imputation and Ensemble Machine Learning

Jun-Lei Tian[1], Jia-Xing Feng[1], Jia-Cong Shen[1], Lei Yao[1], Jing-Yan Wang[1], Tao Wu[1],, *and Yao-Lin Zhao*[2],

[1]Huzhou Key Laboratory of Environmental Functional Materials and Pollution Control, Huzhou University, Huzhou 313000, China
[2]School of Nuclear Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

*Corresponding authors: Tao Wu and Yao-Lin Zhao
Email: twu@zjhu.edu.cn (Tao Wu), zhaoyaolin@mail.xjtu.edu.cn (Yao-Lin Zhao)
Phone: 13655729651
Address: Huzhou University, No. 759 East Erhuan Road, Wuxing District, Huzhou, Zhejiang Province, China, 313000

Physical Parameters

Physical parameters related to the diffusion process can enhance our understanding of diffusion mechanisms. In this study, five physical parameters—including total porosity, ion molar conductivity, ionic radius, and montmorillonite stacking number—were incorporated into the datasets to improve predictive performance and provide deeper insights into the diffusion mechanism. Specifically, the total porosity (tot) is related to the grain density of bentonite (d) and compacted dry density (s), and is defined as:

tot = 1 −

The relationship between the rock capacity factor (α) and porosity ( ) is expressed by:

$\alpha =$ + Kd · d,

The anionic exclusion effect has been recognized by numerous diffusion models [1, 2, 3]. This phenomenon assumes that radionuclide anions cannot enter the interlayer pores of compacted bentonite due to electrostatic repulsion from the negatively charged montmorillonite layers. Consequently, the rock capacity factor for radionuclide anions should be less than the total porosity. However, the calculation of the rock capacity factor in JAEA-DDB neglects the anionic exclusion effect, leading to overestimation for radionuclide anions. In this study, the anionic exclusion effect was properly considered: = tot in Eq. (S1) was employed for radionuclides with Kd > 0, whereas = 1–0.39 · d was used for those with Kd = 0.

The ion molar conductivity (λ) is expressed by the following equation [5]:

Dw · F 2 · |z| R · T

where F is the Faraday constant (96500 C/mol), z is the ion charge, T is the Kelvin temperature (K), and R is the molar gas constant (8.314 J/mol · k).

Micro-parameters such as the ionic radius and montmorillonite stacking number were integrated into the datasets to establish a connection between mesoscopic and microscopic features. The ionic radius (r) is calculated using the classic Stokes-Einstein relation [6]:

k ⋅ T 6 ⋅ π ⋅ ⋅ Dw

where k is the Boltzmann constant (1.380649 × 10^{23} Pa ⋅ m^3/K), and is the viscosity of the medium.

The montmorillonite stacking number is approximately given as [7]:

MW ⋅ Aext ⋅ na ⋅ nb − 2 ⋅ a ⋅ c∗ ⋅ na ⋅ NA − 2 ⋅ b ⋅ c∗ ⋅ nb ⋅ NA

where NA is Avogadro's number (6.022 × 10^{23} molecules/mol). The chemical molecular weight (MW) of a typical montmorillonite is 735 g/mol. The unit-cell dimension of a montmorillonite layer is a × b × c∗ = 0.523 × 0.905 × 0.96 nm^3. The length and width of a montmorillonite layer, na and nb, were approximately

[TABLE:S1]

Table S1 The statistical information of Dataset I (316 instances).

Parameters	Min	Max	Mean	Std	Skw
Ionic strength (I, mol/L)	0.00	5.50	0.45	1.04	2.77
Compacted dry density (d, kg/m^3)	400	2000	1470.47	447.75	-0.47
Montmorillonite content (m, –)	0.00	1.00	0.67	0.33	-0.77
Distribution coefficient (Kd, m^3/kg)	0.00	1.00	0.12	0.28	2.18
Ionic charge (z, –)	-2.00	3.00	0.16	1.69	-0.21
Temperature (T, °C)	5.00	25.00	20.71	5.18	-1.53
Specific surface area (Aext, m^2/g)	0.00	800.00	347.53	247.78	0.06
Grain density (s, kg/m^3)	2650	2800	2754.43	53.69	-0.69
Molecular weight (MW, g/mol)	18.00	735.00	377.95	358.53	-0.53
Species diffusion coefficient in water (LogDw, –)	-9.30	-8.34	-8.79	0.18	-0.82
Total porosity (tot, –)	0.25	0.85	0.47	0.16	0.06

Parameters	Min	Max	Mean	Std	Skw
Rock capacity factor (α, –)	0.00	0.85	0.35	0.31	0.53
Ionic radius (Logr, –)	-10.24	-9.34	-9.84	0.18	-0.21
Ion molar conductivity (λ, $m^2 \cdot S/mol$)	0.00	14.98	7.48	5.27	-0.06
Montmorillonite stacking number (nc, –)	1.00	10.00	5.50	2.87	0.00
pH (–)	6.00	12.00	7.93	1.53	0.69

Std = Standard Deviation; Skw = Skewness

[TABLE:S2]

Table S2 The statistical information of Dataset III (813 instances).

