Analysis of Chornobyl Groundwater Monitoring Data Using Unsupervised and Supervised Machine Learning Algorithms

D. Bugai, D. Hryhorenko

Institute of Geological Sciences, Kyiv, Ukraine

REMPLEX Symposium, 4-7 November, 2025 Pacific North-West National Laboratory, USA

1. Objectives, tasks and analysed dataset

Objectives

Test the potential of Machine Learning algorithms to predict radionuclide concentrations in groundwater of the Chornobyl Exclusion Zone using surface contamination levels, unsaturated zone thickness, sampling depth, and other relevant hydrogeological parameters as input features

Dataset

- Groundwater monitoring data from the Chornobyl Ecocenter radiation monitoring service
- 72 monitoring wells, with measurements collected between 1989 and 2022

Two-Phase Exercise

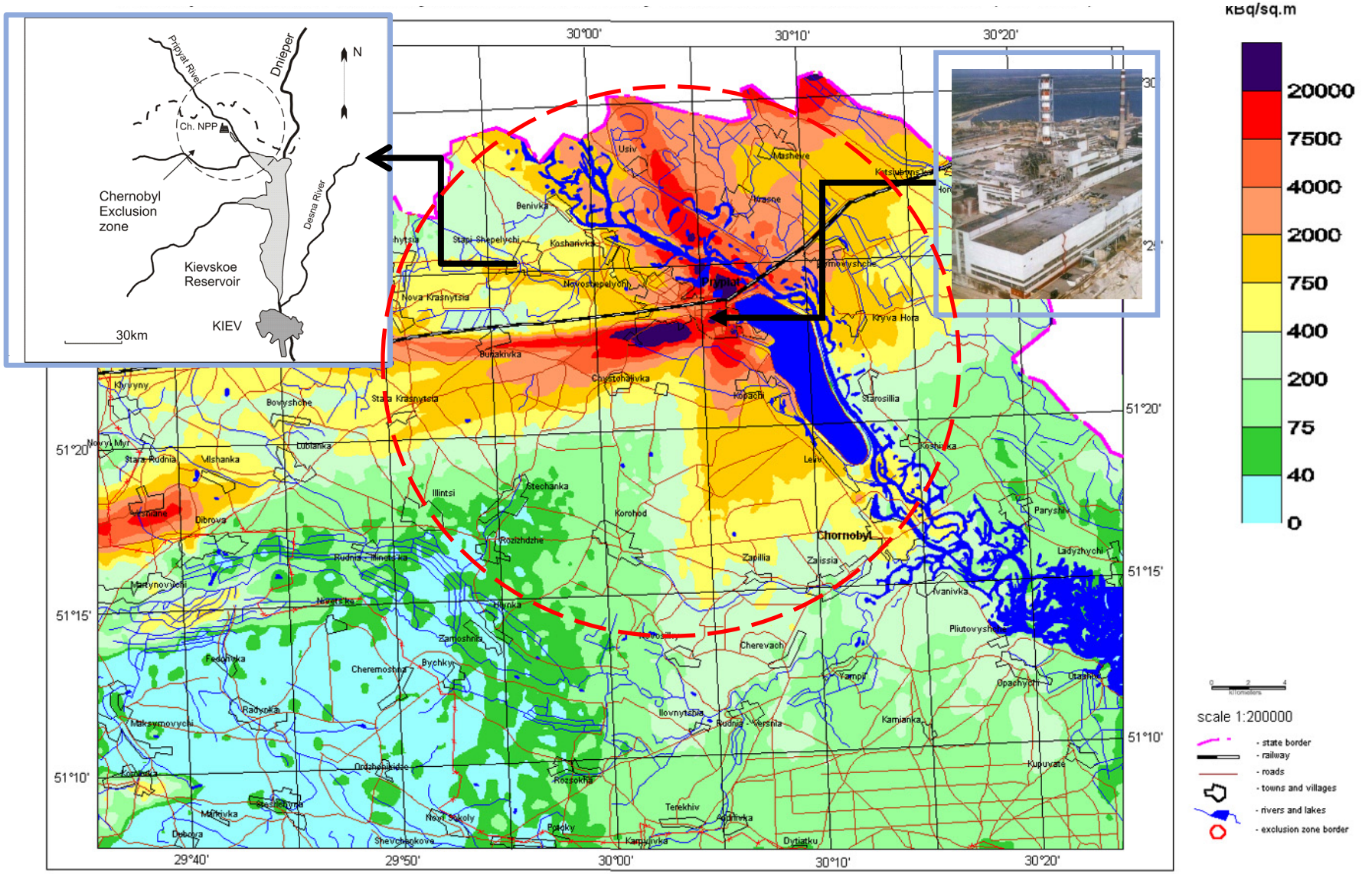
I. Unsupervised learning: explore interdependencies between radionuclide groundwater concentrations (^{137}Cs, ^{90}Sr), hydrogeological, and radiological parameters

II. Supervised learning: assess the potential to predict ^{90}Sr groundwater concentrations using an expanded set of features

2. Chornobyl Exclusion Zone–geography and radionuclide inventory

- Chornobyl Exclusion Zone (CEZ, 2600 km^2) was established shortly after the 1986 Chornobyl Accident;
- CEZ contains fallout radionuclides mostly in the fuel particle form;
- The most highly contaminated part of CEZ is the 10-km zone of ChNPP;
- The contaminated area is situated in the upper reaches of the Pripyat-Dnieper River basins, with the potential of off-site transport of radioactivity to downstream populations.

Map of the Chornobyl Exclusion zone and ^{137}Cs contamination (as of 1997)



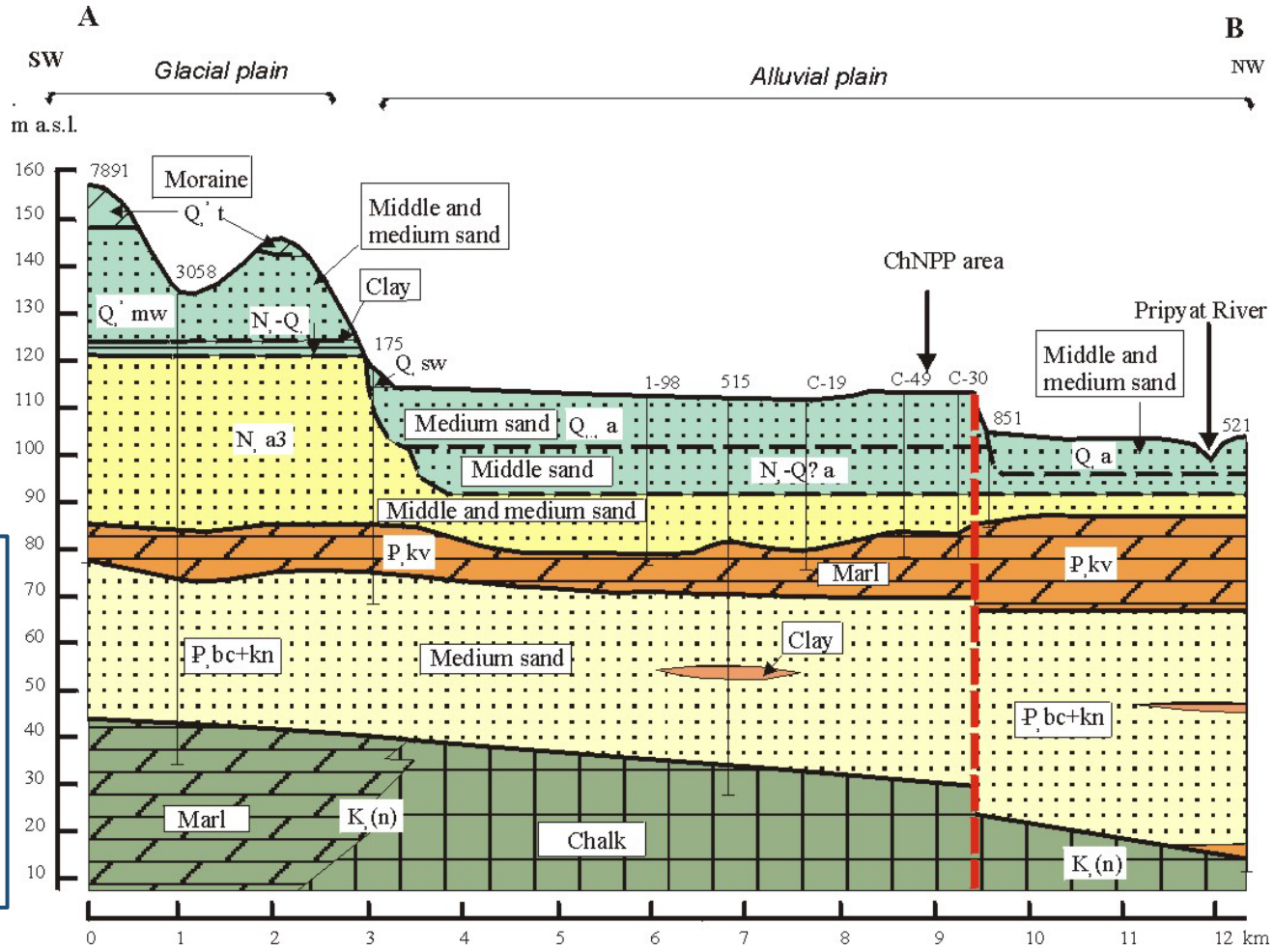
3. Chornobyl Exclusion Zone climate, geology, and hydrogeology

