



Jonne Pohjankukka

Machine Learning Approaches for Natural Resource Data



TURKU CENTRE *for* COMPUTER SCIENCE

TUCS Dissertations
No 232, June 2018

Machine Learning Approaches for Natural Resource Data

Jonne Pohjankukka

*To be presented, with the permission of the Faculty of Science and
Engineering of the University of Turku, for public criticism in
Auditorium XX on June 15th, 2018, at 12 noon.*

University of Turku
Department of Future Technologies
Vesilinnantie 5, 20500 Turku, Finland

2018

Supervisors

Professor Jukka Heikkonen
Department of Future Technologies
University of Turku
Finland

Assistant Professor Tapio Pahikkala
Department of Future Technologies
University of Turku
Finland

Reviewers

Associate Professor Tuomo Kauranne
School of Engineering Science
Lappeenranta University of Technology
Finland

Associate Professor Mikko Vastaranta
School of Forest Sciences
University of Eastern Finland
Finland

Opponent

Professor Markus Holopainen
Department of Forest Sciences
University of Helsinki
Finland

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service

ISBN 978-952-12-3710-2
ISSN 1239-1883

To my family

Omistettu perheelleni

Abstract

Real life applications involving efficient management of natural resources are dependent on accurate geographical information. This information is usually obtained by manual on-site data collection, via automatic remote sensing methods, or by the mixture of the two. Natural resource management, besides accurate data collection, also requires detailed analysis of this data, which in the era of data flood can be a cumbersome process. With the rising trend in both computational power and storage capacity, together with lowering hardware prices, data-driven decision analysis has an ever greater role.

In this thesis, we examine the predictability of terrain trafficability conditions and forest attributes by using a machine learning approach with geographic information system data. Quantitative measures on the prediction performance of terrain conditions using natural resource data sets are given through five distinct research areas located around Finland. Furthermore, the estimation capability of key forest attributes is inspected with a multitude of modeling and feature selection techniques. The research results provide empirical evidence on whether the used natural resource data is sufficiently accurate enough for practical applications, or if further refinement on the data is needed. The results are important especially to forest industry since even slight improvements to the natural resource data sets utilized in practice can result in high saves in terms of operation time and costs.

Model evaluation is also addressed in this thesis by proposing a novel method for estimating the prediction performance of spatial models. Classical model goodness of fit measures usually rely on the assumption of independently and identically distributed data samples, a characteristic which normally is not true in the case of spatial data sets. Spatio-temporal data sets contain an intrinsic property called spatial autocorrelation, which is partly responsible for breaking these assumptions. The proposed cross validation based evaluation method provides model performance estimation where optimistic bias due to spatial autocorrelation is decreased by partitioning the data sets in a suitable way.

Keywords: Open natural resource data, machine learning, model evaluation

Tiivistelmä

Käytännön sovellukset, joihin sisältyy luonnonvarojen hallintaa ovat riippuvaisia tarkasta paikkatietoaineistosta. Tämä paikkatietoaineisto kerätään usein manuaalisesti paikan päällä, automaattisilla kaukokartoitusmenetelmillä tai kahden edellisen yhdistelmällä. Luonnonvarojen hallinta vaatii tarkan aineiston keräämisen lisäksi myös sen yksityiskohtaisen analysoinnin, joka tietotulvan aikakautena voi olla vaativa prosessi. Nousevan laskentatehon, tallennustilan sekä alenevien laitteistohintojen myötä datapohjainen päätöksenteko on yhä suuremmassa roolissa.

Tämä väitöskirja tutkii maaston kuljettavuuden ja metsäpiirteiden ennustettavuutta käyttäen koneoppimismenetelmiä paikkatietoaineistojen kanssa. Maaston kuljettavuuden ennustamista mitataan kvantitatiivisesti käyttäen kaukokartoitusaineistoa viideltä eri tutkimusalueelta ympäri Suomea. Tarkastelemme lisäksi tärkeimpien metsäpiirteiden ennustettavuutta monilla eri mallintamistekniikoilla ja piirteiden valinnalla. Väitöstyön tulokset tarjoavat empiiristä todistusaineistoa siitä, onko käytetty luonnonvara-aineisto riittävän laadukas käytettäväksi käytännön sovelluksissa vai ei. Tutkimustulokset ovat tärkeitä erityisesti metsäteollisuudelle, koska pienetkin parannukset luonnonvara-aineistoihin käytännön sovelluksissa voivat johtaa suuriin säästöihin niin operaatioiden ajankäyttöön kuin kuluihin.

Tässä työssä otetaan kantaa myös mallin evaluointiin esittämällä uuden menetelmän spatiaalisten mallien ennustuskyvyn estimointiin. Klassiset mallinvalintakriteerit nojaavat yleensä riippumattomien ja identtisesti jakautuneiden datanäytteiden oletukseen, joka ei useimmiten pidä paikkaansa spatiaalisilla datajoukoilla. Spatio-temporaaliset datajoukot sisältävät luontaisen ominaisuuden, jota kutsutaan spatiaaliseksi autokorrelaatioksi. Tämä ominaisuus on osittain vastuussa näiden oletusten rikkomisesta. Esitetty ristiinvalidointiin perustuva evaluointimenetelmä tarjoaa mallin ennustuskyvyn mitan, missä spatiaalisen autokorrelaation vaikutusta vähennetään jakamalla datajoukot sopivalla tavalla.

Avainsanat: Avoin luonnonvara-aineisto, koneoppiminen, mallin evaluointi

Acknowledgements

I feel privileged to have been given this opportunity. When I was accepted to study at the Department of Information Technology in University of Turku, I could have never imagined that one day I would be graduating from of a doctoral programme. I had a keen interest to understand computer science and mathematics but I did not really have a clear direction where to strive towards. After being introduced to the field of data analysis, I felt my direction was becoming clear. Data analysis felt very interesting to me with rigorous theoretical background and loads of potential for practical applications in the era of information flood.

A doctoral thesis is not just the achievement of a single person, but the result of work and influence of many people to whom all I owe my gratitude. I want to thank my supervisors Professor Jukka Heikkonen and Assistant Professor Tapio Pahikkala for introducing me the world of data analysis. They are among the first people responsible for me being currently at this stage. I thank Professor Olli Nevalainen and Professor Tapio Salakoski for guiding and believing in me during my studies. I also thank Associate Professor Tuomo Kauranne and Associate Professor Mikko Vastaranta for acting as pre-examiners and constructive feedback allowing me to improve the quality of this thesis. I would also like to express my gratitude to Professor Markus Holopainen for agreeing to act as my opponent.

I want to thank all my colleagues and people I have been in contact with in our department for their support and discussions. These people include Paavo Nevalainen, Anne-Maarit Majanoja, Markus Viljanen, Parisa Movahedi, Ileana Montoya-Perez, Jussi Toivonen, Elise Syrjälä, Pekka Naula and Antti Airola. Special thanks to Anne-Maarit for helping me out through the last stages of my thesis. Lastly, but certainly not least, I thank my family. I could not have done it without you!

