Predicting water permeability of the soil based on open data

Jonne Pohjankukka, Paavo Nevalainen, Tapio Pahikkala, Eija Hyvönen, Maarit Middleton, Pekka Hänninen, Jari Ala-Ilomäki, and Jukka Heikkonen

University of Turku, Computer Science Dept.
{Jonne.Pohjankukka,ptneva,Tapio.Pahikkala}@utu.fi,
{Eija.Hyvonen,Pekka.Hanninen,Maarit.Middleton}@gtk.fi,
Jukka.Heikkonen@utu.fi, jari.ala-ilomaki@metla.fi,
http://www.utu.fi/en/units/sci/units/it/

Abstract. Water permeability is a key concept when estimating load bearing capacity, mobility and infrastructure potential of a terrain. Northern sub-arctic areas have rather similar dominant soil types and thus prediction methods successful at Northern Finland may generalize to other arctic areas. In this paper we have predicted water permeability using publicly available natural resource data with regression analysis. The data categories used for regression were: airborne electro-magnetic and radiation, topographic height, national forest inventory data, and peat bog thickness. Various additional features were derived from original data to enable better predictions. The regression performances indicate that the prediction capability exists up to 120 meters from the closest direct measurement points with concordance index 0.66 at 75 meters. The results were measured using leave-one-out cross-validation with a dead zone between the training and testing data sets.

Key words: load bearing capacity of soil, water permeability, regression, k-nearest neighbor, mobility, sub-arctic infrastructure.

1 Introduction

This paper is about predicting the water permeability of the soil by regression analysis using publicly available multi-source data. Water permeability (also called hydraulic conductivity) is a central soil property related to soil type and soil texture. High permeability means that soil tends to stay dry and traversable most of the year, whereas low permeability creates a risk for mobility when precipitation is high. Mobility in arctic areas is of great interest to many different parties. E.g. the mining industry is interested about the mobility estimates when placing various facilities. The forest industry is interested on the load bearing capacity of the soil, since the route solutions can be adaptive to mobility predictions. Peat bog and mire areas are a high risk for heavy machinery and predetermined knowledge of their locations is required.

Our input data set consists of 44 features which are publicly available. The data is in the raster format with grid resolution ranging from 10 meters to 50 meters.

Water permeability of the soil has been measured in 1788 test spots at Northern Finland provided by the Geological Survey of Finland (GTK). It is an important attribute which, when combined with other features available, helps to determine the soil types. Related studies have been conducted in [1] where soil respiration rates are predicted from temperature, moisture content and soil type. Another related research was published in the paper of P. Scull, J. Franklin and O.A. Chadwick [2]. In their paper they use classification tree analysis for predicting the soil type in desert landscapes. Related work has been done also by R.P.O. Schulte et al.[3] and H. Gao et al. [4]. Other related research was conducted by R.A. Chapuis [5], R. Kiss [6], H.S. Mahmood et al. [7], N.J. McKenzie [8], I.D. Moore et al. [9], A.T. Ramli et al. [10] and J.V.A. Zachary [11]. The main novelty of this paper related to the previous studies is that the prediction is based on wide-area public data. The features used in this paper are basically available through-out the arctic zone.

We use regression analysis to find a mapping between the publicly available data and water permeability of the soil. In the following, we present the regression methods in Ch. 2. Then we introduce the test area, the original data sets and derived features (Ch. 3) and describe the analysis process and results of the analysis (Ch. 4). The last part is for conclusions and future approaches (Ch. 5).

2 Regression methods

Regularized least squares regression (RLS) is well known so we describe it mainly to introduce the variables and the notation used later in the paper. The explanatory variables $x_1, ..., x_p$ consist of given data and dependent variable y is the water permeability. We need to find a set of parameters $\mathbf{w} \in \mathbb{R}^p$ and $b \in \mathbb{R}$ such that the error function:

$$E(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \mathbf{w}^T \mathbf{x}_i - b \Big)^2 + \frac{\lambda}{n} \mathbf{w}^T \mathbf{w} \tag{1}$$

is minimized, where $\mathbf{x}_i \in \mathbb{R}^p$ is the input vector, $y_i \in \mathbb{R}$ is the response value, n is the number of observations and λ is the regularization parameter.

