Semi-Supervised Learning for Grain Size Distribution Interpolation

Konstantin Kobs, Christian Schäfer, Michael Steininger, Anna Krause, Roland Baumhauer, Heiko Paeth, and Andreas Hotho

University of Würzburg, Würzburg, Germany
{kobs,steininger,anna.krause,hotho}@informatik.uni-wuerzburg.de
{christian.d.schaefer,baumhauer,heiko.paeth}@uni-wuerzburg.de

Abstract. High-resolution grain size distribution maps for geographical regions are used to model soil-hydrological processes that can be used in climate models. However, measurements are expensive or impossible, which is why interpolation methods are used to fill the gaps between known samples. Common interpolation methods can handle such tasks with few data points since they make strong modeling assumptions regarding soil properties and environmental factors. Neural networks potentially achieve better results as they do not rely on these assumptions and approximate non-linear relationships from data. However, their performance is often severely limited for tasks like grain size distribution interpolation due to their requirement for many training examples. Semi-supervised learning may improve their performance on this task by taking widely available unlabeled auxiliary data (e.g. altitude) into account.
We propose a novel semi-supervised training strategy for spatial interpolation tasks that pre-trains a neural network on weak labels obtained by methods with stronger assumptions and then fine-tunes the network on the small labeled dataset. In our research area, our proposed strategy improves the performance of a supervised neural network and outperforms other commonly used interpolation methods.

Keywords: spatial interpolation · semi-supervised learning · neural networks.

1 Introduction

The composition of different grain sizes in the soil affects many hydrological processes such as groundwater recharge, infiltration rates or surface flow. For example, soils with dominating clay fractions (grain size $\leq 0.002\,\text{mm}$) retain water better than sandy soils ($0.063\,\text{mm} <$ grain size $\leq 2.000\,\text{mm}$). Given accurate grain size distribution maps, it is possible to estimate hydrological parameters for environmental modelling purposes, e.g. regional climate models. Since sampling is expensive or even impossible due to inaccessible terrain, spatial interpolation methods are used to estimate grain size distributions for unknown locations.

A model for grain size distribution interpolation has the following requirements: (1) The model input is a location with (potentially) additional auxiliary

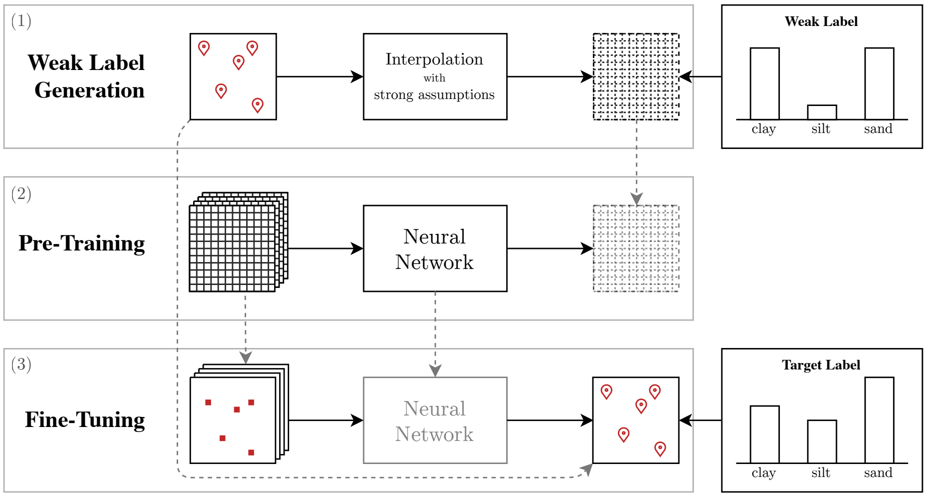


Fig. 1: In our proposed semi-supervised training method, (1) a spatial interpolation method with strong assumptions is trained on the labeled dataset. (2) The neural network is pre-trained on weak labels obtained by applying the interpolation method to the unlabeled data. The network gets locations and auxiliary data as inputs. (3) It is then fine-tuned on the labeled dataset.

data (e.g. altitude). (2) The model outputs distributions across the grain size classes (clay, silt, sand) for each unknown location. (3) The model works with few labeled data points, since soil samples are rare.

