

Probabilistic reconstruction via machine-learning of the Po watershed aquifer system (Italy)

Andrea Manzoni ⁽¹⁾, Giovanni Michele Porta^{*(1)}, Laura Guadagnini⁽¹⁾, Alberto Guadagnini⁽¹⁾,
Monica Riva⁽¹⁾

⁽¹⁾Dipartimento di Ingegneria Civile e Ambientale (DICA), Politecnico di Milano, Piazza
Leonardo da Vinci 32, Milano, 20133, Italy

* Corresponding author. Tel. +390223996256, e-mail address:

giovanni.porta@polimi.it

ESM SECTION S1

The datasets provided by the national research institute (ISPRA 2021) and by the Piemonte region (Arpa Piemonte, 2022) include a qualitative description of the geomaterials detected along each available borehole. A semi-supervised categorization strategy is employed to analyze the data embedded in these sources. The initial step of the analysis relies on grouping information that contain references to soil matrix texture (e.g., elements identified as ‘*cobbles in clay matrix*’ are assigned to clay), with the exception of matrix descriptions preceded by adjectives that suggest a reduced importance of the matrix (e.g., ‘*scarce clay matrix*’). Descriptions containing keywords that indicate the most abundant soil texture (e.g., *prevalent*, *abundant*, *rich*, or *predominant*) are then selected. The remaining well layers are finally grouped based on the first geomaterial mentioned in the description. Additional details on the employed keywords are included in the software available through the dedicated GitHub repository(<https://github.com/manzoniandrea/Geo-reconstruction - ANN/tree/main/dataCategorization>).

Furthermore, it should be noted that the lithostratigraphic layers available along boreholes in the Lombardy (Regione Lombardia 2022) and Emilia-Romagna (Regione Emilia-Romagna 2022) regions were already associated with at least one soil texture category and, sometimes, also with the matrix soil texture. This information is used to group the training points into the six macro-categories illustrated in section ‘*Study area and data pre-processing*’ in the main article. Firstly, all of the categories documented in the datasets (i.e., from coarse to very fine, for gravel, sand silt, and clay) are combined into the corresponding macro-categories (e.g., ‘*fine sand*’ is set to the *sand* macro-category) by prioritizing the soil texture of the matrix, when available. Subsequently, all available rock categories that typically associated with an average hydraulic conductivity larger than 10^{-9} cm/s (e.g., *sandstone*, *limestone*, *gypsum*) and all fractured rocks are grouped into the permeable rocks macro-category. The remaining rock categories (e.g., those termed as, *granite*) are assigned to the impermeable rock macro-category.

ESM SECTION S2

Table S1 Number of parameters, N_{ANN} , (weight and bias terms) of each of the tested ANN architectures. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	31	91	151	301	601	1501
2	61	331	801	2851	10701	64251
3	91	571	1451	5401	20801	127001
4	121	811	2101	7951	30901	189751
5	151	1051	2751	10501	41001	252501
6	181	1291	3401	13051	51101	315251
7	211	1531	4051	15601	61201	378001
8	241	1771	4701	18151	71301	440751

Table S2 Accuracy of each of the ANN architectures obtained with the *k*-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.534	0.565	0.570	0.581	0.586	0.589
2	0.547	0.580	0.586	0.595	0.599	0.596
3	0.548	0.587	0.592	0.598	0.596	0.583
4	0.559	0.590	0.592	0.595	0.591	0.583
5	0.558	0.589	0.595	0.597	0.592	0.589
6	0.564	0.591	0.596	0.597	0.592	0.595
7	0.567	0.595	0.595	0.595	0.595	0.593
8	0.565	0.588	0.596	0.595	0.595	0.588

Table S3 Mean cross-entropy loss function of each of the ANN architectures at the end of the training phase of the cross-validation step (the mean is evaluated over all of the k -iterations). The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	1.047	0.988	0.965	0.937	0.917	0.909
2	1.020	0.941	0.910	0.862	0.809	0.752
3	1.006	0.920	0.883	0.827	0.767	0.672
4	0.993	0.908	0.867	0.813	0.745	0.672
5	0.990	0.899	0.856	0.801	0.740	0.706
6	0.980	0.889	0.848	0.796	0.751	0.836
7	0.976	0.884	0.844	0.793	0.770	0.878
8	0.972	0.880	0.841	0.797	0.791	0.904

Table S4 Mean CPU time [s] required by the training phase during the cross-validation step of each ANN architecture. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	504	968	1628	3351	24273	47306
2	714	2927	19605	38383	83535	273278
3	1416	4209	25233	50718	101696	247476
4	2767	4107	27739	47857	93321	247476
5	2510	5271	31162	58097	79759	249441
6	3940	6411	33818	46203	63202	39089
7	3564	6872	36378	40423	46010	26370
8	5530	6851	32825	42526	31993	24288