Environmental conditions

- Humid continental climate (rainfall ~600 mm/y);
- Relatively flat landscape;
- Shallow groundwater table (often 1-3 m or less);
- Sandy alluvial and fluvioglacial deposits have relatively high permeability and low CEC;
- Groundwater recharge rates 100–250 mm/y; flow velocities ~ 10 m/y.

- Hydrogeologic conditions generally **facilitate radionuclide migration into groundwater**.
- From the earliest days following the accident, the **transport of radionuclides via groundwater to the river network** - and ultimately to downstream populations in the Dnieper Basin- was recognized as **a major hazard**.

Geological cross-section of CEZ

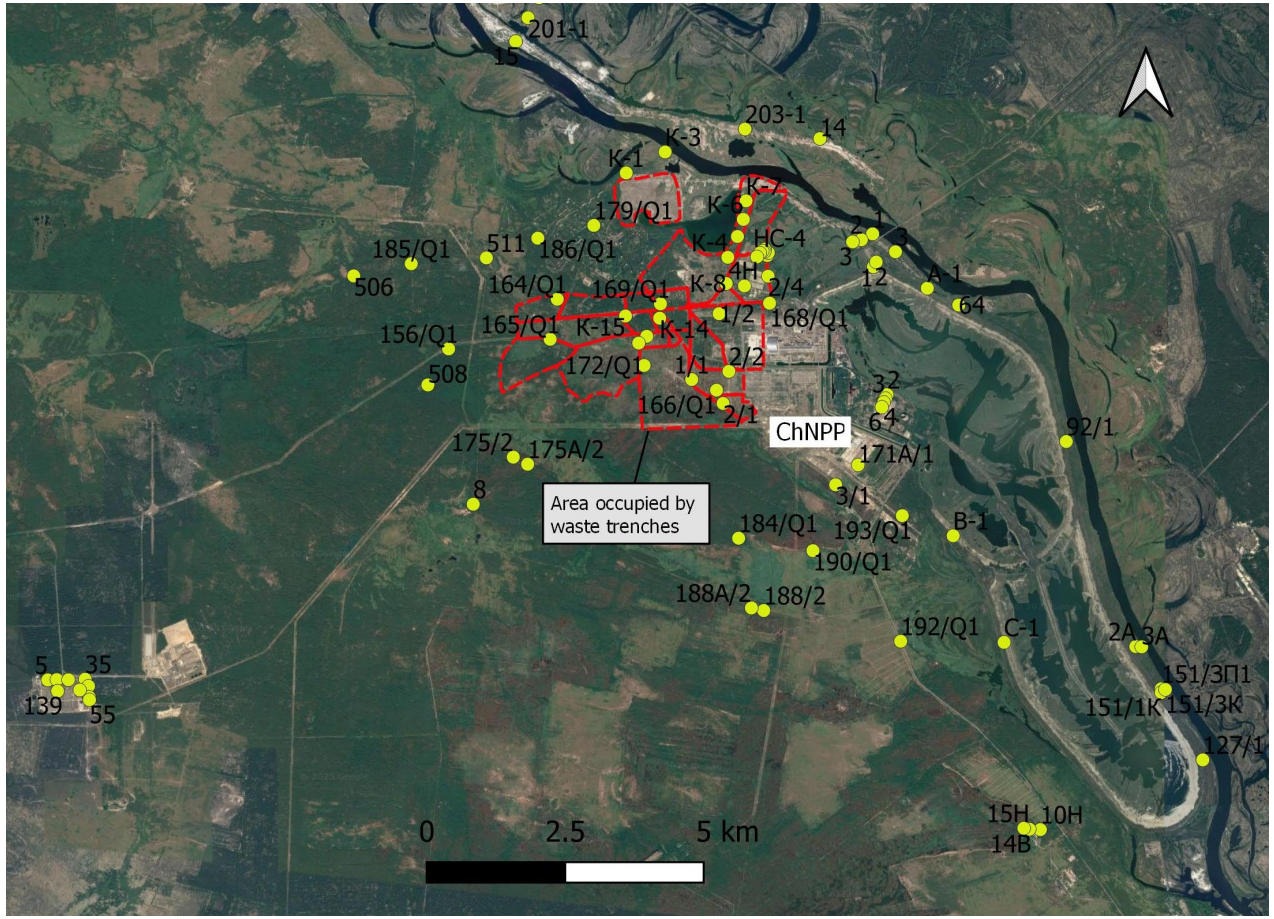


Adapted from [Matoshko, 1999]

4. Groundwater monitoring system in the Chornobyl Zone

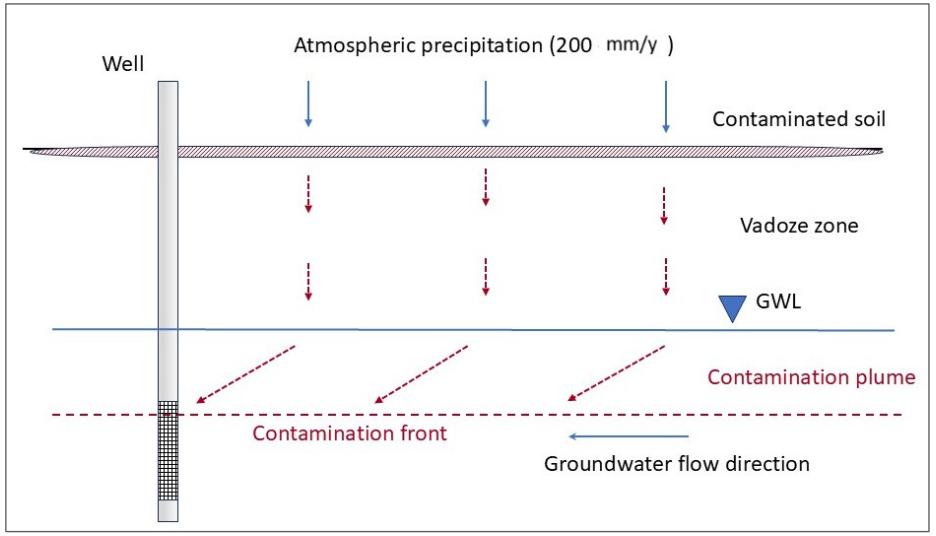
- Regional groundwater monitoring in CEZ is conducted by the State Special Enterprise “Ecocenter”;
- Monitoring observations have been carried out since 1990;
- Groundwater monitoring network comprises ~150 wells, mostly located in the 10-km zone of the ChNPP;
- Monitored parameters are: ^{90}Sr, ^{137}Cs, $^{239+240}Pu$, and ^{241}Am, groundwater levels;
- Analyses are carried out on unfiltered samples;
- Sampling frequency varies from monthly to quarterly and yearly.