Parameters	Min	Max	Mean	Std	Skw
Ionic strength (I, mol/L)	0.00	5.50	0.47	1.04	2.77
Compacted dry density (d, kg/m^3)	400	2000	1470.47	447.75	-0.47
Montmorillonite content (m, –)	0.00	1.00	0.67	0.33	-0.77
Distribution coefficient (Kd, m^3/kg)	0.00	1.00	0.12	0.28	2.18
Ionic charge (z, –)	-2.00	3.00	0.16	1.69	-0.21
Temperature (T, °C)	5.00	25.00	20.71	5.18	-1.53
Specific surface area (Aext, m^2/g)	0.00	800.00	347.53	247.78	0.06
Grain density (s, kg/m^3)	2650	2800	2754.43	53.69	-0.69
Molecular weight (MW, g/mol)	18.00	735.00	377.95	358.53	-0.53
Species diffusion coefficient in water (LogDw, –)	-9.30	-8.24	-8.79	0.18	-0.82
Total porosity (tot, –)	0.25	0.85	0.47	0.16	0.06

Parameters	Min	Max	Mean	Std	Skw
Rock capacity factor (α, –)	0.00	0.85	0.35	0.31	0.53
Ionic radius (Logr, –)	-10.31	-9.33	-9.84	0.18	-0.21
Ion molar conductivity (λ, $m^2 \cdot S/mol$)	0.00	14.98	7.48	5.27	-0.06
Montmorillonite stacking number (nc, –)	1.00	10.00	5.50	2.87	0.00
pH (–)	6.00	12.00	7.93	1.53	0.69
Effective diffusion coefficient (LogDe, –)	-12.57	-8.54	-10.05	0.53	0.21

Std = Standard Deviation; Skw = Skewness

[TABLE:S3]

Table S3 Hyperparameters, tuning range of the machine learning models, and the optimized hyperparameters for Dataset I.

Algorithms	Parameter name	Range	Dataset I
LGBM-CatB	Max_{depth}	1-100	15
	Learning_{rate}	0.01-0.1	0.05
	Min_{{{data}}*{{in}}}{leaf}	1-100	20
	Feature_{fraction}	0.1-1.0	0.8
	Bagging_{freq}	1-100	5
	Bagging_{seed}	1-100	50
	Bagging_{fraction}	0.01-1.0	0.8
	Lambda_{l1}	0.0-1.0	0.0
	Lambda_{l2}	0.0-1.0	0.0
LGBM-XGB	Max_{depth}	1-100	10
	Learning_{rate}	0.01-0.5	0.1
	Min_{{{data}}*{{in}}}{leaf}	1-100	20
	Feature_{fraction}	0.1-1.0	0.8
	Bagging_{freq}	1-100	5
	Bagging_{seed}	1-100	50
	Bagging_{fraction}	0.01-1.0	0.8
	Lambda_{l1}	0.0-1.0	0.0
	Lambda_{l2}	0.0-1.0	0.0
CatB	Depth	1-100	6
	Learning_{rate}	0.01-0.5	0.1

Algorithms	Parameter name	Range	Dataset I
XGB	Iterations	1-100	50
	Subsample	0.01-1.0	0.8
	L2_{{{leaf}}_{{reg}}}	0.0-1.0	0.0
	Random_{seed}	1-100	50
	Max_{depth}	1-100	6
	N_{estimators}	1-100	50
	Gamma	0.0-1.0	0.0
	Lambda	0.0-1.0	0.0
	Subsample	0.01-1.0	0.8
	Reg_{alpha}	0.0-1.0	0.0
	Min_{{{child}}_{{weight}}}	1-100	1
RF	N_{estimators}	1-100	50
	Max_{depth}	1-100	6
	Min_{{{samples}}_{{split}}}	1-100	2
	Min_{{{samples}}_{{leaf}}}	1-100	1
	Min_{{{weight}}*{{fraction}}*{leaf}	0.0-0.5	0.0
	Random_{state}	1-100	50

[TABLE:S4]

Table S4 The optimized hyperparameters used in machine learning models for Dataset II and Dataset III.