The k-nearest neighbors (k-NN) approach predicts the test sample by taking the average from k points nearest to it. Euclidean distance is the used metric in our analysis. Explicitly stated, if $y_1, ..., y_k$ are the response values of the k-nearest points to the test sample, then the response value for the test sample $\hat{y}_t$ is:

$$\hat{y}_t = \frac{1}{k} \sum_{i=1}^{k} y_i.$$

3 Test area, data sets and features

The research area is located in the northern part of the municipality of Sodankylä, which is a part of Finnish Lapland. The size of the target area is 18432 km^2. The center point of the rectangular target area is at ETRS-TM35FIN coordinates 7524 kmN, 488 kmE, zone 35.

The data set consists of aerial gamma-ray spectroscopy data (referred later as gamma-ray data, AGR) combined with electromagnetic (AEM), topographical (Z), peat bog mask (PBM) and The National Forest Inventory 2011 (VMI[1]) data when predicting the qualities and characteristics of the soil, namely its type and water permeability (WP). Gamma-ray data is inversely related on the amount of water on the soil, which can be used to predict the type of the soil. The forest inventory data describes the profile of tree species, their maturity and foresting state. Albeit this kind of data is not directly available elsewhere in northern subarctic areas (e.g. Russia, Canada), several studies are underway to predict the main characteristics of the forest by remote measurement methods [12]. These methods include LiDAR and various satellite measurements.

The data providers are:

Provider	Data	Grid size
Geological Survey of Finland (GTK)	AGR, AEM WP	50 m
Finnish Forest Research Institute (Metla)	VMI, PBM	20 m
National Land Survey of Finland (NLS)	Z	10 m

Table 1: Data providers, data and the grid size.

When considering all the derived features used in the analysis we get a total of 96 data layers.

The test site has 1788 sample points, where many mechanical and electrochemical properties of the soil were measured, see [13]. The water permeability is a theoretical value derived from the soil particle size distribution of the soil.

We now present our data sources and donors.

3.1 Forest inventory data

The National Forest Inventory (VMI) holds the state of Finnish forests. The data is updated once in two years. The parameters are derived from various remote sensing sources, and several spotwise verification and calibration methods are applied to it before publishing the data [14]. 44 numerical features include green mass, trunk dimensions and tree density per specie category. These multi-source features exhibit built-in dependencies, thus the final number of useful features is lower.

[1] VMI2011: http://www.metla.fi/ohjelma/vmi/vm11-info-en.html

3.2 Aerial gamma-ray data

The aerial gamma-ray data was provided by the Geological Survey of Finland (GTK). The raster data is based on gamma-ray flux from potassium, which is the decay process of the naturally occurring chemical element potassium (K). This data indicates many significant characteristics of the soil, including the tendency to stay moist after precipitation and tendency to frost heaving. Also the soil type, especially density, porosity, grain size and humidity of the soil have an effect to gamma-ray radiation. In Fig. 1 we present the gamma-ray data from Sodankylä target area. The bright end of the gray scale is for the high gamma radiation and hence less water in the locality of the pixel.

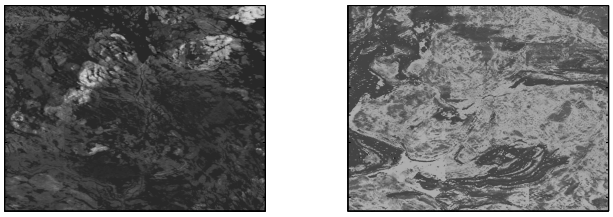

Fig. 1: Aerial data: gamma-ray (left) and electromagnetic data (right). Air-borne electromagnetic data is sensitive to geological properties to depth of hundreds of meters, but it also indicates some features of the top soil.

3.3 Electromagnetic properties of soil

The air-borne electromagnetic (AEM) data was provided by the Geological Survey of Finland (GTK). Primary AEM components, in-phase and quadrature, were transformed to apparent resistivity values by using a half-space model [15]. The apparent resistivity gives information on different kind of soil conductors. The apparent resistivity is governed by grain size distribution, water and electronic conductors content of soil and cumulative weathering.