Layout of monitoring wells of the Ecocenter in the 10-km zone of the ChNPP

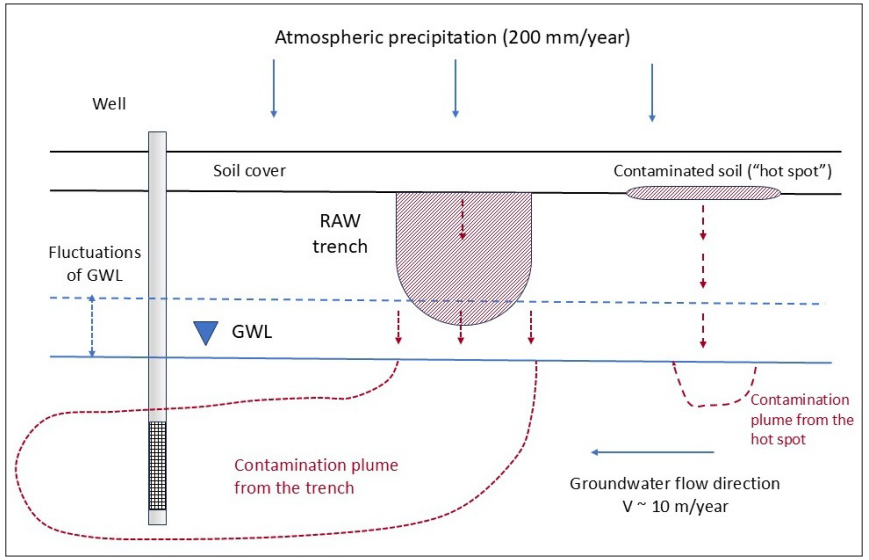


5. Sources of radioactivity to groundwater in Chornobyl Zone

Soils contaminated by radioactive fallout

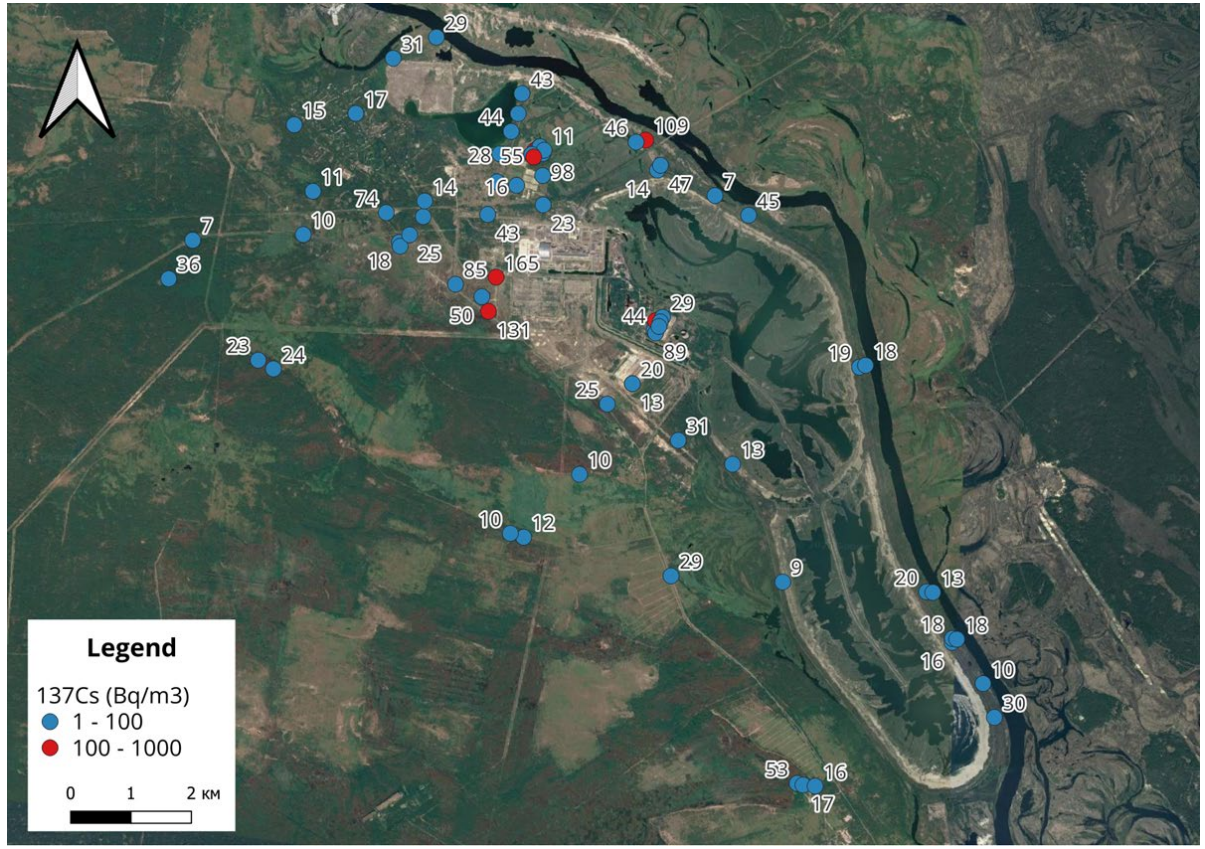


Waste trenches

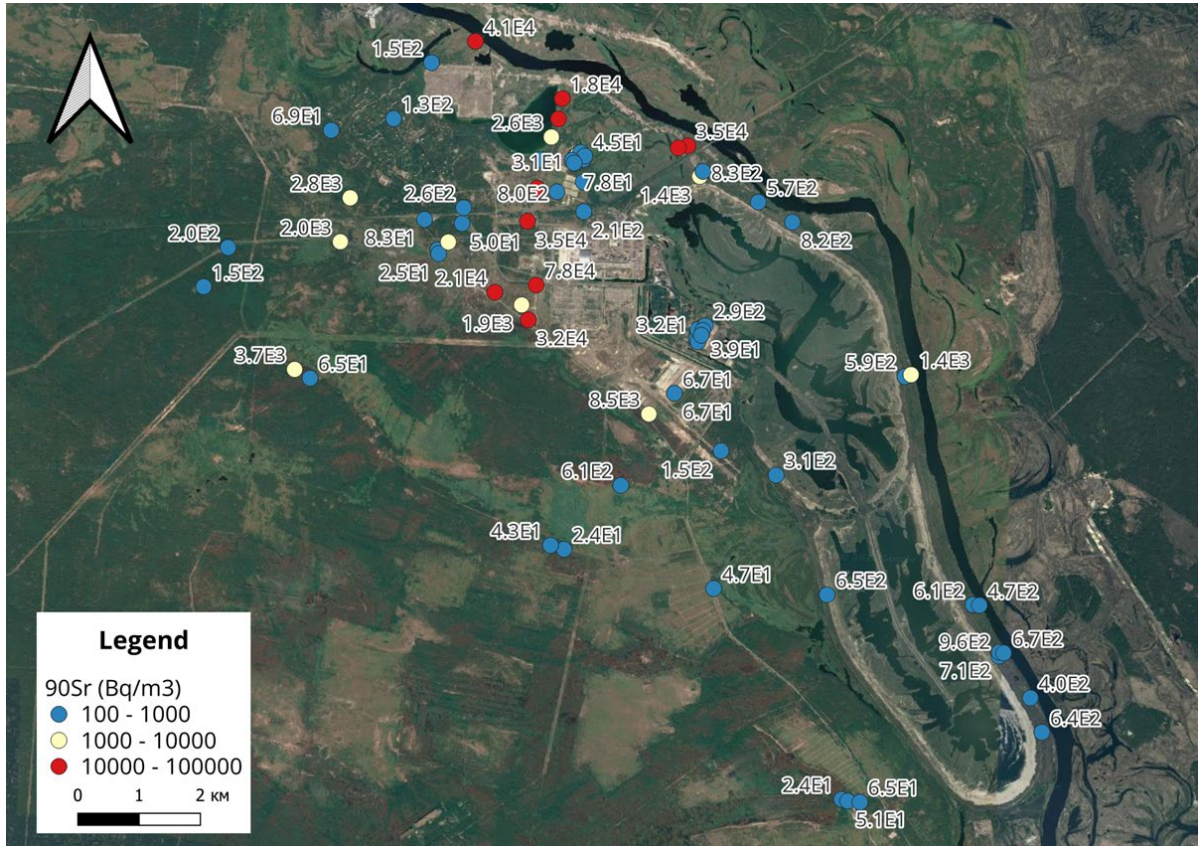


6. Examples of 2D maps of ^{137}Cs and ^{90}Sr in groundwater in 2024 (data of the SSE “Ecocenter”)

^{137}Cs



^{90}Sr

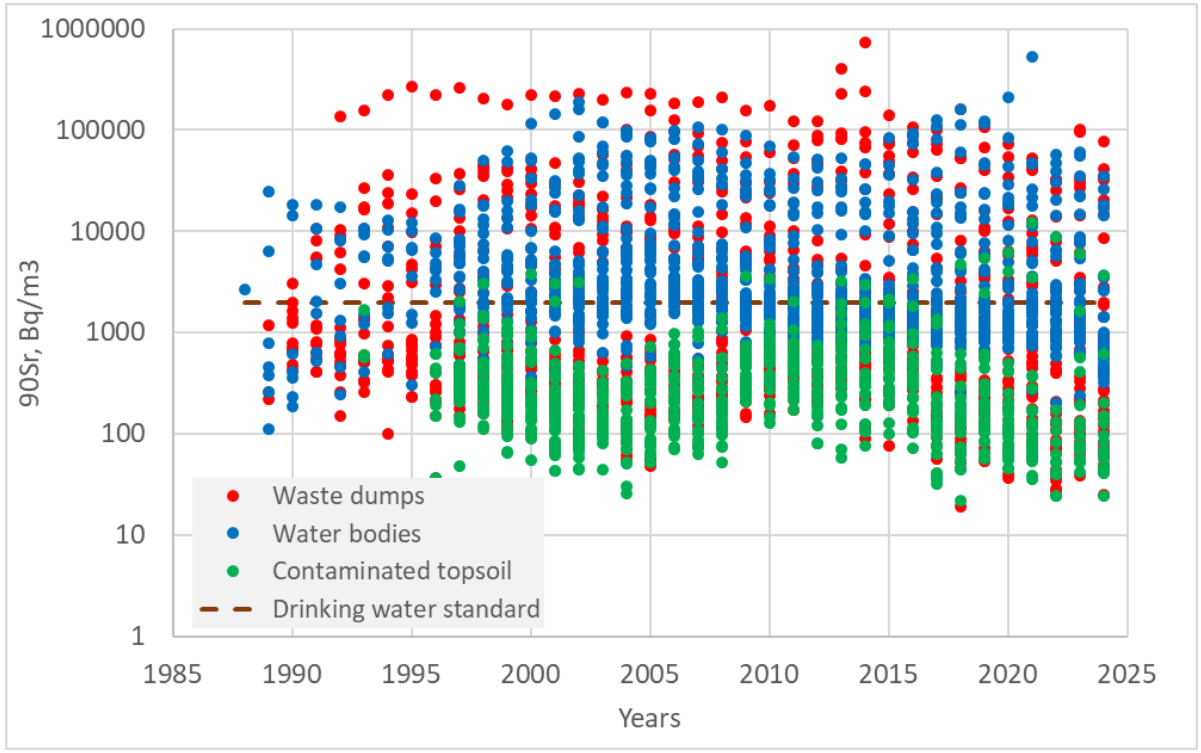


The ^{137}Cs concentrations in the unconfined aquifer vary in the range 10 – 100 Bq m^{-3}, which is below the Ukrainian drinking water standard (DWS) of 2000 Bq m^{-3}

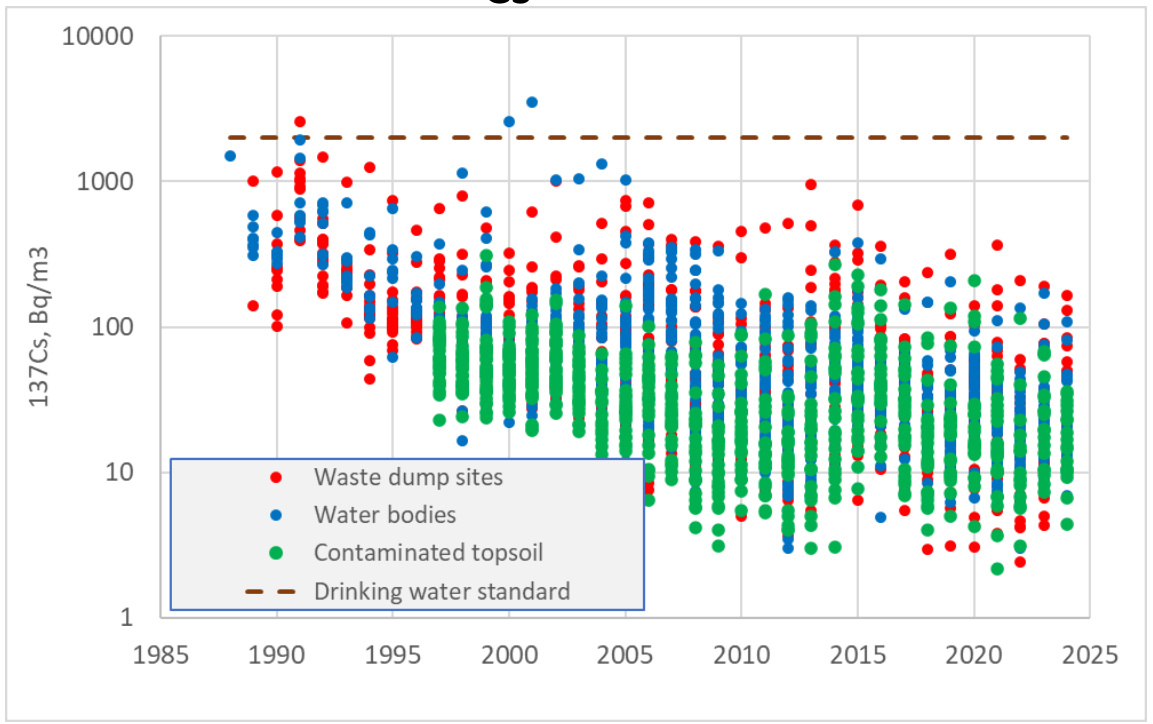
The ^{90}Sr concentrations in groundwater range from 10^2 to 10^5 Bq m^{-3}, and often exceed the Ukrainian DWS of 2000 Bq m^{-3}

7. Time series of radionuclide activity concentrations in groundwater in all monitoring wells

^{90}Sr



^{137}Cs



Time trends of ^{90}Sr are consistent with the conceptual model assuming the gradual dissolution of nuclear fuel particles and ^{90}Sr downward transport through the vadose zone to groundwater

The mechanisms determining dynamics of relatively low mobile ^{137}Cs in monitoring wells are less clear

8. Application of unsupervised learning using PyLEnM: input data – parameters and methods

A pilot 'unsupervised learning' exercise is based on the application of the PyLEnM software (Meray et al., 2022) to analyze factors potentially influencing radionuclide concentrations in groundwater at the ChEZ