Algorithms	Parameter name	Dataset II	Dataset III
LGBM-CatB	Max_{depth}	15	15
	Learning_{rate}	0.05	0.05
	Num_{leaves}	31	31
	Min_{{{data}}*{{in}}*{leaf}	20	20
	Feature_{fraction}	0.8	0.8
	Bagging_{freq}	5	5
	Bagging_{seed}	50	50
	Bagging_{fraction}	0.8	0.8
	Lambda_{l1}	0.0	0.0
	Lambda_{l2}	0.0	0.0
LGBM-XGB	Max_{depth}	10	10
	Learning_{rate}	0.1	0.1
	Num_{leaves}	31	31
	Min_{{{data}}*{{in}}*{leaf}	20	20
	Feature_{fraction}	0.8	0.8
	Bagging_{freq}	5	5
	Bagging_{seed}	50	50
	Bagging_{fraction}	0.8	0.8
	Lambda_{l1}	0.0	0.0

Algorithms	Parameter name	Dataset II	Dataset III
	Lambda_{l2}	0.0	0.0
LGBM	Max_{depth}	15	15
	Learning_{rate}	0.05	0.05
	Num_{leaves}	31	31
	Min_{{{data}}*{{in}}}{leaf}	20	20
	Feature_{fraction}	0.8	0.8
	Bagging_{freq}	5	5
	Bagging_{seed}	50	50
	Bagging_{fraction}	0.8	0.8
	Lambda_{l1}	0.0	0.0
	Lambda_{l2}	0.0	0.0
CatB	Depth	6	6
	Learning_{rate}	0.1	0.1
	Iterations	50	50
	Subsample	0.8	0.8
	L2_{{{leaf}}_{{reg}}}	0.0	0.0
	Random_{seed}	50	50
XGB	Max_{depth}	6	6
	N_{estimators}	50	50
	Gamma	0.0	0.0
	Lambda	0.0	0.0
	Subsample	0.8	0.8
	Reg_{alpha}	0.0	0.0
	Min_{{{child}}_{{weight}}}	1	1
	Colsample_{bytree}	0.8	0.8
ANN	Epochs	100	100
	Learning_{rate}	0.001	0.001
	Hidden layers	2	2
	Number of neurons	64	64
	Activation_{function}	PReLU	PReLU
	Dropout	0.2	0.2
DNN	N-epoch	100	100
	Learning_{rate}	0.001	0.001
	Hidden layers	3	3
	Kernel_{{{Regularizer}}_{{L2}}}	0.01	0.01
	Number of neurons	128	128
	Activation_{function}	PReLU	PReLU
RF	N_{estimators}	50	50
	Max_{depth}	6	6
	Min_{{{samples}}_{{split}}}	2	2
	Min_{{{samples}}_{{leaf}}}	1	1
	Min_{{{weight}}*{{fraction}}}{leaf}	0.0	0.0
	Random_{state}	50	50
SVM	Cache_{size}	200	200

Algorithms	Parameter name	Dataset II	Dataset III
	Gamma	0.1	0.1
	Kernel	RBF	RBF
	Epsilon	0.1	0.1

[TABLE:S5]

Table S5 Mean values of performance metrics utilizing five-fold cross-validation.

Algorithms	Datasets	MSEcv (Training)	MSEcv (Validation)
LGBM-CatB	Dataset II (n=316)	0.021	0.045
LGBM-XGB	Dataset II (n=316)	0.023	0.048
LGBM-RF	Dataset II (n=316)	0.025	0.052
LGBM-CatB	Dataset III (n=813)	0.018	0.038
LGBM-XGB	Dataset III (n=813)	0.020	0.041
LGBM-RF	Dataset III (n=813)	0.022	0.044

[TABLE:S6]

Table S6 Detailed statistics of the additional instances from four reported literatures.