Lieto, May 2018
Jonne Pohjankukka

List of original publications

- I Jonne Pohjankukka, Paavo Nevalainen, Tapio Pahikkala, Eija Hyvönen, Raimo Sutinen, Pekka Hänninen and Jukka Heikkonen. Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. In Magnus Borga, Anders Heyden, Denis Laurendeau, Michael Felsberg, and Kim Boyer, editors, Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014), pages 1822–1827. IEEE, 2014.
- II Jonne Pohjankukka, Paavo Nevalainen, Tapio Pahikkala, Eija Hyvönen, Pekka Hänninen, Raimo Sutinen, Jari Ala-Ilomäki and Jukka Heikkonen. Predicting water permeability of the soil based on open data. In Lazaros Iliadis, Ilias Maglogiannis, and Harris Papadopoulos, editors, Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014), volume 436 of IFIP Advances in Information and Communication Technology, pages 436–446. Springer, 2014.
- III Jonne Pohjankukka, Henri Riihimäki, Paavo Nevalainen, Tapio Pahikkala, Jari Ala-Ilomäki, Eija Hyvönen, Jari Varjo and Jukka Heikkonen. Predictability of boreal forest soil bearing capacity by machine learning. *Journal of Terramechanics*, 68:1–8. Elsevier, 2016.
- IV Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen and Jukka Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019. Taylor & Francis, 2017.
- V Jonne Pohjankukka, Sakari Tuominen, Juho Pitkänen, Tapio Pahikkala and Jukka Heikkonen. Comparison of estimators and feature selection procedures in area-based forest inventory based on airborne laser scanning and digital aerial imagery. *Scandinavian Journal of Forest Research*, Accepted. Taylor & Francis, 2018.
- VI Antti Airola, Jonne Pohjankukka, Johanna Torppa, Maarit Middleton, Vesa Nykänen, Jukka Heikkonen and Tapio Pahikkala. Reliable AUC estimation of spatial classifiers, with application to mineral prospectivity mapping. *Data Mining and Knowledge Discovery*, Accepted. Springer, 2018.

Contents

1	Introduction	1
1.1	Data analysis with open data	1
1.2	Motivation for the research	3
1.3	Main objectives of the research	4
1.4	Organization of the thesis	5
2	Theoretical foundations	7
2.1	Geographic information systems	7
2.2	Remote sensing methods	8
2.2.1	Sensors	8
2.2.2	Data collection techniques	10
2.3	Data representation	12
2.4	Open data sets	12
2.5	Spatial data analysis	14
2.5.1	Spatial autocorrelation	14
2.5.2	Measures of spatial autocorrelation	16
2.6	Machine learning	17
2.6.1	The learning setup	18
2.6.2	Types of learning	20
2.7	Model complexity selection	22
2.7.1	Overfitting	23
2.7.2	Bias-variance trade-off	24
2.7.3	Information criteria	25
2.7.4	Regularization	31
2.7.5	Cross validation	34
2.7.6	Spatial k-fold cross validation	36
2.8	Feature selection	37
2.8.1	Greedy forward/backward selection	38
2.8.2	Genetic algorithm	39
2.8.3	Automatic relevance determination	40

3	Research studies and results	43
3.1	Research publications	43
3.1.1	Publication I: Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data	43
3.1.2	Publication II: Predicting water permeability of the soil based on open data	45
3.1.3	Publication III: Predictability of boreal forest soil bearing capacity by machine learning	47
3.1.4	Publication IV: Estimating the prediction performance of spatial models via spatial k-fold cross validation	49
3.1.5	Publication V: Comparison of estimators and feature selection procedures in forest inventory based on airborne laser scanning and digital aerial imagery	52
3.1.6	Publication VI: Reliable AUC estimation of spatial classifiers, with application to mineral prospectivity mapping	55
3.2	Research results	58
3.2.1	(RQ1): Are the provided open natural resource data sets applicable in predicting terrain conditions and forest attributes in Finland?	58
3.2.2	(RQ2): How to evaluate the prediction performance of a model involving spatially dependent data?	59
4	Conclusions	61
4.1	Summary of the thesis	61
4.2	Discussion and outcomes	62

List of Abbreviations

IC	information criterion
AEM	airborne electromagnetic
AIC	Akaike information criterion
ALS	airborne laser scanning
ARD	automatic relevance determination
BIC	Bayesian information criterion
C-index	concordance index
CRISP-DM	cross-industry standard process for data mining
CV	cross validation
DAI	digital aerial imagery
DEM	digital elevation model
EM	electromagnetic
FGI	Finnish Geospatial Research Institute
FMI	Finnish Meteorological Institute
GA	genetic algorithm
GBS	greedy backward selection
GFS	greedy forward selection

GIS	geographic information system
GPS	global positioning system
GTK	Geological Survey of Finland
i.i.d.	independent and identically distributed
K-L	Kullback-Leibler
KCV	k-fold cross validation
kNN	k-nearest neighbor
LBP	local binary pattern
LiDAR	light detection and ranging
LOO-SCV	leave-one-out spatial cross validation
LOOCV	leave-one-out cross validation
LOOCVDZ	leave-one-out cross validation with a dead zone
LPO-SCV	leave-pair-out spatial cross validation
LS	least squares
LUKE	Natural Resource Institute Finland
MAP	maximum a posteriori
MCMC	Markov chain Monte Carlo
MDL	minimum description length
ML	machine learning
MLE	maximum likelihood estimation
MLP	multilayer perceptron
MLP-ESC	multilayer perceptron early-stopping committee
MPM	mineral prospectivity mapping

MS-NFI	multi-source national forest inventory
MSE	mean squared error
NCV	nested cross validation
NGFS	nested greedy forward selection
NLS	National Land Survey of Finland
NRMSE	normalized root mean squared error
RADAR	radio detection and ranging
RLS	regularized least squares, ridge regression
RMSE	root mean squared error
RQ1, RQ2	research questions 1 and 2
RS	remote sensing
RVM	relevance vector machines
SAC	spatial autocorrelation
SAR	synthetic aperture radar
SE	squared error
SKCV	spatial k-fold cross validation
SKCV-RLO	spatial k-fold cross validation random-leave-out
SVM	support vector machines
TIC	Takeuchi's information criterion

Nomenclature

Φ	design matrix
ϕ	vector of basis functions

Θ	set of parameter vectors
θ	model parameter vector
θ_0	true unknown parameter vector
$\ell(\theta)$	log-likelihood function of θ
ϵ	random noise function
$\hat{\theta}$	estimator of θ_0
$\hat{\mathcal{Y}}$	the set of estimated outputs
$\hat{\mathbf{y}}$	vector of predicted output values
\hat{f}	estimator function of g
\hat{y}	estimated output value
λ	regularization parameter
\mathcal{D}	data space
\mathcal{E}	objective/performance function
\mathcal{F}	feature set
\mathcal{H}	hypothesis set
\mathcal{I}	index set
\mathcal{N}	normal distribution
\mathcal{V}	set of cross validation folds
\mathcal{X}	set of explanatory variable vectors
\mathcal{Y}	set of output values
μ	expected value, average
$\Omega(\theta)$	penalty/constraint function of θ
ϕ_j	basis function

σ^2, α_j	variance, scale parameter
B	set of binary vectors
b	binary vector
c	geographical location vector
d	geographical data point
X	matrix of explanatory variable vectors
x	vector of explanatory variables
X_{ij}	$(i, j)^{\text{th}}$ element of matrix X
y	vector of output values
ε	random noise term
$b(G)$	bias with respect to distribution G
D	data set
D_c	geographical data set
D_e	test data set
D_t	training data set
D_v	validation data set
E	expected value
e	metric/Euclidean distance function
E_G	expected value with respect to distribution G
F	approximation distribution
f	approximation function
$f_r(x)$	probability density function of X_r
G	true unknown distribution

g	true unknown function
r_δ	dead zone radius
X	random variable
$x^{(n)}$	set of n realizations of X
X_r	random variable with distance r
x_{wp}	water permeability exponent
y	output value

Chapter 1

Introduction

1.1 Data analysis with open data

We live in an era of data flood with information being continuously stored by millions of devices world wide. These devices include e.g. mobile phones, cars, satellites, industrial machines, medical instruments and washing machines, with more and more of these devices forming Internet of things networks together. The data collected are used for a wide variety of purposes and applications. Some of these collected data sets have strict privacy legislations or rules behind them, which means access to this data is limited. Other data sets are categorized as *open data*, meaning this data has no restrictions on its usage. Open data is based on the idea that data should be freely available to everyone to use and republish as they wish, without restrictions from patents, copyright or other mechanisms of control. Due to its nature, open data can offer greater opportunities than private data, since it is available to larger groups of people. This allows broader utilization and evaluation of the data from many different disciplines and parties.