The following parameters were analyzed

- ^{137}Cs and ^{90}Sr concentrations in groundwater
- Land surface contamination densities by ^{137}Cs and ^{90}Sr
- Distance from the source of release - damaged Chernobyl Unit 4
- Groundwater levels
- Depth of the well screen

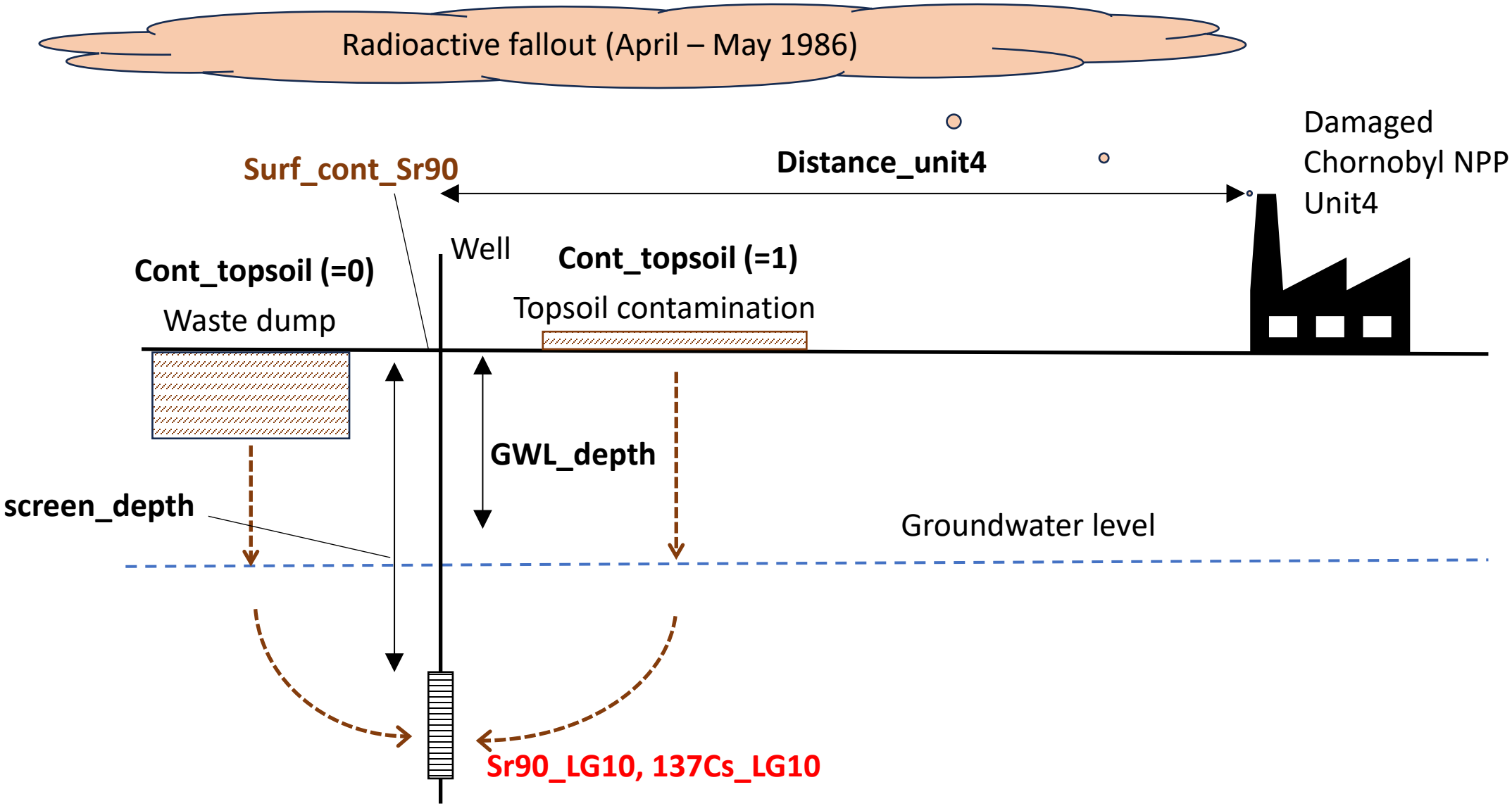
Methods

- The initial Excel database was restructured as a Python DataFrame (MySQL format).
- Radionuclide concentrations in groundwater were log-transformed to account for skewed distributions and to reduce the influence of potential outliers.
- Correlation analysis was then performed using yearly averaged concentrations of ^{137}Cs and ^{90}Sr, together with groundwater level measurements.

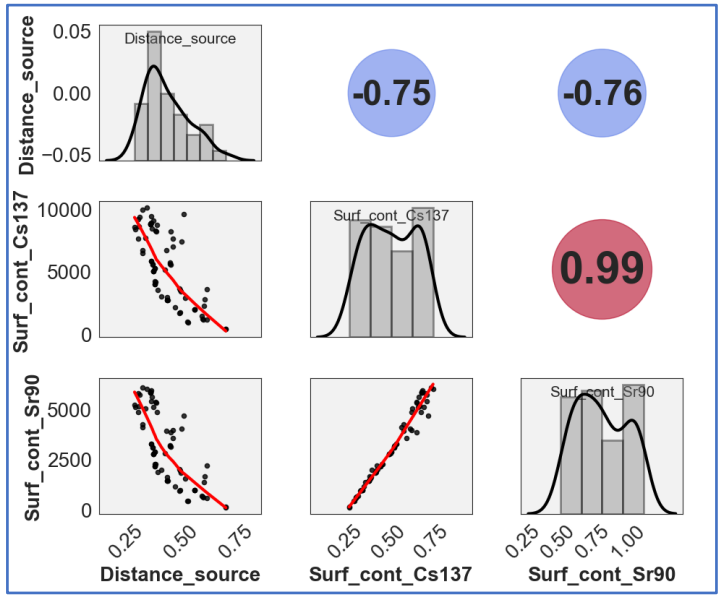
- Radionuclide land **surface concentrations at individual locations** were estimated based on soil sampling data set of Kashparov et al. (1997) using kriging interpolation

Reference: Meray A et al. (2022) A Machine Learning Framework for Long-Term Groundwater Contamination Monitoring Strategies. *Environ. Sci. Technol.*, 56, 5973−5983 https://doi.org/10.1021/acs.est.1c07440