Species	I (mol/L)	d (kg/m^3)	m (–)	Kd (m^3/kg)	z (–)	pH (–)	T (°C)	Aext (m^2/g)	s (kg/m^3)	MW (g/mol)	Dw (m^2/s)	tot (–)	α (–)	r (Å)	λ (m^2·S/mol)	nc (–)	De (m^2/s)	Reference
CeEDTA	1×10^{-3}	1600	0.75	0	-1	7.5	25	562	2780	404	5.48×10^{-10}	0.42	0.42	1.92×10^{-11}	7.48×10^{-3}	3.50	1.17×10^{-11}	[1]
CeEDTA	1×10^{-3}	1600	0.75	0	-1	7.5	25	562	2780	404	5.48×10^{-10}	0.42	0.42	1.92×10^{-11}	7.48×10^{-3}	3.50	1.09×10^{-11}	[1]
CeEDTA	1×10^{-4}	1600	0.75	0	-1	7.5	25	562	2780	404	5.48×10^{-10}	0.42	0.42	1.92×10^{-11}	7.48×10^{-3}	3.50	8.8×10^{-12}	[1]
CeEDTA	1×10^{-4}	1600	0.75	0	-1	7.5	25	562	2780	404	5.48×10^{-10}	0.42	0.42	1.92×10^{-11}	7.48×10^{-3}	3.50	2.5×10^{-11}	[1]
$HSeO_3^-$	5×10^{-4}	2000	0.50	0	-1	7.5	25	760	2780	128	7.77×10^{-10}	0.28	0.28	1.51×10^{-10}	7.48×10^{-3}	3.50	5.1×10^{-11}	[2]
$HSeO_3^-$	1×10^{-3}	2000	0.50	0	-1	7.5	25	760	2780	128	7.77×10^{-10}	0.28	0.28	1.51×10^{-10}	7.48×10^{-3}	3.50	1.4×10^{-11}	[2]
$HSeO_3^-$	3×10^{-4}	2000	0.50	0	-1	7.5	25	760	2780	128	7.77×10^{-10}	0.28	0.28	1.51×10^{-10}	7.48×10^{-3}	3.50	9.5×10^{-12}	[2]
$HSeO_3^-$	3×10^{-4}	2000	0.50	0	-1	7.5	25	760	2780	128	7.77×10^{-10}	0.28	0.28	1.51×10^{-10}	7.48×10^{-3}	3.50	6.4×10^{-12}	[2]

ChinaRxiv 中X

Species	I (mol/L)	d (kg/m³)	m (–)	Kd (m³/kg)	z (–)	pH (–)	T (°C)	Aext (m²/g)	s (kg/m²)	MW (g/mol)	Dw (m²/s)	tot (–)	α (–)	r (Å)	λ (m²·S/mol)	nc (–)	De (m²/s)	Reference
$HCrO_4^-$	2×10^{-9}	2000	0.500	- 1	7.5	25	760	2780	117	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	2.03×10^{-10}	7.48×10^{3}	1.50	6.18×10^{-11}	[3]
$HCrO_4^-$	3×10^{-9}	2000	0.500	- 1	7.5	25	760	2780	117	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	2.03×10^{-10}	7.48×10^{3}	1.50	4.2×10^{-11}	[3]
$HCrO_4^-$	3×10^{-9}	2000	0.500	- 1	7.5	25	760	2780	117	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	2.03×10^{-10}	7.48×10^{3}	1.50	8×10^{-11}	[3]
$HCrO_4^-$	3×10^{-9}	2000	0.500	- 1	7.5	25	760	2780	117	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	2.03×10^{-10}	7.48×10^{3}	1.50	6.4×10^{-11}	[3]
$HCrO_4^-$	3×10^{-9}	2000	0.500	- 1	7.5	25	760	2780	117	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	2.03×10^{-10}	7.48×10^{3}	1.50	4.4×10^{-11}	[3]
$HCrO_4^-$	3×10^{-9}	2000	0.500	- 1	7.5	25	760	2780	117	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	2.03×10^{-10}	7.48×10^{3}	1.50	2.8×10^{-11}	[3]
$CoEDTA^{2-}$	5×10^{-10}	1600	0.750	- 2	7.5	25	562	2780	409	4.59×10^{-3}	9.42×10^{-1}	6.42×10^{-1}	1.49×10^{-10}	7.48×10^{3}	1.50	1.85×10^{-11}	[4]
$CoEDTA^{2-}$	5×10^{-10}	1600	0.750	- 2	7.5	25	562	2780	409	4.59×10^{-3}	9.42×10^{-1}	6.42×10^{-1}	1.49×10^{-10}	7.48×10^{3}	1.50	1.43×10^{-11}	[4]
$CoEDTA^{2-}$	5×10^{-10}	1600	0.750	- 2	7.5	25	562	2780	409	4.59×10^{-3}	9.42×10^{-1}	6.42×10^{-1}	1.49×10^{-10}	7.48×10^{3}	1.50	6.8×10^{-12}	[4]
$CoEDTA^{2-}$	5×10^{-10}	1600	0.750	- 2	7.5	25	562	2780	409	4.59×10^{-3}	9.42×10^{-1}	6.42×10^{-1}	1.49×10^{-10}	7.48×10^{3}	1.50	6.4×10^{-12}	[4]
$CoEDTA^{2-}$	5×10^{-10}	1600	0.750	- 2	7.5	25	562	2780	409	4.59×10^{-3}	9.42×10^{-1}	6.42×10^{-1}	1.49×10^{-10}	7.48×10^{3}	1.50	5.6×10^{-12}	[4]
$HSeO_3^-$	3.5×10^{-4}	2000	-0.500	- 1	7.5	25	760	2780	128	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	1.51×10^{-10}	7.48×10^{3}	1.50	4.4×10^{-11}	[5]
$HSeO_3^-$	3.26×10^{-4}	2000	-0.500	- 1	7.5	25	760	2780	128	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	1.51×10^{-10}	7.48×10^{3}	1.50	3×10^{-11}	[5]
$HSeO_3^-$	3.2×10^{-4}	2000	-0.500	- 1	7.5	25	760	2780	128	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	1.51×10^{-10}	7.48×10^{3}	1.50	2×10^{-11}	[5]
$HSeO_3^-$	3.2×10^{-4}	2000	-0.500	- 1	7.5	25	760	2780	128	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	1.51×10^{-10}	7.48×10^{3}	1.50	5.6×10^{-11}	[5]
$HSeO_3^-$	3.4×10^{-4}	2000	-0.500	- 1	7.5	25	760	2780	128	7.77×10^{-3}	9.28×10^{-1}	6.28×10^{-1}	1.51×10^{-10}	7.48×10^{3}	1.50	6.6×10^{-11}	[5]