Distance based interpolation methods such as k Nearest Neighbors or Inverse Distance Weighting can output distributions and are applicable to small labeled datasets due to their strong assumptions. However, they do not take auxiliary data into account which can benefit performance [11,17]. Neural networks can learn non-linear relationships from data, are able to incorporate additional auxiliary inputs, and are able to output distributions across grain size classes. However, they usually need many labeled training data points [15]. The idea of semi-supervised learning utilizes large unlabeled datasets to support network training [8]. In recent years, most methods for semi-supervised learning were designed for image classification, which are not applicable to our setting.

Therefore, in this paper, we bring semi-supervised learning specifically to the task of grain size distribution interpolation for spatial inputs. We propose a training strategy that makes use of weak labels produced by an interpolation method with stronger modeling assumptions. Figure 1 gives a schematic overview of our proposed three-step process. In our experiments for the region of Lower Franconia, we show that our approach improves the performance of a supervised neural network and outperforms other common interpolation methods. Furthermore, we analyze the effects of the proposed training strategy on model performance.

Our contributions are: (1) We describe a semi-supervised training strategy for neural networks in the spatial domain to interpolate grain size distributions. (2) We compare our strategy to supervised training and common interpolation approaches and show that it outperforms them in our research area. (3) We analyze the resulting model to understand what factors are important for its performance.

2 Related Work

There are various spatial interpolation techniques with different properties used in environmental sciences, e.g. k Nearest Neighbors, Inverse Distance Weighting, or Kriging [16]. Neural networks have been successfully applied in such tasks since they allow auxiliary data as input features and can model non-linear relationships [5,20,23]. However, to obtain robust performance, they need many labeled data points not available in most spatial interpolation tasks [15]. Semi-supervised training promotes the use of large unlabeled datasets to support the training of neural networks with few labeled data points [8]. For image classification, which is the most popular semi-supervised learning task, domain-specific strategies such as image augmentation have been proposed, which are not trivial to apply in our setting. Classification specific approaches such as using the softmax output of the network as confidence for a weak label [26] are not directly applicable to our task, since our desired output is a distribution and not a class.

For our semi-supervised training strategy, we adapt so-called “distant supervision” from other domains [10,14] by training the network on weak labels. Obtaining weak labels from more traditional interpolation methods and fine-tuning the network on labeled data afterwards is a new approach in this area.

3 Research Area and Dataset

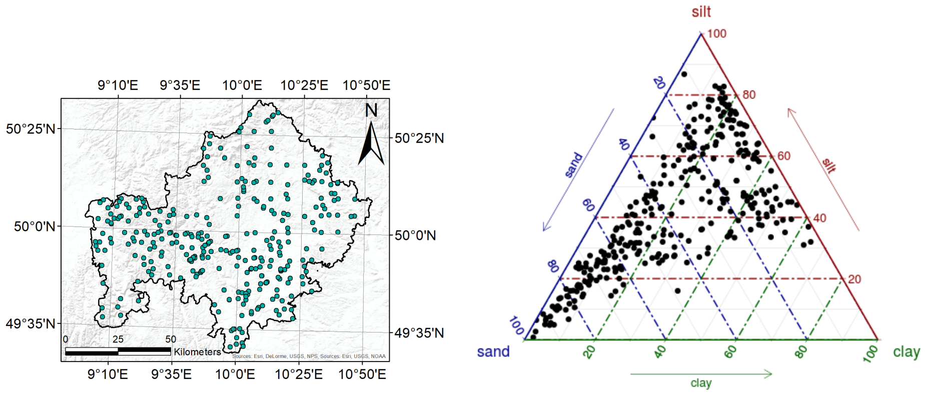
In this section, we describe the research area and the dataset we use for the interpolation task. Inputs to the interpolation models are the *latitude*, *longitude*, and multiple features from different auxiliary data sources that we suspect to have an influence on or are influenced by the grain size distribution. While only 315 locations have a target grain size distribution, the auxiliary data is widely available in a fine grid of 25 m × 25 m cells (overall 11 952 963 grid cells).

The research area is Lower Franconia, northern Bavaria, Germany. It covers 8530 km^2 and falls within 49.482°N to 50.566°N and 8.978°E to 10.881°E. The topography of this region is characterized by alluvial zones with surrounding low mountain ranging from 96 m to 927 m in altitude.