3.4 Topographical height data

Topographical data provided by the National Land Survey of Finland (NLS) was included in the analysis. The data from NLS server is basically similar to aerial laser measurements (LiDAR) except LiDAR can reach denser grid. Instead of raw height alone we used local height difference, flow accumulation area, confluence and inclination described in [16]. These four derived features are more efficient for prediction than raw height data alone.

3.5 Peat bog mask

Peat bog mask is created from GTK aero-radiometric data and is courtesy of NLS and METLA. The grid size is 20 m and the value 1 indicates that peat thickness is over 60 cm. Value 0 indicates thickness less than 60 cm. The limit chosen is practical for mobility prediction.

3.6 Derived features

The following features were derived from gamma-ray and electromagnetic data:

- Mean and variance over 3×3 window
- Mean and variance over Gabor filter with 8 orientations, see [17]
- Local Binary Pattern (LBP) with pixel radii $r \in \{1, 2\}$, see [18]

From topographical height we derived the following features: local height difference, ground inclination, convergence index and flow accumulation area. The definition of these features is at [6]).

There are several additional attributes possible to derive from topographical height data, and more geomorphological features will be employed in the future.

The regression methods use total of 44 original and 52 derived features, including the constant feature. The derived features are useful only if the original feature is continuous enough. E.g. the Forest Inventory data often has locally constant zones with abrupt changes and the derived features do not help much.

3.7 Water permeability exponent

This is the subject of prediction. Basically, the water permeability indicates the nominal vertical speed of water through the soil sample. The measurement of this quantity is indirect, based on soil particle size distribution, and the actual speed highly depends on the inhomogeneities (roots, rocks) and micro-cracks in the soil. This is why this quantity is descriptive and theoretical. In our analysis we are using a logarithmic quantity x_{wp} derived from water permeability speed v. For purposes of this presentation it is called as the water permeability exponent and defined as:

$$x_{wp} = -\log_{10} v, \; [v] = \frac{m}{sec}, \tag{2}$$

This formula has v as the vertical speed of water flow through the soil.

4 Analysis and results

We are looking for methods which predict water permeability on areas, where there may not be direct water permeability measurements nearby. Therefore, we developed a modification of the leave-one-out cross-validation (LOOCV) for measuring the degree of spatial dependency from the nearby direct measurements, which we refer to as LOOCV with dead zone. Namely, the approach works on the

measurement data just like an ordinary LOOCV in which each measurement at a time is removed from the training set and used as a test point, except that we also remove from the training set all points that are within geographical distance r from the test point. This approach is illustrated in Fig. 2. By varying r, we can measure how far from the test area we assume the closest measurements to be at the very least. In addition, the results can be helpful in deciding how dense grid of direct measurement one should use in order to obtain a certain level of prediction performance.

We perform the regression of water permeability with the following three feature sets:

- location only
- features + location
- features only

where location refers to the geographical coordinates (e.g. latitude and longitude) and features to the ones described in Section 3. Note that one can not rely on the location information if there are no nearby direct measurements at all, and therefore we measure the prediction performance separately with these.

The prediction performances with the different feature sets as a function of the radius r of the dead zone are depicted in the two leftmost graphs in Fig. 3 on p. 8. The generic version based on feature data only gives weaker results, since the sample point arrangement at Sodankylä (see sample sets A and B in Fig. 2) and perhaps the phenomenon itself induce spatial dependency. No good generic regression method for this data set has been found, instead the problem is about how much additional samples are needed per target area to make the prediction useful.

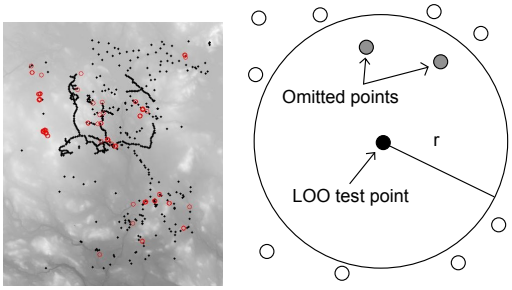

Fig. 2: *Left*: 1788 sample points. Set A (1187 points, marked with red circles, distance to the nearest neighbor $d_{NN} \leq 86m$) is tightly packed and set B is very sparse (601 points, marked with black dots, aver. $d_{NN} \approx 1.1\,km$). *Right*: the dead zone (with radius r) around the leave-one-out test point (black circle). The gray circles are omitted from the training set (white circles). Both the k-NN and RLS method address the training data only, e.g. the k nearest neighbours are selected from outside the circle.