References

[1] Y. Fukatsu, K. Yotsuji, T. Ohkubo, et al., Diffusion of tritiated water, $^{137}Cs^+$, $^{125}I^-$ in compacted Ca-montmorillonite: Experimental and modeling approaches. *Appl. Clay Sci.* **106176** (2021). https://doi.org/10.1016/j.clay.2021.106176

[2] T. Wu, Y. Hong, D. Shao, et al., Experimental and modeling study of the diffusion path of Ce(III)-EDTA in compacted bentonite. *Chem. Geol.* **636**,

121639 (2023). https://doi.org/10.1016/j.chemgeo.2023.121639

[3] T. Wu, Y. Yang, Z. Wang, et al., Anion diffusion in compacted clays by pore-scale simulation experiments. *Water Resour. Res.* e2019WR027037 (2020). https://doi.org/10.1029/2019wr027037

[4] I.C. Bourg, A.C. Bourg, G. Sposito, Modeling diffusion and adsorption in compacted bentonite: a critical review. *J. Contam. Hydrol.* **61**, 293–302 (2003). https://doi.org/10.1016/S0169-7722(02)00128-6

[5] P. Vanysek, Ionic conductivity and diffusion at infinite dilution. In: *CRC Handbook of Chemistry and Physics*, 83rd ed., pp. 76–78 (2000).

[6] Y. Ishii, T. Murakami, N. Ohtori, Molecular shape effects: Tracer diffusion and the Stokes-Einstein relation. *J. Mol. Liq.* **346**, 118235 (2022). https://doi.org/10.1016/j.molliq.2021.118235

[7] C. Tournassat, C. Appelo, Modelling approaches for anion-exclusion in compacted Na-bentonite. *Geochim. Cosmochim. Acta* (2011). https://doi.org/10.1016/j.gca.2011.04.001

[8] Z. Feng, J. Tian, X. Shi, et al., Analyzing porosity of compacted bentonite via through-diffusion method. *J. Radioanal. Nucl. Chem.* (2024). https://doi.org/10.1007/s10967-024-09368-y

[9] T. Wu, J. Tian, X. Shi, et al., Predicting anion diffusion in bentonite using hybrid machine learning model and correlation of physical quantities. *Sci. Total Environ.* **946**, 174363 (2024). https://doi.org/10.1016/j.scitotenv.2024.174363

[10] X. Shi, J. Tian, J. Shen, et al., Application of machine learning in predicting the apparent diffusion coefficient of Se(IV) in compacted bentonite. *J. Radioanal. Nucl. Chem.* (2024). https://doi.org/10.1007/s10967-024-09637-w

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.