Since data is collected by millions of devices automatically worldwide, it is clear that the amount of this data is massive. It is estimated that around 2.5 quintillion (10^{30}) bytes of data is created every day. Furthermore, the data comes in a variety of different forms (e.g. photos, videos, databases) with different requirements on the processing speed (e.g. periodic, real-time). Data sets which fit into this so-called 3Vs (volume, variety, velocity) characterization are called *big data* (see e.g., Chen et al., 2014). It is obvious that automation, i.e. intelligent data processing by computers is needed to extract useful information from the data. *Machine learning* (ML) is a subfield of computer science focused on this, i.e. on the development and application of intelligent data processing systems. With big data available from many domains, the utilization of this data via ML approaches offers many interesting applications. In ML projects, data analysis experts usually

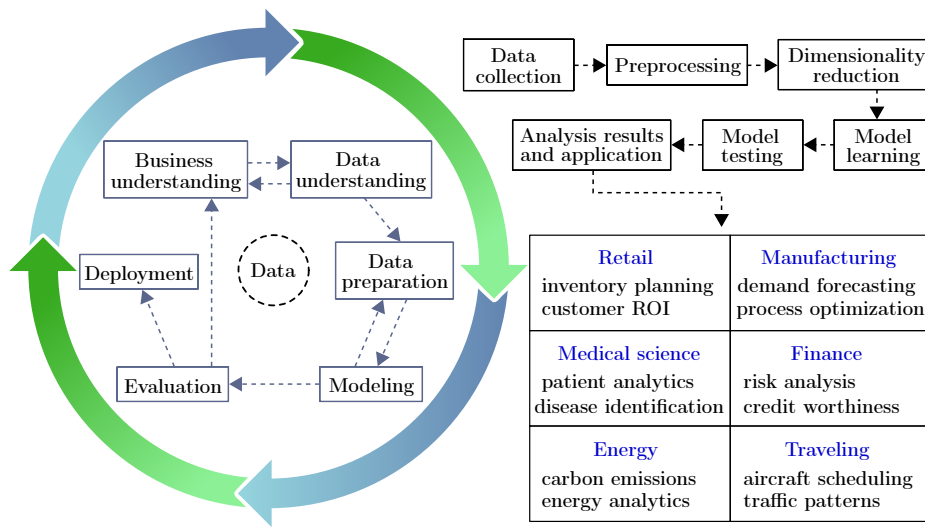


Figure 1.1: *Left*: the CRISP-DM process model. *Right*: the general flow of modeling in ML with example applications given.

follow the phases described by the *cross-industry standard process for data mining* (CRISP-DM, Shearer, 2000). The CRISP-DM model consists from the following six phases (see Figure 1.1):

1. *Business understanding*. In the first phase, the focus is on understanding the project objectives and requirements from a business perspective, and then converting this into a data mining problem definition.
2. *Data understanding*. In this phase, initial data collection and inspection are implemented to discover first insights into the data.
3. *Data preparation*. In data preparation phase the collected raw data is cleaned and transformed (e.g. dimensionality reduction) into the final data set suitable for the modeling tools.
4. *Modeling*. In here, various modeling techniques are selected and applied, and their parameter values are optimized.
5. *Evaluation*. At this stage, the model is evaluated carefully and checked if it achieves the business objectives. A key task is to determine if there are some important business issues which have not been sufficiently considered.
6. *Deployment*. In the final phase, the model is deployed into the business application. Monitoring and a maintenance plan are implemented, and project review is documented.

With increasing computational power and decreasing cost per byte of storage ML approaches are nowadays a hot topic both for companies and private

consumers. It is therefore clear that with the current development of technology and data-driven goals, data analysis is one of the key subjects to be studied now and in the future.

1.2 Motivation for the research

This thesis focuses on ML problems involving the prediction of forest soil characteristics, forest inventory attributes and the evaluation of spatial models, i.e. models which use geographically distributed data samples. The data in this research consists from a combination of various *geographic information system* (GIS) open natural data sets collected both manually and by *remote sensing* (RS) methods like satellites and airborne scanning techniques. Most of these natural resource data sets are collected for some specific purposes in their main applications, but this is not the subject of investigation in this thesis. In this work, we focus on investigating the potential additional value these open data sets can provide in other applications. Fairly recently, many governmental institutions have shown interest on data analytical solutions applying GIS data. In Finland for example, which has a big industry in forestry, harvest route planning in foresting operations is one of the key concerns. It has been discussed lately if GIS data could be used to improve the safety and efficiency in these operations. It is estimated that challenging trafficability conditions cause a yearly costs of 100 million euros for the industry in Finland alone (Tamminen, 1991), so even small improvements offered by ML approaches can result in substantial saves in operation costs.

Research motivation for this thesis originates from the use of open natural data in forestry. Finland is one of the world's largest producer of paper and cardboard and one the of largest producers of sawn timber in Europe. Around 20% of Finnish exports is produced by the forest industry which employs approximately 160,000 people in Finland. The Natural Resource Institute Finland (LUKE, formerly METLA), a key party in Finnish forest research, collects annually and biannually various GIS data sets to keep track on soil and forest conditions in Finland. LUKE also carries out statutory government work such as monitoring natural resources, certifying plant production, supporting natural resource policies and producing Finland's official food and natural resource statistics. One of the collected natural resource data sets is called the *multi-source national forest inventory* (MS-NFI, Tomppo et al., 2008). This data set consists from various forest state attributes such as tree volume, tree basal area, tree diameter and tree height per hectare. The MS-NFI is based on combining field measurements (i.e. MS-NFI sample plots) with satellite imagery and map data (Mäkisara et al., 2016). In addition to these data sources, the MS-NFI employs also e.g.

land-use maps and elevation models as other digital data sources (Zevenbergen and Thorne, 1987; Wood, 1996). By utilizing satellite images, forest characteristics can be estimated for geographical areas laying between relatively sparse network of MS-NFI samples. By then combining the sampled and estimated MS-NFI forest characteristics various statistics in the form of thematic maps can be produced for any given area. An example of an MS-NFI thematic forest map is the volume of growing stock in the whole of Finland.

Among other environmental data sets, the MS-NFI data is used in applications such as forestry decision making and strategic large-area forest planning. It is interesting to study how data sets like the MS-NFI could be used e.g. in operative planning of forest operations. In forest harvesting for example, the location of a harvest operation might be determined based on the number of trees and the trafficability of the harvesting routes in that area. The role of the MS-NFI data here would be to provide estimations (using ML) on tree quantities and the safety of the routes in the corresponding harvest area. Due to its possible uses like in the example presented, it is therefore important to study the performance of the MS-NFI data (and other data sets like it) in these tasks. This work aims to provide these performance measures by building and evaluating ML models using the open natural resource data sets provided by LUKE and other similar governmental institutions of Finland such as National Land Survey of Finland (NLS), Geological Survey of Finland (GTK), Finnish Meteorological Institute (FMI), and the Finnish Geospatial Research Institute (FGI).

1.3 Main objectives of the research

Regarding the matters considered above, the main objectives of this thesis are: to assess the prediction capability of the provided natural resource data sets, to identify the most relevant predictor features needed for forest attribute estimation, and to develop a model evaluation method for applications involving spatial data. These objectives are summarized in the following main research questions:

(RQ1): Are the provided open natural resource data sets applicable in predicting terrain conditions and forest attributes in Finland?

(RQ2): How to evaluate the prediction performance of a model involving spatially dependent data?

The research results obtained for question **(RQ1)** will provide empirical evidence which indicates if the current natural resource data collection and preprocessing procedures need to be modified or not before they can be

used in corresponding applications. The results will also help to identify the useful features in these data sets, which can decrease the computation time and data collection costs of the modeling processes in future applications. The results for question **(RQ2)** provide a method for evaluating spatial models by taking into account the geographical dependencies in the data, which is many times disregarded by classical model evaluation methods.