The common k-NN method has one essential parameter, the number of neighbors k. The spatial dependency can be probed by adding the dead zone radius r

to avoid the optimisitc effect of the nearest neighbors. Fig. 2 depicts the modified leave-one-out arrangement, where k nearest points outside the dead zone of radius r are used for teaching. By varying r one gets a varied dataset and a rough estimate on how dense it should be for it to predict well in new circumstances.

The same parameterized dead zone leave-one-out arrangement was used with regression, too.

4.1 Predicting water permeability

As mentioned before, the prediction subject is the water permeability exponent x_{wp}. The values used for regularization parameter λ ranged from $2^{-15}, ..., 2^{15}$. k-NN parameter had $k \in \{1, 3, 6, 12, 22\}$. Two different error measures were used for estimating prediction performance: mean absolute error (MAE) and concordance index (CI) [19]. Explicitly, the error measures are:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad CI = \frac{1}{N}\sum_{y_i<y_j} h(\hat{y}_i - \hat{y}_j) \tag{3}$$

MAE baseline $\tilde{y}$ is the best possible prediction under the assumption that the prediction will be constant: MAE baseline $= \arg\min_{\tilde{y}} \frac{1}{n}\sum_{i=1}^{n}\left|(y_i - \tilde{y})/y_i\right|$ The prediction performance should be better than this to be useful. The corresponding persentage values (MAPE and MAPE baseline) have been used in the rest of the text.

In equation (3) we denote $N = \mid \{(i,j) \mid y_i > y_j\} \mid$ as the normalization constant which equals to the number of data pairs with different label values and $h(u)$ is the step function returning 1.0, 0.5 and 0.0 for $u > 0$, $u = 0$ and $u < 0$, respectively. The values of the C-index range between 0.0 and 1.0, where 0.5 corresponds to a random predictor and 1.0 to the perfect prediction accuracy in the test data.

4.2 Results

The results for regression analysis can be seen in Fig. 3 on p. 8.

Both MAPE and C-index indicate rather good prediction performance to the distance of 120 m from the nearest soil sample point. This is seen both with k-NN and RLS methods. When MAPE is higher than the baseline, it is better to use baseline average than the prediction. MAPE baseline is the horizontal line in the lower figures in Fig. 3.

The dead zone radius $r > 0$ simulates a situation, where the test point is at least r distance away from the given training points. $r = 0$ is traditional LOO test arrangement and measures best the properties of the predicted value within the training set itself. It may be too optimistic, since we seek for generalization. A large radius $r \approx \infty$ is overly pessimistic, since it would use only tiny fragments of the training set and would completely distort the prediction.

The prediction performance near $r = 0$ seems to indicate rather good generalization ability, but the performance reduces drastically over the dead zone