3.1 Target Variable: Grain Size Distribution

Soils are compositions of grain sizes. To get soil conditions for the research area, we use a soil profile database of the Bavarian Environment Agency (BEA)[1]. The

[1] unpublished data; reference: https://www.lfu.bayern.de/umweltdaten/



(a) Labeled data point locations. Map tiles by ESRI, USGS, NOAA, data by BEA.

(b) Distribution of target grain size distributions.

Fig. 2: Map showing labeled locations and distribution of the labels.

database covers detailed information on in-depth grain size distribution on 431 sites in Lower Franconia. The sampling took place in-between 1989 and 2017 and exposes grain size distributions of the fine earth fraction per soil-horizon through combined sieve and pipette analysis [12]. The method of sampling varies between drill cores and complete profile excavations.

While each observed location lists multiple layers, we limit the interpolation task to two dimensions by only using soil information from 14 cm–15 cm as most recorded layers span across this range. This common approach [6] results in 315 labeled locations, shown in Figure 2a.

Given the detailed grain sizes, we represent each location as a composition of three grain size classes [1]: **clay** (grain size $\leq 0.002\,\text{mm}$), **silt** ($0.002\,\text{mm} <$ grain size $\leq 0.063\,\text{mm}$), and **sand** ($0.063\,\text{mm} <$ grain size $\leq 2.000\,\text{mm}$). Each label is a three dimensional distribution vector, e.g. 20 % clay, 50 % silt, and 30 % sand. The label distribution is shown in Figure 2b. The task is to estimate this distribution for a location given other locations and auxiliary data.

3.2 Auxiliary Data

While there are only 315 labeled data points, auxiliary data is available for all locations in Lower Franconia (11 952 963 grid cells). For this work, we use a Digital Elevation Model (DEM) and meteorological data to generate ten features for each grid cell: *latitude*, *longitude*, *altitude*, *slope*, *Multi-Scale Topographic Position Index (minimum, mean, and maximum)*, *Topographic Wetness Index*, *temperature*, and *precipitation*, that are explained in the following.

The used DEM provided by the BEA[2] reflects the *altitude* of the terrain surface, excluding buildings and vegetation, resampled to our grid's spatial reso-

[2] https://geodatenonline.bayern.de/geodatenonline/seiten/dgm_info

lution of 25 m. We derive five additional features through topographic, morphometric and hydrographic analysis [25].

Slope. In basic terrain analysis, *slope* represents the change in elevation over a given distance. For a cell with altitude alt, we calculate the mean altitude over the neighboring cells in north and south direction $\overline{\mathrm{alt}}_{\mathrm{NS}}$ and in west and east direction $\overline{\mathrm{alt}}_{\mathrm{WE}}$. The slope ranges from 0° (a horizontal plane) to 90° and is calculated using $\mathrm{slope} = \frac{180}{\pi \cdot \sqrt{\left(\overline{\mathrm{alt}}_{\mathrm{NS}} - \mathrm{alt}\right)^2 + \left(\overline{\mathrm{alt}}_{\mathrm{WE}} - \mathrm{alt}\right)^2}}$.

Multi-Scale Topographic Position Index. The Topographic Position Index (TPI) [24] is defined as the altitude difference between a location of interest and the mean altitude of a square area around it, giving values that indicate local ridges and valleys. We obtain TPIs on multiple scales by altering the side length of the square from 3 grid cells (75 m) to 41 grid cells (1025 m) in steps of two cells, having the current location in the square's center. From the resulting 19 TPIs, we take the *minimum*, *mean*, and *maximum* as features. They describe the morphology of our study area at different scales as numeric factors.

Topographic Wetness Index. To represent spatial variations of soil moisture content and soil water drainage, a *terrain-based wetness index (TWI)* is computed [4]. The index is high for locations where water normally collects due to the topographic setting. It is calculated as a tangent function of the cell's slope angle w.r.t. the cell's area ($625\,\mathrm{m}^2$): $\mathrm{TWI} = \ln\left(\frac{625}{\tan(\mathrm{slope})}\right)$.

Meteorological Data. In addition to terrain based features described above, we also obtain meteorological data provided by the German Meteorological Service (DWD). The data reflects the 30-year (1971–2000) means of the monthly averaged mean daily air *temperature* 2 m above the ground and *precipitation*.[3] The grid-based data was obtained by accurate interpolation methods for temperature and precipitation at a resolution of $1\,\mathrm{km}^2$ [19] and resampled to the target grid size of 25 m using nearest neighbor interpolation.