9. Conceptual scheme

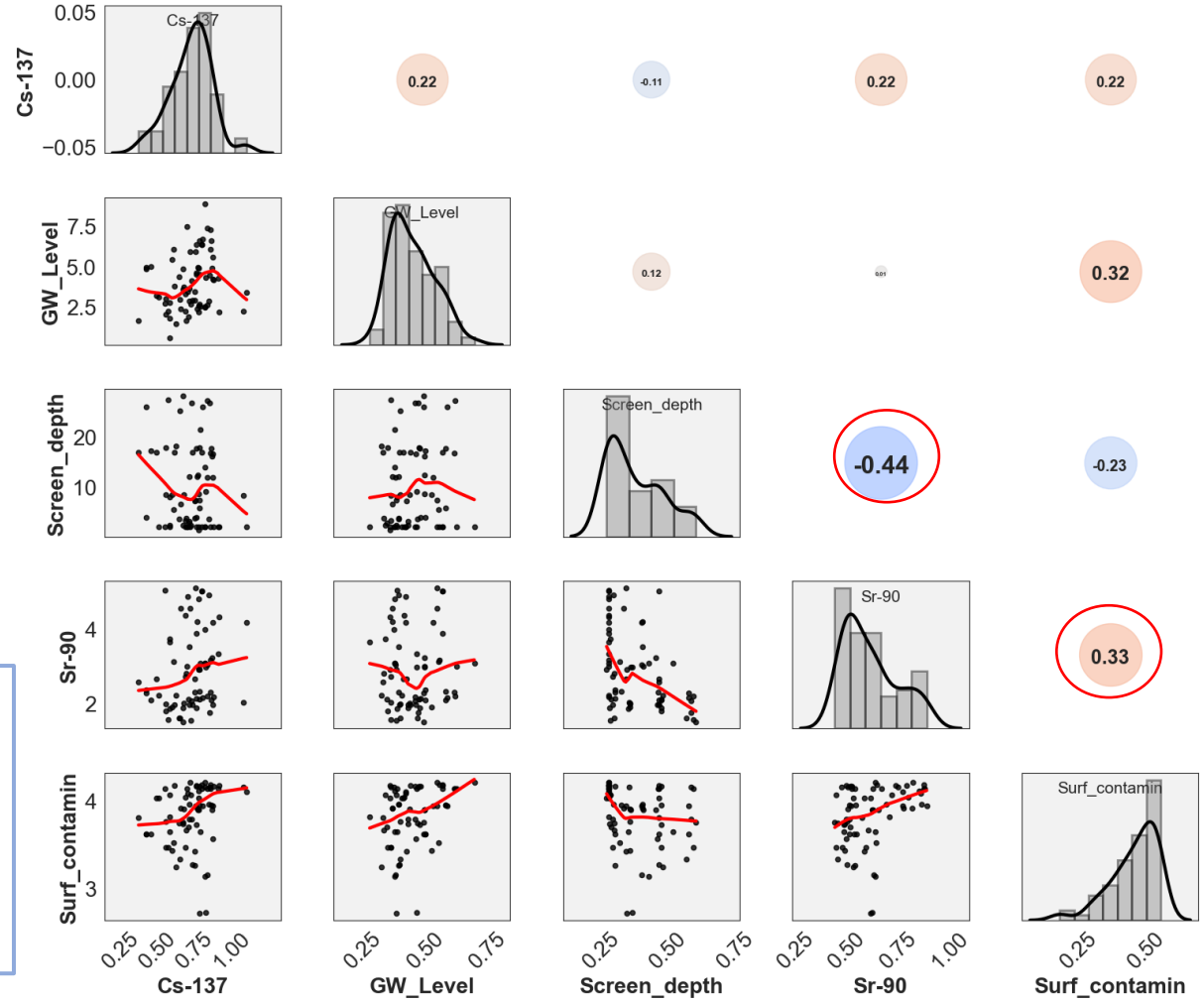


10. Correlation analysis of monitoring data using PyLEnM



- Statistical analysis shows good correlation between surface contamination by ^{137}Cs and ^{90}Sr, and inverse correlation to distance to source
- Subsequent correlation analyses used only 1 parameter 'Surface contamination'

Correlation analysis of monitoring data and sampling locations characteristics for 2017

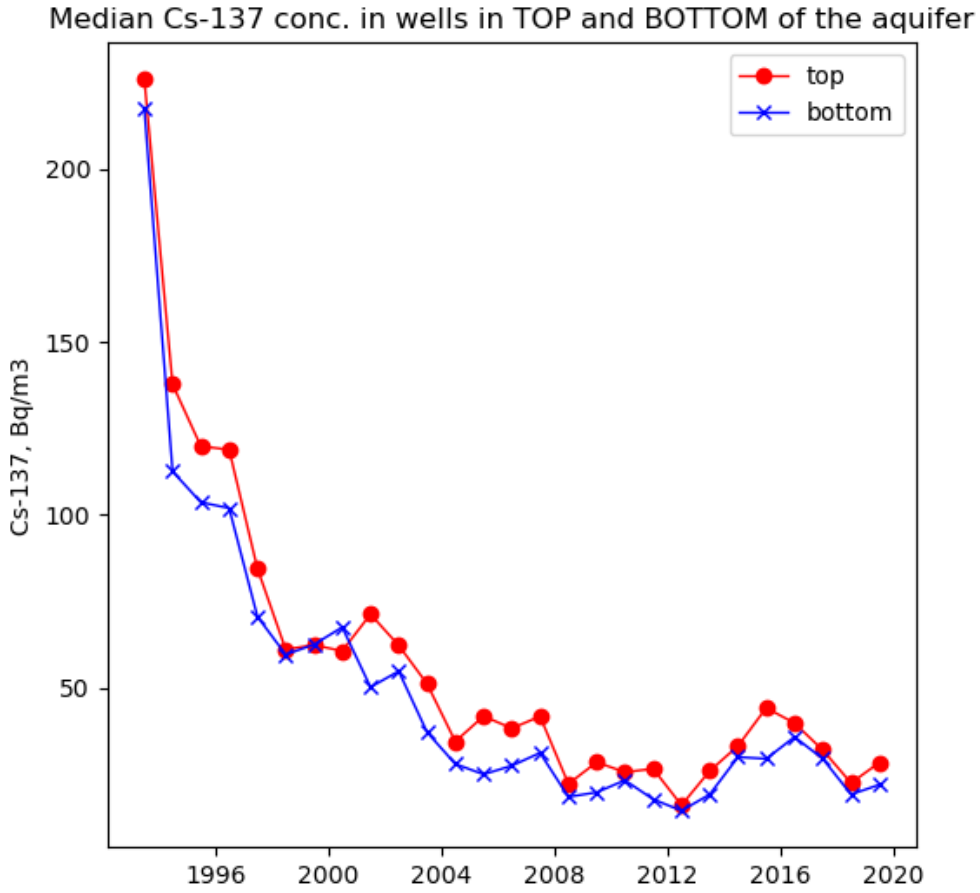


Date: 2017-07-01

Wells: 72
Samples used: 360

11. Summary of the PyLEnM correlation analysis

- No strong positive correlation was found between radionuclide groundwater concentrations and land surface contamination.
- **The limited set of characteristics** used in the analysis **is likely insufficient to explain the variability** of observed radionuclide concentrations in groundwater.
- For ^{90}Sr, the observed negative correlation with sampling depth in groundwater is consistent with a **conceptual model assuming migration from a near-surface source term**.
- In contrast, the correlation results **for ^{137}Cs are not consistent** with a conceptual model of advective–dispersive transport from a near-surface source.
- The lack of concentration differentiation between the upper and lower parts of the aquifer, combined with the decreasing time trend of ^{137}Cs in groundwater, may indicate **contamination of wells during drilling**.



12. Supervised Machine Learning Analyses of ^{90}Sr Dataset

Testing the possibility of **predicting ^{90}Sr concentrations** in groundwater using Random Forest algorithm based on the **extended set** of hydrogeological and radiological features of monitoring locations

Analyzed parameters

Target parameter:

- Sr-90 concentration in groundwater

Features

- Soil surface contamination by Sr-90
- Distance from the source of release (damaged Unit 4)
- Groundwater level depths
- Depth of the well screen
- Length of well screen

Binary data

- Type of source (contaminated topsoil/waste dump)
- Top/bottom part of the aquifer
- Long/short screen