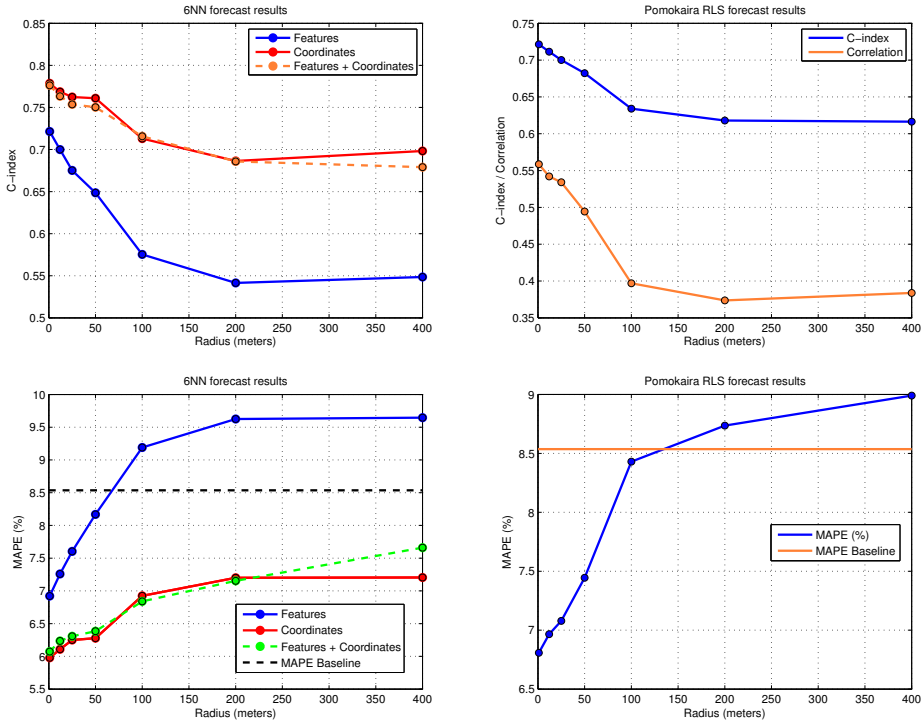

Fig. 3: *Left*: k-NN results with $k = 6$ and 3 different feature sets. *Right*: RLS results on features-only case. C-index and Pearson correlation at top and MAPE below. The prediction performance is adequate below 120 meters.

distance r. Further study, both theoretical and practical, must be done to properly address classical geoinformatics concepts such as spatial autocorrelation and spatial semivariance together with the general prediction ability. The problem is new, since spatial analysis in geosciences is usually applied to one variable only (e.g. topological height), but this type of problem has many (in this case 96+2) features from which only 2 are locational coordinates.

5 Conclusions and future work

The results indicate that the chosen five data sources (forest inventory, gamma-ray, air-borne electromagnetic, topographical data and peat bog mask) can be used to estimate the water permeability to a certain range from known measurements. This range seems to be c. 120-150 m. The best results come from the k-NN method based on the location of the sample points only. This method is naturally unavailable for general prediction.

The prediction efficiency in the features-only case is relatively low, but since the target application is the forest mobility, road and route planning, even a small local prediction power leads to cumulative results over long periods. Also, the forest harvesters will be capable of indirect load bearing capacity measurementsduring the operation. The heavier load hauling harvesters may be able to improve their track based on the earlier observations of the light-weight logging harvesters.

The mapping from water permeability to soil types is not unique, see [20]. Given the water permeability prediction, a special majority rule could be used to select the dominant soil type from neighboring grid point predictions. Such expert rules would require additional features like sophisticated geomorphological categories.

For forestry applications, the most practical tool for direct load bearing capacity measurement is the spiked shear vane [21]. It yields shear modulus of soil, which can be converted to modulus of elasticity. Modulus of elasticity can be directly used as input to wheel sinkage calculus. Spiked shear vane output was however not available in the field test data.

There is also a possibility to use aerial Light Detection and Ranging (LiDAR) data instead of the topographical and forest inventory data. This would enable the expansion of the scope of this study to any location at the arctic zone, where only aerial and satellite measurements are economical. Also LiDAR has more potential for derived features like geological morphology [22] and soil water budget modeling [12]. The final goal is to predict the water permeability, soil types, approximate water budget and the load bearing capacity of the terrain in relation to the given weather forecast, while the model is based on remote measures only. This remains a subject of the further study.

The potential applications aim to wide-area routing and location planning. In this regard, even a modest prediction power could yield a cumulative effect on route decision.