1.4 Organization of the thesis

This thesis consists of two separate parts. Part I of the thesis includes the chapters 1-4. Chapter 1 gives a general introduction to the subject and provides the motivations and research questions of this thesis. Chapter 2 presents an overview of the theoretical background and Chapter 3 gives a summary of the included research publications, results and the author's corresponding contributions. Conclusions and discussion are then presented in Chapter 4. Part II presents the six original research publications that were written during the doctoral studies at the university.

Chapter 2

Theoretical foundations

In this chapter, we will go through the background information related to the research conducted in the included publications. GIS and RS methods are introduced since the provided data is mostly collected using these techniques. Also, the nature of spatial data is considered and an introduction to the ML paradigm and standard model selection techniques is presented. A novel method developed during the research project is also introduced, which aims to tackle a problem that many classical model selection methods have with natural data sets.

2.1 Geographic information systems

Data today comes from many different sources. Multiple information systems collect data about weather, traffic, stock trade, logistics, consumer grocery behavior, agriculture, forests et cetera. Knowing all these activities is important but it is equally important to know where they happen. Geographic location is an important attribute of activities, policies and strategies. GIS is a special class of information systems which keeps track about events and things, and also where these events and things happen or exist (Longley et al., 2005). In each day millions of people utilize geographic information in their work and daily lives. For example we routinely check the weather forecasts from our neighborhoods in order to plan suitable schedules for our activities. GIS is also critically important for officials such as the police, hospitals and fire departments.

Geographical data collection is naturally subject to measurement errors due to plenty of distorting factors and can lead to inaccurate inferences while analyzing the data. In aerial imaging for example, wind causes measurement devices to momentarily change imaging angles and thus making some data samples slightly incompatible with each other. Other distortions to the data can be the effect of natural phenomena like for example clouds,

rain or shadows caused by trees. The measurement errors will always be present in any natural data sets we collect and must be used as such. It is impossible to make a perfect representation of the world, so uncertainty about it is inevitable in practice. Fortunately, many times the measurement errors in the data are negligible for the purpose of the corresponding GIS application and also many methods exist which help to remedy the distortions in the data. The extent of the measurement errors should however always be acknowledged as stated by Walter Lewin, a former professor of physics in Massachusetts Institute of Technology, "Any measurement that you make without the knowledge of its uncertainty is completely meaningless". By uncertainty here, Lewin means the magnitude of measurement errors.

2.2 Remote sensing methods

When working on a GIS project, the first issue or decision that we have to face is how to incorporate data into our analysis. Data collection is one of the most time consuming and laborious processes in GIS projects, making the decisions on how we should collect the data, one of the most important ones. It is crucially important to have a carefully planned and implemented data collection process for the sake of a successful GIS project. Nowadays, GIS data is collected via multitude of techniques such as satellites, aircraft and remote sensors. The measurements are carried out usually by a laser pulse and multi- or hyperspectral scanning and imaging methods. These methods usually include devices like image sensors, optical sensors, interferometers and spectrometers. In this thesis, we are mainly concerned with natural resource data collected via RS techniques, since the included publications use GIS data collected almost completely in RS manner (Pohjankukka et al., 2014a,b, 2016, 2017, 2018; Airola et al., 2018).

2.2.1 Sensors

A sensor is a device which detects events or changes in its environment and sends the information to other devices. Sensor devices can be divided into two main groups: passive sensors and active sensors. Passive sensors depend on external energy sources, like the Sun or Earth. These sensors gather data through the detection of energy such as light, heat or radiation. Examples of passive sensors are the photographic camera and a thermometer. Active sensors have their own energy source and can therefore be controlled more than passive sensors. They work by emitting a controlled beam of energy to a surface and measure the amount of energy reflected back to the sensor. Active sensors include methods like RADAR (radio detection and ranging) and LiDAR (light detection and ranging). In RS applications both passive and active sensors are used and they usually measure energy in different

intervals of the electromagnetic (EM) spectrum. These sensors measuring EM energy can further be divided into optical and non-optical sensors.

Optical sensors

Optical sensors measure EM energy in the wavelength range of approximately 400-700 nanometers (nm), which is the interval of the EM spectrum called light. An example of an optical sensor is the LiDAR which is an active RS sensor utilizing pulsed laser to measure distances to Earth (Weitkamp, 2005). LiDAR uses visible-light, ultraviolet and near infrared spectra to perform the imaging. The laser pulses can be used to generate a three-dimensional point cloud representation about the shape of the measurement object. Airborne LiDAR measurements are collected with airplanes or helicopters by emitting a laser pulse to the target surface. The echo of this surface reflected laser pulse is then saved by a scanner instrument in the airplane. Currently there exists a wide variety of LiDAR applications. For example, autonomous vehicles use LiDAR to create a 3D-model of their surroundings for navigation purposes and quadcopter drones use LiDAR to identify specific cereal species in crop fields.

Non-optical sensors

Non-optical sensors measure EM energy in the range outside the wavelength range of light. These sensors measure energy of e.g. microwaves, gamma waves, ultraviolet waves, infrared waves and radio waves. Notice that since LiDAR also measures ultraviolet and near infrared spectra it can be considered as both an optical and a non-optical sensor. Other examples which utilize non-optical sensors are gamma-ray spectrometry and RADAR. The gamma-ray spectrometry involves measuring the amount of very short wavelength (picometers) gamma rays emitted by the upper soil or rock layers due to radioactive decay (Bakker et al., 2009). Gamma rays are measured mainly in mineral explorations because the measured energy of specific wavelengths provide information on the abundance of specific minerals. Gamma-ray spectrometry must be measured close to the Earth's surface (within a few hundred meters) because of large atmospheric absorption of these waves. RADAR sensors are one of the most commonly used active microwave sensors originally developed and used by the military. Nowadays, they are also widely used in civil applications. RADAR applications include e.g. environmental monitoring, aviation, marine navigation and meteorology. Classified examples of devices and techniques using active/passive and optical/non-optical sensors are shown in Table 2.1. Note that these classifications are not completely fixed.

	Active	Passive
Optical	LiDAR structured light	Photographic camera Video camera Multispectral scanner
Non-optical	RADAR altimeter Imaging RADAR Laser scanner	Gamma-ray spectroscopy Thermal scanner Imaging spectrometer Radiometer

Table 2.1: Example devices and techniques utilizing passive/active and optical/non-optical sensors (see e.g., Bakker et al., 2009).

2.2.2 Data collection techniques

Many different techniques exist today for data collection purposes. Sensors can be attached e.g. to automobiles, aircraft, ships and satellites for multiple different purposes. For example automobiles can collect data about air pollution by making real time measurements of nitrogen dioxide levels. The aircraft meteorological data relay (AMDAR, 2018) is an example of a program initiated by the World Meteorological Organization (WMO) in which meteorological data is collected worldwide by using commercial aircraft. One of the common modern ways to collect data from the surface of the Earth is to use a satellite. Satellites contain various measurement instruments like for example multispectral scanners, which measure the reflected EM energy from Earth’s surface resulting as digitalized pixel image data (Bakker et al., 2009). The Landsat (originally Earth Resources Technology Satellite ETRS) program is a series of satellite imaging missions started during the 1970s. The missions are operated by NASA and the U.S. Geological Survey. The program so far has consisted from eight satellites Landsat 1-8. In the course of these missions millions of high resolution images have been acquired by diverse set of instruments aboard the Landsat satellites (Landsat, 2018). The collected data contain beneficial information regarding agriculture, geology, forestry, resource detection, state of oceans, regional planning et cetera. In Figure 2.1 is presented the Landsat 8 satellite and an example image produced by the corresponding satellite. Another example of satellite data collection is the Sentinel-1 mission which comprises from two polar-orbiting satellites (Sentinel-1 team, 2013). Its mission is the joint initiative of the European Commission and the European Space Agency or ESA. It is based on data received from Earth observing satellites and ground-based information, and it provides short revisit times, dual polarization capabil-