Methodology

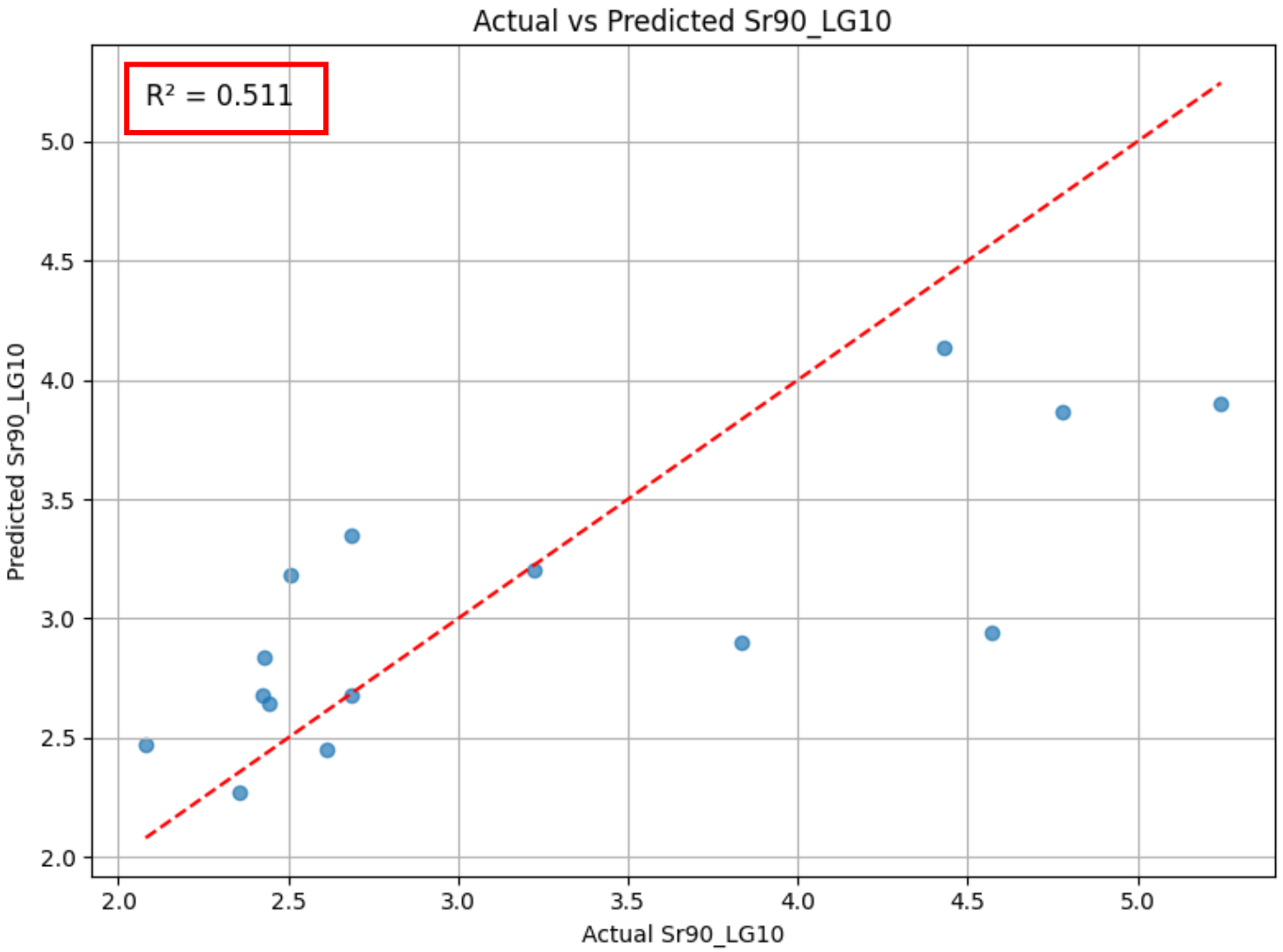
- Analyses used median values of hydrogeological parameters in individual wells (^{90}Sr, groundwater level) for 1989-2020
- The **scikit-learn Python library** was used for **Random Forest** algorithm (in regressor and classification modes)
- The data set was split into 80% training vs 20% testing sub-sets

13. Target parameter and attributes (for RF regressor mode)

Sr90_LG10	GWL_DEPTH	SCREEN_DEPTH	SURF_CONT_SR90	DISTANCE_UNIT4	AQUIFER_TOP	CONT_TOPSOIL	LONG_SCREEN
5,25	2,13	2,1	5044,5	1,66	1	0	1
4,75	5,26	2,1	5805,4	1,05	1	0	1
4,44	2,78	2,1	4264,2	1,46	1	0	1
2,89	2,73	17	4103,8	1,43	0	0	1
4,43	2,8	2,1	4862,3	1,02	1	0	1
3,83	0,52	2,1	2290,9	2,57	1	0	1
2,62	0,73	17	2262,9	2,60	0	0	1
2,30	2,51	3,25	2158,6	2,57	1	0	0
2,22	2,49	2,45	1890,3	2,79	1	0	0
2,44	3,04	26,7	2158,6	2,57	0	0	0
2,51	4,11	4	5830,0	2,12	1	0	0
2,44	3,99	17,9	5821,4	2,13	0	0	0
2,43	3,94	26,8	5615,2	2,29	0	0	0
2,96	1,73	1,5	3915,3	3,62	1	0	0
3,73	1,38	2,5	3944,9	4,19	1	0	0
4,24	1,77	2,1	5927,4	2,23	1	0	1
2,43	3,5	2,1	5845,7	2,14	1	0	1
2,37	3,515	17	5877,6	2,12	0	0	1
2,68	3,85	2,1	5329,6	2,65	0	0	1
4,88	5,26	17	5370,6	2,62	1	0	1
2,79	8,58	2,1	6019,7	1,82	1	0	1
3,81	5,91	2,1	5396,1	2,18	1	0	1
3,46	1,81	2,1	5112,9	2,44	1	0	1

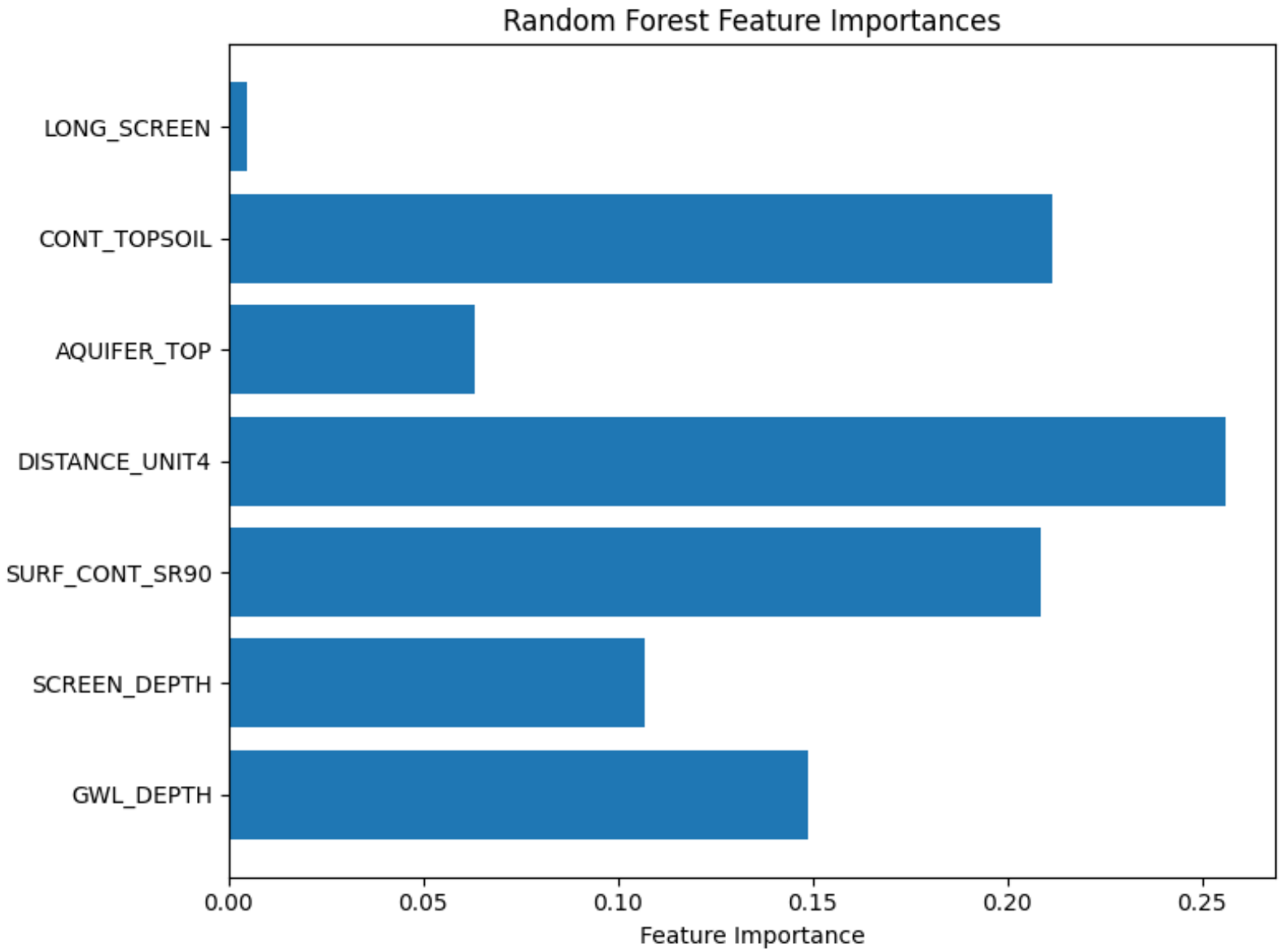
Analysis used data from 73 monitoring wells