4 Methodology

Given the data described above, we now have a large dataset of unlabeled data as well as a small labeled dataset. A neural network should now learn to estimate the grain size distribution of a location based on the ten input features. To make use of the large unlabeled dataset, we propose a three step semi-supervised training strategy that pre-trains the neural network on weak labels created by an interpolation method with stronger assumptions:

1. Weak Label Generation. We apply a common interpolation method such as Inverse Distance Weighting (IDW) on the small labeled dataset. Note that these methods usually do not take auxiliary data into account. Due to the strong modeling assumptions of such algorithms, they are able to work with

[3] https://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/ *air_temperature_mean* and *precipitation*

small datasets. The trained model then estimates the target labels for the large unlabeled dataset, which are used as weak labels in the next step.

2. Pre-Training. The neural network is pre-trained using the large amount of available weakly labeled data, thus being exposed to the property assumptions of the weak label generator. This way, the network learns representations from all input features, including the auxiliary data, and is guided to create more realistic outputs. Since interpolation methods such as IDW represent the location information as distances, the network has to learn from different features, as we will show in Section 6.1. Calculating the euclidean distance from locations is hard for the network, therefore it tries to find other correlations as well.

3. Fine-Tuning. The pre-trained network is fine-tuned on the labeled dataset. This reinforces or weakens some correlations the network has found. For fine-tuning, a smaller learning rate is used in order to keep the previously trained weights intact. The resulting model can then be used on all locations.

5 Experiments

Now, we compare our self-supervised training strategy to the traditional supervised method and other common interpolation methods on the grain size distribution task. Note that not all methods can output distributions, so we will only apply methods that are able to handle this task-specific output type.

5.1 Methods

Mean. Always predicts the mean of all training examples. As the average of multiple distributions is also a distribution, the prediction is valid.

k Nearest Neighbors (kNN). Calculates the average label of the nearest k training locations [2]. We set $k = 3$ based on a parameter search on validation data for $k \in \{1, \dots, 10\}$.

Inverse Distance Weighting (IDW). Same as kNN, but the average is inversely weighted based on the distance to a labeled location [22]. A parameter search for $k \in \{1, \dots, 10\}$ results in $k = 7$.

Multilayer Perceptron (MLP). Trains a Multilayer Perceptron on the labeled dataset in a supervised learning setting. The ten-dimensional input is normalized to zero mean and unit variance. It is then fed through three hidden layers with 256 neurons each with ReLU activation functions [9] in a batch of size 1024. The three-dimensional output is then converted to a probability distribution by applying the softmax activation function. These hyperparameters have been found on validation data. The standard cross entropy loss function is used that allows distributions as targets. The network is optimized with Adam [13] and a learning rate of 10^{-1} for at most 1000 epochs. Early stopping [18] stops the training if the validation loss does not improve at least 10^{-5} for ten epochs.

Semi-supervised MLP (SemiMLP). We apply our semi-supervised training strategy to the same MLP architecture as above. We generate weak labels using

Table 1: Test results (mean ± standard deviation) for each model. Best values are written in bold.

	MAE	**MSE**	**JSD**
Mean	0.5210 ± 0.0384	0.1337 ± 0.0183	0.0549 ± 0.0076
kNN	0.4267 ± 0.0412	0.1011 ± 0.0223	0.0398 ± 0.0090
IDW	0.4188 ± 0.0417	0.0954 ± 0.0225	0.0381 ± 0.0090
MLP	0.4361 ± 0.0552	0.1068 ± 0.0251	0.0426 ± 0.0088
SemiMLP (after pre-training)	0.4781 ± 0.0577	0.1296 ± 0.0283	0.0497 ± 0.0099
SemiMLP (after fine-tuning)	**0.4078 ± 0.0445**	**0.0952 ± 0.0195**	**0.0377 ± 0.0077**

the IDW baseline with $k = 7$ as it achieved the best baseline validation results. We train the network with learning rates 10^{-1} and 10^{-3} for pre-training and fine-tuning, respectively.