References

1. Azzalini, A., Diggle, P.: Prediction of soil respiration rates from temperature, moisture and soil type. Journal of the Royal Statistical Society - Series C: Applied Statistics **43** (1994) 505–526
2. Sculla, P., Franklin, J., Chadwick, O.: The application of classification tree analysis to soil type prediction in a desert landscape. Ecological Modelling **181** (2005) 1–15
3. Schulte, R., Diamond, J., Finkele, K., Holden, N., Brereton, A.: Predicting the soil moisture conditions of irish grasslands. Irish Journal of Agricultural and Food Research **44** (2005) 95–110
4. Gao, H., Tang, Q., Shi, X., Zhu, C., Bohn, T.J., Su, F., Sheffield, J., Pan, M., Lettenmaier, D.P., Wood, E.F.: Water budget record from variable infiltration capacity (vic) model. Algorithm Theoretical Basis Document for Terrestrial Water Cycle Data Records (2010)
5. Chapuis, R.: Predicting the saturated hydraulic conductivity of soils: a review. Bulletin of Engineering Geology and the Environment **71** (2012) 401–434

6. Kiss, R.: Determination of drainage network in digital elevation models, utilities and limitations. Journal of Hungarian Geomathematics **2** (2004) 16–29
7. Mahmood, H., Hoogmoed, W., van Henten E.J.: Proximal gamma-ray spectroscopy to predict soil properties. Sensors **13** (2013) 16263–16280
8. McKenzie, N., Ryan, P.: Spatial prediction of soil properties using environmental correlation. Geoderma **89** (1999) 67–94
9. Moore, I., Gessler, P., Nielsen, G., Peterson, G.: Soil attribute prediction using terrain analysis. Soil Science Society of America Journal **57** (1993) 443–452
10. Ramli, A., Rahman, A., Lee, M.: Statistical prediction of terrestrial gamma radiation dose rate based on geological features and soil types in kota tinggi district, malaysia. Applied Radiation and Isotopes **59** (2003) 393–405
11. Zachary, J.: Using topographic and soils data to understand and predict field scale moisture patterns. Master's thesis, Iowa State University (2012)
12. Leutner, B., Müller, H., Wegmann, M., Beierkuhnlein, C.: Modelling biodiversity and forest structure using hyperspectral. In: 41st Annual Meeting of the Ecological Society of Germany. (2011)
13. Hyvönen, E., Pänttäjä, M., Sutinen, M.L., Sutinen, R.: Assessing site suitability for scots pine using airborne and terrestrial gamma-ray measurements in finnish lapland. Canadian Journal of Forest Research **33-5** (2003) 796–806(11)
14. Tomppo, E., Katila, M., Mäkisara, K., Peräsaari, J.: Multi-source national forest inventory methods and applications. Volume 18 of Managing Forest Ecosystems. Springer (2008)
15. Hautaniemi, H., Kurimo, M., Multala, J., H., L., Vironmäki, J.: The three in one aerogeophysical concept of gtk in 2004. Geological Survey of Finland, Special Paper **39** (2005) 21–74
16. Schwanghart, W., Kuhn, N.: Topotoolbox: a set of matlab functions for topographic analysis. Environmental Modelling & Software **25** (2010) 770–781
17. Weldon, T.P., Higgins, W.E., Dunn, D.F.: Efficient gabor filter design for texture segmentation. Pattern Recognition (1996) 2005 – 2015 This could be removed, a more arcaic Gabor reference instead!
18. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture. IEEE Transactions on Pattern Analysis and Machine Intel- **24** (2002) 971987
19. Gönen, M., Heller, G.: Concordance probability and discriminatory power in proportional. Biometrika **92** (2005) 965–970
20. Pohjankukka, J., Nevalainen, P., Pahikkala, T., Hyvönen, E., Sutinen, R., Hänninen, P., Heikkonen, J.: Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. Proceedings of the 22nd International Conference on (ICPR 2014) (2014) to appear.
21. Ala-Ilomäki, J.: Spiked shear vane - a new tool for measuring peatland top layer strength. Mires and Peat **64(2-3)** (2013) 113–118
22. Sutinen, R., Hyvönen, E., Middleton, M., Ruskeeniemi, T.: Airborne lidar detection of postglacial faults and pulju moraine. Global and Planetary Change (2014) 24–32

ACKNOWLEDGEMENTS

This work is done as a part of ULJATH project, which is funded by the *Finnish Funding Agency for Technology and Innovation* (TEKES). ULJATH stands for *New Computational Methods For The Efficient Utilization of Public Data*.