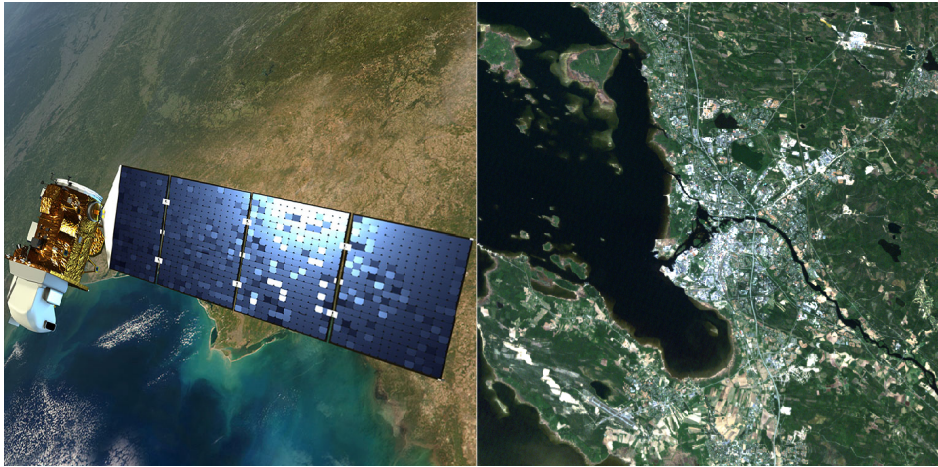


Figure 2.1: *Left:* Landsat 8 satellite. *Right:* Landsat image of Oulu, Finland. Image copyright © NASA and U.S. Geological Survey.

ity and rapid product delivery. Due to the synthetic aperture radar (SAR) technology, the Sentinel-1 satellites can acquire imagery regardless of the weather. SAR has the advantage of operating at wavelengths not impeded by lack of illumination or cloud cover and can acquire data over a site under all weather conditions during day or night time. In Table 2.2 is listed other examples of satellites with different specifications.

Satellite	Resolution	Bands	λ region (μm)	Application
IKONOS	0.82m	5	0.45-0.9	mining and exploration
TerraSAR-X	1m	1	31066	infrastructure planning
SPOT-7	1.5m	4	0.45-0.89	deforestation
RADARSAT	8m	1	56564	oil and gas
Sentinel-2A	10m	13	0.44-2.19	agriculture
LANDSAT 8	15m	11	0.43-12.51	vegetation analysis

Table 2.2: Examples of satellites providing RS natural resource data (see e.g., Satellite imaging corporation, 2018; Earth observation portal, 2018; European space agency, earth portal, 2018). Resolution refers to the maximum resolution of the corresponding satellite. Number of used EM frequency bands, wavelength ranges (λ) and example applications for the satellites are also shown.

2.3 Data representation

In order to analyze real world phenomenas we first need to store information about it in some format. This information is ordinarily stored either in raster or ascii-file format in digital computers. A raster data set corresponds to a set of images (such as GeoTIFF) representing e.g. satellite data. An image file contains a discretized representation of a target area with each pixel value corresponding to an attribute of interest like topographical height for example. The other format, i.e. an ascii-file, is normally a text- or a csv-file containing the target variable and geographical location data. Data resolution size plays an important role when digitalizing real world phenomenas to a set of discrete data points. Any geographic area contains potentially an infinite amount of information so finding the suitable resolution size for the corresponding application should be carefully inspected. The Figure 2.2 depicts an example raster data image of a topographic height image map. Darker values in the image correspond to areas with lower topographical height and bright areas to higher topographical height respectively.

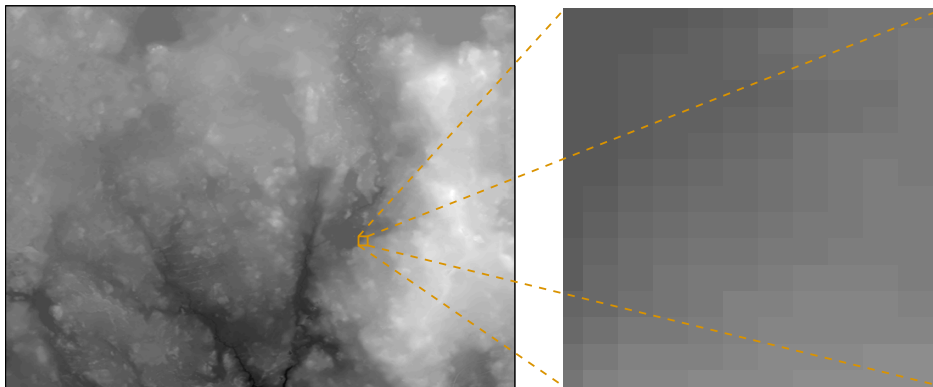


Figure 2.2: Raster topographical height data image from Parkano, Finland. The potentially infinite amount of geographical information is discretized into a finite set of pixel units.

2.4 Open data sets

This section presents some of the open natural resource data sets used in the included publications. The data were mainly collected using RS techniques such as satellite and airborne imaging methods. Note that not all of the available RS data is provided in raw format, but some of the data sets are the products (e.g. map data) of measured raw data. The data formats are mostly GeoTIFF-images with varying resolutions. Examples of these data sets are presented in Table 2.3. Digital elevation model (DEM) consists

Data	Provider	Publication
Digital elevation model	NLS	I-IV
Gamma-ray spectroscopy	GTK	I-IV
Airborne EM	GTK	II, IV
Peatland	LUKE	II-IV
Weather information	FMI	III-IV
Aerial imagery	NLS	V
MS-NFI	LUKE	II-V

Table 2.3: Examples of open natural resource data sets used in the publications included into this thesis. Data provider is given in the center column and the last column shows in which publication the data was used.

from a numeric representation of the Earth’s surface which contains height points representing the topography, and a method for calculating elevations between the height points. DEM data is usually stored as a regular grid or a triangulated irregular network (Wood, 1996). Gamma-ray spectroscopy data is based on gamma-ray flux from potassium, which is the decay process of the naturally occurring potassium element. This data indicates many characteristics of the soil, such as the tendency to frost heaving and tendency to stay moist after precipitation (Hyvönen et al., 2003). Factors like soil type, porosity, density, humidity and grain size affect the amount of gamma-ray radiation. Soil with high gamma-radiation tends to have lower moisture than soil with low gamma-radiation. The peatland data is a binary raster image with 0/1 values corresponding to non-peatland/peatland areas. It is compiled using open geographic information data derived from NLS topographic database (NLS, 2014). Airborne EM (AEM) data is collected by transmitting an EM signal from a sensor attached to an aircraft. Depending on the system used and the surface conditions, AEM techniques can detect changes in the conductivity of soil to depths of hundreds of meters. The conductivity response in the ground is commonly caused by the presence of e.g. graphite, salt or clays which are electrically conductive materials. Weather data provided by the FMI includes temperature and rainfall information. Aerial imagery contains RGB and color-infrared images acquired with digital camera sensors. The biannually updated MS-NFI data (introduced in Section 1.2) holds the state of Finnish forests in high spatial resolution. For more information on the corresponding data sets see the included publications (Pohjankukka et al., 2014a,b, 2016, 2017, 2018; Airola et al., 2018).