Table S5 Accuracy of each of the ANN architectures in predicting the gravel macro-category resulting from the *k*-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.451	0.501	0.502	0.516	0.533	0.534
2	0.469	0.518	0.526	0.550	0.566	0.568
3	0.477	0.531	0.550	0.566	0.569	0.554
4	0.494	0.545	0.553	0.564	0.557	0.554
5	0.494	0.544	0.553	0.556	0.559	0.561
6	0.510	0.549	0.556	0.557	0.557	0.556
7	0.513	0.553	0.561	0.565	0.561	0.546
8	0.497	0.540	0.556	0.565	0.560	0.537

Table S6 Accuracy of each of the ANN architectures in predicting the sand macro-category resulting from the k-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.511	0.547	0.556	0.565	0.577	0.577
2	0.538	0.568	0.577	0.588	0.582	0.587
3	0.526	0.574	0.576	0.581	0.590	0.569
4	0.547	0.583	0.585	0.584	0.586	0.569
5	0.541	0.580	0.586	0.596	0.581	0.574
6	0.546	0.575	0.587	0.588	0.586	0.589
7	0.551	0.580	0.595	0.590	0.595	0.592
8	0.555	0.583	0.595	0.591	0.586	0.577

Table S7 Accuracy of each of the ANN architectures in predicting the silt macro-category resulting from the k-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.133	0.313	0.302	0.309	0.301	0.285
2	0.196	0.316	0.312	0.281	0.263	0.219
3	0.164	0.280	0.296	0.278	0.240	0.204
4	0.133	0.328	0.285	0.268	0.233	0.204
5	0.000	0.274	0.263	0.248	0.222	0.217
6	0.000	0.310	0.252	0.242	0.234	0.251
7	0.000	0.290	0.261	0.249	0.249	0.293
8	0.206	0.280	0.264	0.250	0.258	0.261

Table S8 Accuracy of each of the ANN architectures in predicting the clay macro-category resulting from the *k*-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.600	0.619	0.634	0.642	0.637	0.643
2	0.611	0.637	0.638	0.642	0.648	0.641
3	0.613	0.642	0.638	0.641	0.638	0.634
4	0.612	0.634	0.630	0.638	0.636	0.634
5	0.618	0.631	0.640	0.645	0.645	0.641
6	0.618	0.639	0.645	0.647	0.640	0.641
7	0.628	0.641	0.632	0.636	0.634	0.637
8	0.628	0.634	0.637	0.634	0.642	0.640

Table S9 Accuracy of each of the ANN architectures in predicting the permeable rock macro-category resulting from the *k*-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.401	0.479	0.467	0.498	0.479	0.479
2	0.391	0.458	0.498	0.464	0.452	0.473
3	0.407	0.464	0.480	0.481	0.463	0.463
4	0.415	0.466	0.484	0.457	0.460	0.463
5	0.414	0.490	0.482	0.463	0.465	0.471
6	0.392	0.455	0.463	0.459	0.457	0.495
7	0.368	0.480	0.479	0.466	0.486	0.480
8	0.412	0.464	0.462	0.469	0.450	0.478

Table S10 Accuracy of each of the ANN architectures in predicting the impermeable rock macro-category resulting from the *k*-fold cross-validation method. The value associated with the selected ANN architecture is highlighted in bold.

N° of hidden layers	N° of hidden nodes per hidden layer					
	5	15	25	50	100	250
1	0.401	0.479	0.467	0.498	0.479	0.479
2	0.391	0.458	0.498	0.464	0.452	0.473
3	0.407	0.464	0.480	0.481	0.463	0.463
4	0.415	0.466	0.484	0.457	0.460	0.463
5	0.414	0.490	0.482	0.463	0.465	0.471
6	0.392	0.455	0.463	0.459	0.457	0.495
7	0.368	0.480	0.479	0.466	0.486	0.480
8	0.412	0.464	0.462	0.469	0.450	0.478

ESM References

Arpa Piemonte (2022) Geoportale Arpa Piemonte (Geoportal of Arpa Piemonte).
<https://geoportale.arpa.piemonte.it/app/public/>. Accessed on October 20th, 2022.

ISPRA (2021) Dati geognostici e geofisici (Geognostic and geophysical data).
<https://www.isprambiente.gov.it/it/banche-dati/banche-dati-folder/suolo-e-territorio/dati-geognostici-e-geofisici> . Accessed on October 20th, 2022.

Regione Lombardia (2022) Banca dati geologica del sottosuolo (Geological data base).
https://www.geoportale.regione.lombardia.it/it/metadati?p_p_id=detailSheetMetadata_WAR_gptmetadataportlet&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&_detailSheetMetadata_WAR_gptmetadataportlet_uuid=%7BD9AF98B21-3257-4D23-9D53-5AECC966D872%7D Accessed on October 20th, 2022.

Regione Emilia-Romagna (2022) Portale minERva. https://datacatalog.regione.emilia-romagna.it/catalogCTA/dataset?groups=geologia&tags=stratigrafia&publisher_name=Region e+Emilia-Romagna Accessed on October 20th, 2022.