5.2 Evaluation

To evaluate the methods described above, we perform a ten-fold cross-validation (i.e. 31 or 32 examples per fold) using the labeled dataset. We average over 50 repetitions to account for the random initialization of the neural networks. Three metrics are used for evaluation: **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **Jensen-Shannon Divergence (JSD)**. While MAE and MSE compute the mean (absolute and squared) deviation from the correct values, JSD is specifically designed to measure the difference between two distributions [7]. Note that MAE and MSE sum the errors up for an example before averaging over all examples.

6 Results

Table 1 shows the test results for all models. The model with our training strategy (SemiMLP) yields the best test results. While the supervised MLP performs worse than kNN, the fine-tuned SemiMLP even improves the performance of the IDW baseline. In fact, a Wilcoxon signed rank test ($\alpha = 0.01$) on the MSE indicates that the improvement w.r.t. IDW is significant. We suspect that the network's improvement comes from having direct access to locations as well as auxiliary data that it uses during training, while IDW only relies on distances between locations as inputs.

6.1 Analysis

Pre-training matters. For our experiments, we altered the MLP baseline by adding the pre-training step to obtain SemiMLP, while the architecture and preprocessing were fixed. Thus, SemiMLP's better performance compared to

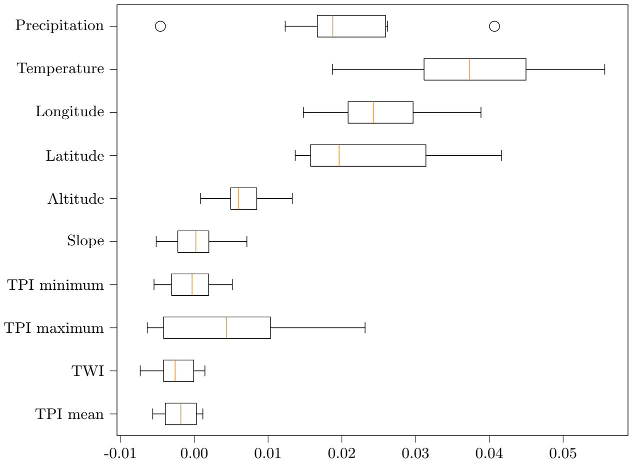


Fig. 3: Drop in MAE performance when the feature column was permuted.

MLP (cf. Table 1) shows that pre-training has a positive effect on SemiMLP. Pre-training the network seems to build better representations for the downstream task than random initialization.

Fine-tuning matters. While it helps, pre-training alone does not give superior performance. Table 1 shows that only pre-training on weak labels gives worse performance than most baselines and the supervised MLP. This indicates that the network is not able to imitate the IDW baseline, which generated the weak labels. This may be due to IDW using distances between new and labeled locations to assess its predictions. SemiMLP does not get distance information as input and is not able to directly access the labeled dataset. Thus, it learns a surrogate function that fits the training data but will not exactly match IDW's output for new data points. Also, SemiMLP gets more features than IDW, increasing the chance that the network exploits other correlations to predict the output. After the fine-tuning step, the method is superior to all baselines.

Auxiliary data matters. The features that may be influenced by or influence the target variable also have an effect on the performance. To investigate this, we apply the permutation importance for feature evaluation method [3] that permutes the values of a feature to see how much the predictive quality of the trained model changes. The more important a feature is, the higher the drop in performance if its input is altered. We average the features' importances for each test fold over ten different permutations to get more robust results.

Figure 3 shows the resulting feature importances. Besides location, the features temperature, precipitation, and altitude have the largest influence. According to previous research, soil is formed by the alteration of present bedrock under the influence of *climate*, *relief*, *organisms*, and *human activity* over time [21]. Since we do not provide features describing organisms and human activity, the model focuses on climatic (30-year means of temperature and precipitation) and relief-based (altitude) influences. While we expected other relief-based features such as TPI or TWI to be more important for the model, altitude and location seem to be descriptive enough.

7 Discussion

Neural networks make no modeling assumptions for the interpolation task. Compared to common interpolation methods, the network can model non-linear relationships in the data and can utilize any kind of auxiliary data. Our method circumvents the necessity of large training datasets by guiding the network towards more realistic outputs using weak labels before fine-tuning on few real labels. It is very easy to replace the weak label generator with a potentially better interpolation method. The required pre-training of the network on weakly labeled data takes extensively longer. However, depending on the neural network architecture, input data, and size of the research area, inference can be faster than other approaches, as we can compute outputs in batches on specialized hardware without any distance calculations.