2.5 Spatial data analysis

A set of spatially distributed data points in geographical space forms a spatial data set. These data points contain information about the characteristics of the corresponding geographical locations. Data analysis involved with data sets like this is called *spatial data analysis* (see e.g., Cressie, 2015). Plenitude of real world data sets contain geographical locations. Take for example a data set about tree attributes. This data set could involve attributes such as tree height, diameter, volume et cetera, but it also includes the locations of the trees. Another example is provided by the global positioning system (GPS). The GPS system keeps track of the locations of millions of vehicles around the world. All navigation systems use some sort of positioning data and are therefore involved with spatial data analysis. Geographical data samples naturally contain dependencies with each other as a function of distance between them. The first law of geography and fundamental principle in geostatistical analysis, according to Waldo Tobler (Tobler, 1970), states that: "Everything is related to everything else, but near things are more related than distant things". This property of closer things being more similar to one another is called *spatial autocorrelation* (SAC). Because of SAC, special caution has to be taken in many statistical methodologies, which often rely on the assumption of independent and identically distributed (i.i.d.) data samples. When dealing with spatial data sets this assumption can result in optimistically biased approximations. In the next section we will give a formal definition for SAC using the well-known *autocorrelation function* (see e.g., Shumway and Stoffer, 2005).

2.5.1 Spatial autocorrelation

Let X_r denote a random variable with a corresponding probability density function $f_r(x)$ where x is a realization of X_r . The index r denotes either time lag (Shumway and Stoffer, 2005) or distance. In this thesis, we are dealing with spatial data sets and therefore $r \in \mathbb{R}^+$. For defining the autocorrelation function we need the definitions for the *mean* and *autocovariance functions*. The mean function μ_r for random variable X_r is defined as the expected value:

$$\mu_r = E(X_r) = \int_{-\infty}^{\infty} x f_r(x) dx. \quad (2.1)$$

The physical interpretation of μ_r is the average value of realizations of the random variable X_r , which are located r distance units away from some spatial reference point (see left side of Figure 2.3). The reference point corresponds to the black point in the center of the circle. In practice, we usually never have data points which are exactly r distance units away from the center point, and therefore some tolerance level Δr must be used. That

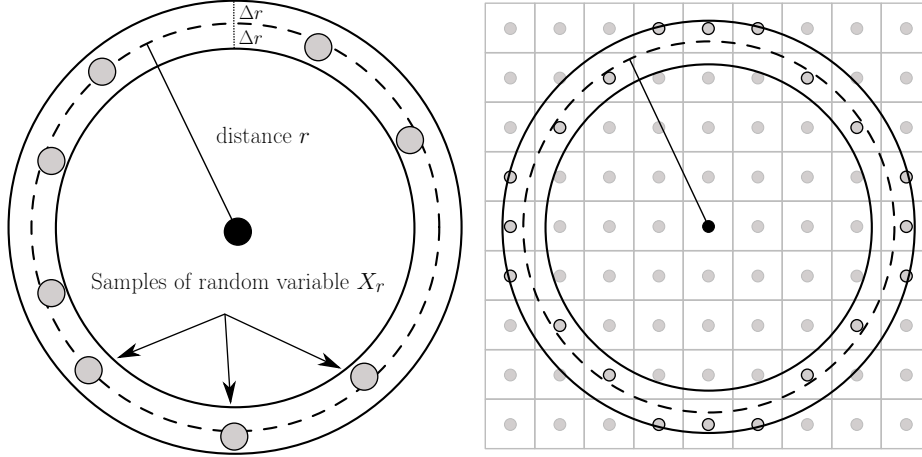


Figure 2.3: *Left*: illustration on how μ_r would be estimated with real data. The realizations of X_r are inside the surrounding shell with thickness $2\Delta_r$. *Right*: illustration of how $\rho(r)$ is calculated with raster data when the realizations of X_r correspond to the center values of the pixels which are inside the shell.

is, we must allow the distance to the surrounding data points to deviate slightly from r by some Δr amount. In the figure, this corresponds to having a circular shell of thickness $2\Delta r$ which contains the realizations of X_r . The autocovariance function $\gamma(r, s)$ for random variables X_r and X_s (with indexes r and s correspondingly) is defined as the second moment product:

$$\gamma(r, s) = \text{Cov}(X_r, X_s) = E[(X_r - \mu_r)(X_s - \mu_s)], \quad (2.2)$$

where $\text{Cov}(X_r, X_s)$ stands for the covariance of random variables X_r and X_s . By normalizing the autocovariance function we get the autocorrelation function, which is defined as:

$$\rho(r, s) = \frac{\gamma(r, s)}{\sqrt{\gamma(r, r)\gamma(s, s)}} = \frac{\gamma(r, s)}{\sqrt{\sigma_r^2\sigma_s^2}}, \quad (2.3)$$

where σ_r^2 and σ_s^2 are the variances of random variables X_r and X_s . To measure the autocorrelation corresponding to the two dimensional setting in Figure 2.3 we could fix $s = 0$, so the center of the circle would correspond to random variable X_0 . The autocorrelation value is now a function of only the distance r away from the center: $\rho(r, 0) = \rho(r)$. In practice with GIS raster data, the calculation of $\rho(r)$ would be implemented as illustrated in the right side of Figure 2.3. Next, we will go through other examples on how the degree of SAC in data can be estimated.

2.5.2 Measures of spatial autocorrelation

Variogram

The variogram (see e.g., Cressie, 2015) is a function describing the degree of spatial dependence of a stochastic process. Let $s \in \mathcal{S}$ denote a location point, and $X(s)$ a random variable X of a stochastic process at location s . Then the *variogram function* between locations $s_1, s_2 \in \mathcal{S}$ is defined as:

$$\begin{aligned} 2v(s_1, s_2) &= E \left[((X(s_1) - \mu(s_1)) - (X(s_2) - \mu(s_2)))^2 \right] \\ &= \text{Var}(X(s_1) - X(s_2)), \end{aligned} \quad (2.4)$$

where $\mu(s_i)$ denotes the expected value of random variable $X(s_i)$. The function $v(s_1, s_2)$ itself is called the *semivariogram function*. If the stochastic process of the random variable X is both isotropic (uniform in orientation) and stationary (independent of time or location shift), then the variogram can be represented as a function of only the distance h between s_1 and s_2 . In this case we have $v(s_1, s_2) = v(h)$, where $h = e(s_1, s_2)$ and e is some metric function. The distance h is usually called the lag term as in time series analysis (Shumway and Stoffer, 2005). In practice, our data set consists from a set of observed sample points $\{x(s_1), x(s_2), \dots, x(s_n)\}$ of the corresponding random variables. In this situation we estimate the variogram $2v(h)$ with the *empirical variogram* $2\hat{v}(h)$, which is calculated by:

$$2\hat{v}(h) \equiv \frac{1}{|N(h)|} \sum_{N(h)} (x(s_i) - x(s_j))^2, \quad (2.5)$$

where $N(h) \equiv \{(i, j) : e(s_i, s_j) = h \wedge i \leq j\}$ and $|N(h)|$ is the cardinality of $N(h)$. That is, $N(h)$ is a set that contains all the index pairs (i, j) of data points that have distance h between them. The condition $i \leq j$ makes sure that each pair (i, j) is not included twice into the summation in Equation 2.5 because data point pairs corresponding to index pairs (i, j) and (j, i) are equal.

Moran index

Moran index (or Moran's I, Longley et al., 2005) is a method for measuring the degree of spatial similarity in locational data. It tests to see if spatial phenomena are clustered or are randomly spread throughout space. A central component in Moran's I is the weight matrix $\mathbb{W} \in \{0, 1\}^{n \times n}$, where n denotes the number of data points. When $i \neq j$ the entries of \mathbb{W} are $w_{ij} = 1$ if data at locations i and j are similar, and $w_{ij} = 0$ otherwise. For all entries where $i = j$ we set $w_{ij} = 0$. The similarity between data points can be

specified depending on the phenomenon in question. The similarity index value itself, denoted by I , is calculated by the formula:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{w \sum_{k=1}^n (y_k - \bar{y})^2}, \quad (2.6)$$

where y_i stands for attribute of interest at location i , $w = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$ and \bar{y} is the average value of the y_i s. When $I > 0$ nearby data points tend to be similar in attributes. When on the other hand $I < 0$ they tend to be more dissimilar, and when $I = 0$ attribute values y_i are arranged randomly and independently in space. Moran's I is one of the most popular measures of SAC and is widely used in the field of geography and in applications involving GIS data.