14. Performance of Random Forest in regressor mode



15. Importance of various features

Type of source – contaminated topsoil/ waste dump →



16. Target parameter and attributes (for RF classifier mode)

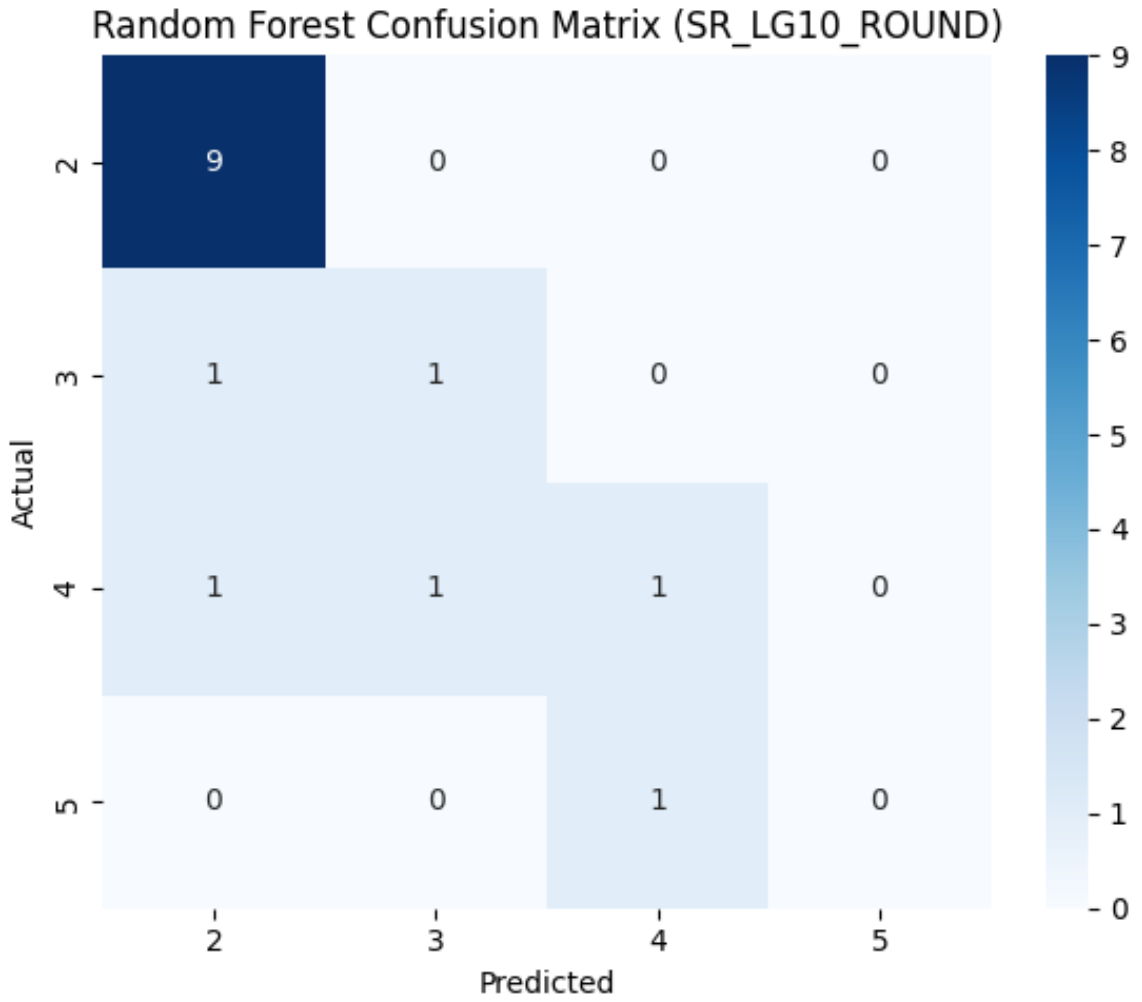
Log10 of Sr-90 concentration in groundwater was rounded to the nearest integer and treated as a **categorical variable** (values 2, 3, … 6)

SR90_LG10_ROUND	GWL_DEPTH	SCREEN_DEPTH	SURF_CONT_SR90	DISTANCE_UNIT4	AQUIFER_TOP	CONT_TOPSOIL	LONG_SCREEN
6	2,13	2,1	5044,5	1,66	1	0	1
5	5,26	2,1	5805,4	1,05	1	0	1
5	2,78	2,1	4264,2	1,46	1	0	1
3	2,73	17	4103,8	1,43	0	0	1
5	2,8	2,1	4862,3	1,02	1	0	1
4	0,52	2,1	2290,9	2,57	1	0	1
3	0,73	17	2262,9	2,60	0	0	1
3	2,51	3,25	2158,6	2,57	1	0	0
3	2,49	2,45	1890,3	2,79	1	0	0
3	3,04	26,7	2158,6	2,57	0	0	0
3	4,11	4	5830,0	2,12	1	0	0
3	3,99	17,9	5821,4	2,13	0	0	0
3	3,94	26,8	5615,2	2,29	0	0	0
3	1,73	1,5	3915,3	3,62	1	0	0
4	1,38	2,5	3944,9	4,19	1	0	0
5	1,77	2,1	5927,4	2,23	1	0	1
3	3,5	2,1	5845,7	2,14	1	0	1
3	3,515	17	5877,6	2,12	0	0	1
3	3,85	2,1	5329,6	2,65	0	0	1
5	5,26	17	5370,6	2,62	1	0	1
3	8,58	2,1	6019,7	1,82	1	0	1
4	5,91	2,1	5396,1	2,18	1	0	1
4	1,81	2,1	5112,9	2,44	1	0	1
5	2,41	2,1	4854,9	2,72	1	0	1

Analysis used data from 73 monitoring wells

17. Accuracy metrics of RF model in classifier mode

Metric	Value	Interpretation
Accuracy	**73%**	Low to moderate accuracy
Balanced Accuracy	**0,46**	Poor balanced accuracy
Cohen's Kappa	**0,48**	Moderate agreement
Matthews Corr Coeff	**0,50**	Moderate classification quality
R^2 (ordinal)	**0,53**	Moderately close predictions



Conclusions

- The **unsupervised machine learning** analysis of the Chernobyl groundwater monitoring dataset using PyLEnM provided valuable insights for interpreting the data and for **verifying relevant conceptual models** of radionuclide migration.
- The machine learning analysis of the **^{137}Cs** groundwater dataset suggests that observed behaviour may be attributable to **well contamination during drilling**.
- **For ^{90}Sr** groundwater concentrations, supervised machine learning (Random Forest classifier) achieved **moderate predictive accuracy** when accounting for a combination of features describing contaminant source intensity and type, as well as hydrogeological parameters.
- The developed model can be applied **for order-of-magnitude predictions** of groundwater contamination levels in areas of the Chernobyl zone not currently covered by the groundwater monitoring network.