As stated in Section 3, we restrict this work to the two-dimensional case of grain size distribution interpolation. While depth information is expected to increase performance, it is not trivial to use it in the weak label generation methods. Labeled locations usually have large distances (hundreds to thousands of meters), while labeled soil layers have very small distances (millimeters to few centimeters). Distance based approaches such as IDW will only take the nearest labeled location into account and average its soil layers as these are overall the closest to the desired location. While this is not resolved, building a model for each depth layer is the simplest approach that we can apply in practice.

8 Conclusion

In this paper we have proposed a semi-supervised training method for spatial interpolation tasks. For our grain size distribution task, additional pre-training on weak labels improved the network's performance compared to supervised learning and common interpolation methods. Testing other weak label generators and sampling strategies to optimize pre-training remains future work. Mixing weak labels from methods with different modeling assumptions might enrich the learned representations of the network. Future challenges include adding the depth dimension, allowing the exploitation of soil layer relations. Further, we will evaluate the interpolated map in a soil-hydrological simulation model.

Acknowledgements This research was conducted in the BigData@Geo project supported by the European Regional Development Fund (ERDF).

References

1. Ad-hoc-AG Boden: Bodenkundliche Kartieranleitung. Schweizerbart, 5 edn. (2005)

2. Altman, N.S.: An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician **46**(3) (1992)
3. Breiman, L.: Random forests. Machine learning **45**(1) (2001)
4. Böhner, J., Selige, T.: Spatial prediction of soil attributes using terrain analysis and climate regionalization. Gottinger Geographische Abhandlungen **115** (2002)
5. Dai, F., Zhou, Q., Lv, Z., Wang, X., Liu, G.: Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecological Indicators **45** (2014)
6. Deshmukh, K.K., Aher, S.P.: Particle Size Analysis of Soils and Its Interpolation using GIS Technique from Sangamner Area, Maharashtra, India **3** (2014)
7. Endres, D.M., Schindelin, J.E.: A new metric for probability distributions. IEEE Trans-IT **49**(7) (2003)
8. van Engelen, J.E., Hoos, H.H.: A survey on semi-supervised learning. Machine Learning **109**(2) (2020)
9. Glorot, X., Bordes, A., Bengio, Y.: Deep Sparse Rectifier Neural Networks. In: 14th AISTATS (2011)
10. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N project report, Stanford **1**(12) (2009)
11. Hengl, T.: A practical guide to geostatistical mapping. 2. extended edn. (2009)
12. ISO Central Secretary: Soil quality — determination of particle size distribution in mineral soil material — method by sieving and sedimentation. Tech. rep. (2009)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs] (2017)
14. Kobs, K., Zehe, A., Bernstetter, A., Chibane, J., Pfister, J., Tritscher, J., Hotho, A.: Emote-controlled: Obtaining implicit viewer feedback through emote-based sentiment analysis on comments of popular twitch.tv channels. ACM TSC **3**(2) (2020)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015)
16. Li, J., Heap, A.D.: Spatial interpolation methods applied in the environmental sciences: A review. Environmental Modelling & Software **53** (2014)
17. Meyer, S.: Climate change impact assessment under data scarcity. Dissertation, LMU München (2016)
18. Prechelt, L.: Early Stopping - But When? In: Neural Networks: Tricks of the Trade, vol. 1524 (1998)
19. Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., Gratzki, A.: A central european precipitation climatology part i: Generation and validation of a high-resolution gridded daily data set (hyras). Meteorologische Zeitschrift **22**(3) (2013)
20. Rezaei, K., Guest, B., Friedrich, A., Fayazi, F., Nakhaei, M., Beitollahi, A., Fatemi Aghda, S.M.: Feed forward neural network and interpolation function models to predict the soil and subsurface sediments distribution in Bam, Iran. Acta Geophysica **57**(2) (2009)
21. Semmel, A.: Relief, Gestein, Boden. Wiss. Buchges. (1991)
22. Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: 23rd ACM national conference (1968)
23. Tarasov, D., Buevich, A., Sergeev, A., Shichkin, A.: High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging. Applied Geochemistry **88** (2018)
24. Weiss, A.: Topographic position and landforms analysis. In: Poster presentation, ESRI user conference, San Diego, CA. vol. 200
25. Wilson, J.P.: Terrain analysis. Wiley (2000)
26. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: IEEE/CVF CVPR (2020)