2.6 Machine learning

We now go through an introduction to the general ML paradigm, which is applied in all the included publications. ML is a subfield of computer science dedicated to the research and development of methods for recognizing and learning dependency relationships in data. It is about making computers to modify and adapt their actions to reflect the correct output to given input data. The ML field contains a wide range of methodologies for achieving the learning goals with its fundamental roots originating from statistics, information theory, neuroscience, physics and mathematical optimization. ML bears many similarities with statistics and the two can be considered to be almost exactly equal, since they both aim for the same goal: to learn from data. They differ however on the things they emphasize. Statistics is more about formal statistical inference like constructing confidence intervals, hypothesis tests et cetera, whereas ML is more focused on making accurate predictions and is less strict on testing assumptions (see e.g., Breiman, 2001b; Shmueli, 2010; James et al., 2014). It must also be added that the statistical methods used in ML is not limited only to frequentist approaches, but applies also many methods of the Bayesian framework (see e.g. Bishop, 1996). Furthermore, the models used in ML and statistics can be divided into two classes: *parametric* and *nonparametric* models. A parametric model contains a fixed number of parameters which are to be tuned so that the model fits the data well. A nonparametric model does not have a fixed number of parameters but they vary depending on the available data. Nonparametric models can be considered as parametric models with potentially infinite number of parameters (see e.g., Sheskin, 2007).

2.6.1 The learning setup

The basic setup in any ML process involves more or less the following components: observed data set $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$, hypothesis set \mathcal{H} , objective function \mathcal{E} and a learning algorithm \mathcal{A} . The vectors $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ correspond to explanatory variables and the values $y_i \in \mathcal{Y} \subset \mathbb{R}$ correspond to the outputs. The set of outputs \mathcal{Y} can also be missing in some cases (e.g. in clustering problems). The hypothesis set is defined as $\mathcal{H} \subset \{f \mid f : \mathcal{X} \rightarrow \mathcal{Y}\}$, i.e. it is a subset of functions (or models) f which map $\mathbf{x} \mapsto f(\mathbf{x}) = y$. The learning algorithm \mathcal{A} is defined as the mapping $\mathcal{A} : \mathcal{D} \rightarrow \mathcal{H}$ ($D \mapsto \hat{f}$), where \mathcal{D} is the set of all possible data sets (or data space, i.e. $D \subset \mathcal{D} = \mathcal{X} \times \mathcal{Y}$) and \hat{f} is selected by \mathcal{A} , based on D . This general learning process of ML is illustrated in Figure 2.4. Examples of \mathcal{A} are e.g. the k-nearest neighbor (kNN) algorithm, backpropagation algorithm in artificial neural networks (ANN, Marsland, 2014) or Markov chain Monte Carlo (MCMC) sampling in Bayesian modeling.

In many cases, the goal of a ML process is to find a model $f \in \mathcal{H}$ such that the objective function \mathcal{E} is minimized, i.e. we select the model \hat{f} such that:

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \mathcal{E}(f, D). \quad (2.7)$$

In the case of a simple linear regression, the model takes the form $f(\mathbf{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_i) = \sum_{j=1}^d \theta_j \phi_j(\mathbf{x}_i)$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$ is a parameter vector defining the linear model (i.e. linear with respect to the parameters) and each ϕ_j is some function of \mathbf{x}_i . In linear regression, we take the optimal model $\hat{f} = f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$ to be such that $\hat{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is matrix with $(i, j)^{\text{th}}$ element being $\mathbf{X}_{ij} = \phi_j(\mathbf{x}_i)$ and $\mathbf{y} \in \mathbb{R}^n$ is a column vector of the output values. The simplest case for the functions ϕ_j is to set $\phi_j(\mathbf{x}_i) = x_j$,

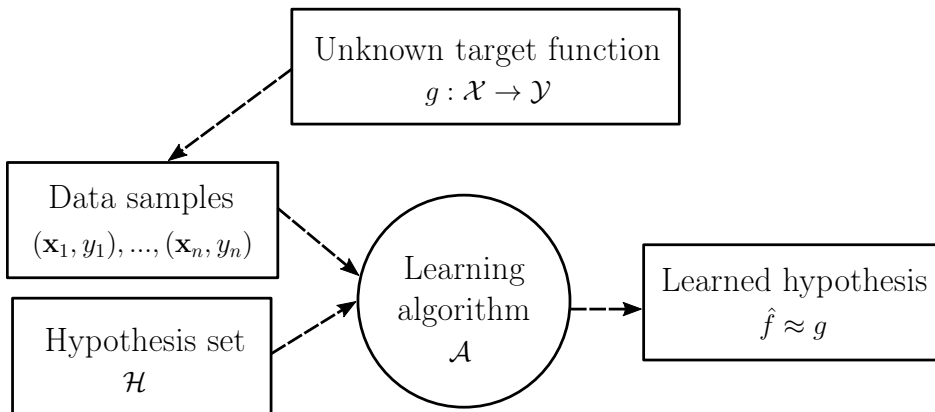


Figure 2.4: Illustration of a general learning process in ML.

where x_j is the j th element of \mathbf{x}_i . The optimal parameter vector $\hat{\boldsymbol{\theta}}$ here is called the *least squares* (LS) estimator, which minimizes the sum of *squared errors* (SE) formula:

$$\text{SE} = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2. \quad (2.8)$$

If the sum in Equation 2.8 is further multiplied with the reciprocal n^{-1} then the equation is called the *mean squared error* (MSE) formula.

Another option for implementing learning in ML is to use probabilistic approaches such as Bayesian methods, which are based on the famous Bayes' theorem:

$$p(\boldsymbol{\theta} | D) = \frac{p(D | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(D)} \propto p(D | \boldsymbol{\theta}) p(\boldsymbol{\theta}), \quad (2.9)$$

where the constant normalizing factor $p(D)$ can be disregarded. The $p(D | \boldsymbol{\theta})$ is the likelihood of the data and $p(\boldsymbol{\theta})$ is the prior distribution of the parameters. In this approach, we are interested in finding a parameter vector $\boldsymbol{\theta}$ which maximizes the posterior distribution:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} p(\boldsymbol{\theta} | D). \quad (2.10)$$

Parameter vector $\hat{\boldsymbol{\theta}}$ that satisfies Equation 2.10 is called the *maximum a posteriori* (MAP) estimator. The advantage of the probabilistic methods are that they allow the computation of statistics like posterior intervals, expected values, posterior predictive variance et cetera. Other difference between non-probabilistic and probabilistic methods is on how predictions can be made. In the non-probabilistic approach, predictions \hat{y} for output value y are made by inputting a new vector of explanatory variables \mathbf{x} to the learned model, i.e. $\hat{y} = f(\mathbf{x}, \hat{\boldsymbol{\theta}})$. In the probabilistic approach, predictions can be calculated as $\hat{y} = E[y | \mathbf{x}, D]$ using the posterior predictive distribution:

$$p(y | \mathbf{x}, D) = \int p(y | \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} | D) d\boldsymbol{\theta}. \quad (2.11)$$

That is, \hat{y} is now calculated as the expected value of y with respect to distribution $p(y | \mathbf{x}, D)$. The prior distribution $p(\boldsymbol{\theta})$ in Equation 2.9 is subject to fundamental debate in statistics between frequentist and Bayesian perspectives. With large data sets however the prior does not play a big role since it essentially factors out in this case working simply as a catalyst, and the likelihood $p(D | \boldsymbol{\theta})$ determines the model completely (Abu-Mostafa et al., 2012). There are also possibilities to select non-informative priors, which are designed to have no or minimal effect to the posterior distribution (Gelman et al., 2013). In addition, the process where the model parameters $\hat{\boldsymbol{\theta}}$ are chosen so that:

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} p(D | \boldsymbol{\theta}), \quad (2.12)$$

then $\hat{\boldsymbol{\theta}}$ is known as the *maximum likelihood estimator* and the process is known as *maximum likelihood estimation* (MLE). It can be easily shown that the process of minimizing SE is equivalent to MLE under Gaussian noise distribution assumption (see e.g., Konishi and Kitagawa, 2007). When the prior does not affect the selection of $\boldsymbol{\theta}$ then MLE and MAP produce identical results. Useful properties of the MLE estimator $\hat{\boldsymbol{\theta}}$ are that: $\Pr(\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0) \rightarrow 1$ and $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightarrow \mathcal{N}(0, \mathfrak{J}(\boldsymbol{\theta}_0)^{-1})$ as $n \rightarrow \infty$, where $\boldsymbol{\theta}_0$ is true unknown parameter vector and $\mathfrak{J}(\boldsymbol{\theta}_0)$ is the *Fisher information* matrix at $\boldsymbol{\theta}_0$. These properties of MLE are known as *asymptotic consistency and normality*. Moreover, the MLE estimator is *efficient* (has asymptotically smallest variance) and is *invariant* under functional transformation, i.e. if $\hat{\boldsymbol{\theta}}$ is the MLE estimator of $\boldsymbol{\theta}_0$, then $z(\hat{\boldsymbol{\theta}})$ is the MLE estimator of $z(\boldsymbol{\theta}_0)$, where $z = z(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ (Fisher, 1922, 1925; Lehmann and Casella, 1998; Mukhopadhyay, 2000). Note that these asymptotic properties of the MLE are valid only in methods which have a probabilistic interpretation.

Considering that the topic of this thesis is related to forest attribute estimation via ML, it is worth mentioning that many studies have been conducted about applying Bayesian nonparametric modeling (Bayesian models with an infinite-dimensional parameter space) to forestry applications. Bayesian nonparametric modeling includes many effective methods such as mixture models, latent feature models, Gaussian processes, Dirichlet processes, hidden Markov models (HMM) and hierarchical models (see e.g. Gelman et al., 2013; Bishop, 2006, 1996). One can find examples of such forestry applications from works such as in a summary volume (Maltamo et al., 2014). The corresponding case studies include applications such as segmentation of forest to tree objects, estimation of biomass components, tree species recognition, tree diameter distribution estimation, estimation of canopy cover, forest fuel assessment and fire prevention, biodiversity assertion et cetera.

2.6.2 Types of learning

The ML methods can be categorized into classes based on what kind of learning is implemented. They are most commonly categorized by whether they belong into *supervised learning* or *unsupervised learning* methods. There are also hybrids of the two mentioned like *semi-supervised learning* and *reinforcement learning* methods, which have characteristics from both supervised and unsupervised learning classes.

Supervised learning

When we are dealing with labeled data, i.e. data which contains both the explanatory input data set \mathcal{X} and the output data set \mathcal{Y} (i.e. labels), we

are doing supervised learning. In a supervised learning problem we are always given the ground truth values y_i for output variables, which can be compared with the outputs $\hat{y}_i = \hat{f}(\mathbf{x}_i)$ of the estimator model \hat{f} . A learning method belongs to a class of supervised learning methods when it involves comparing the observed output values y_i with the model's estimates \hat{y}_i . Usually an objective function like the SE in Equation 2.8 is used in this comparison. When the output values y_i belong to a set $\mathcal{Y} \subset \mathbb{Z}$ we are usually dealing with a *classification problem*, i.e. a ML system will learn to classify data. In cases where y_i s belong to set $\mathcal{Y} \subset \mathbb{R}$ we are dealing with a *regression problem*. An example of supervised classification learning problem is an object recognition system. Consider a data set which contains images of human and non-human faces. The labels of this data set are a set of boolean 1/0 (true/false) values indicating whether the image represents a human face or not. The ML system will be trained to recognize the characteristics of a human face and to link these characteristics with a positive (true) output value. An example of supervised regression problem is the prediction of the future value of a stock in NASDAQ.

Unsupervised learning

Whereas in supervised learning we had the set of output values \mathcal{Y} available, in unsupervised learning we do not have this set. Unsupervised learning is implemented without feedback using set \mathcal{Y} on whether we have made a good estimation or not. In this kind of learning the focus is mostly concerned with finding similarities between the input data points in \mathcal{X} so that inputs that have something in common are categorized together. An example of unsupervised learning is data clustering using e.g. k-means algorithm (Theodoridis and Koutroumbas, 2008) or statistical density estimation. Consider a web-based system providing movie streaming services for consumers like Netflix. Websites like these collect viewer data in order to categorize each user's movie taste. Based on these categorizations the website can offer new movies to the viewers similar to their liking. This will increase the probability that the user will not unsubscribe the websites services, and hence can increase profits.

Semi-supervised and reinforcement learning

In semi-supervised learning problem we have only some (small) subset $\mathcal{U} \subset \mathcal{Y}$ of the output values available and modeling is implemented by using both labeled and unlabeled data. In reinforcement learning, the set of output values \mathcal{Y} is not available at all but some scoring function is implemented to guide the learning process. An example of reinforcement learning is playing a game like chess, where one learns by simply playing the game.

2.7 Model complexity selection

In the previous sections we discussed about modeling with ML. We pointed out, that many of the learning tasks can be represented as an optimization problem in the form of Equation 2.7. It therefore seems reasonable to suggest, that we should always select such a model f which minimizes the right side of this equation. It turns out however that even though this argument makes sense mathematically, it is not a good approach in most real world problems. The reason for this is that the data we use for analysis contains almost certainly an added noise. This is simply due to the fact that absolute precision is impossible to achieve in most cases of physical measuring. Explicitly stated, this means that if our variable of interest y has a true dependency relation $y = g(\mathbf{x})$, where g is some unknown function, then the observed value always has the form $y = g(\mathbf{x}) + \varepsilon$. Here the term ε is called a *noise term*, which by definition is a variable that can not be learned. Therefore, it is important for the goal of the learning process to have $\hat{f} \approx g$ and not $\hat{f} \approx g + \varepsilon$, where ε is a random function generating the noise term ε .

It must also be mentioned that even though noise causes problems in modeling, it is not necessarily the sole factor. Problems can also be caused by e.g. a too complex hypothesis set \mathcal{H} or a small data set D . If \mathcal{H} contains a large number of equally likely models that solve Equation 2.7, then it is unlikely that generalization can be achieved. This is due to the fact that when we have a huge set of equally good models to choose from, then the chances of selecting the correct one is small. Another problem can arise if D is very small and the true model g is a highly complex function. In this case, in order to have any success in generalization, \hat{f} should be selected from a simpler hypothesis set than the set g belongs to. The drawback here of course is that now the learned model \hat{f} can not approximate g well. The simplicity of a hypothesis set \mathcal{H} can be measured with e.g. a concept known as the *Vapnik-Chervonenkis* (VC, Vapnik, 1998) dimension. The VC dimension is a measure of the capacity (complexity, flexibility) of \mathcal{H} that can be learned by a statistical classification algorithm. It can be shown (see e.g. Hoeffding, 1963) that if \mathcal{H} has a finite VC dimension, then generalization with \mathcal{H} is possible.

In the following subsections, we will have closer discussions about the relationship between model fitting and generalization capability, and we will also consider some of the most widely used methods for model complexity selection in ML. Furthermore, at the end of this section we will present a new model evaluation and selection method proposed in the included publication (Pohjankukka et al., 2017), which takes into account the effects of SAC in the data. As we noted earlier, SAC is an intrinsic property in many natural data sets, so model evaluation and selection using the standard methods is not always suitable in all situations.

2.7.1 Overfitting

The situation where $\hat{f} \approx g + \epsilon$ is caused by *overfitting* the model to the observed data D . What this means is that we have trained the model f too much, i.e. we have forced the model to learn non-existing patterns in the data created by the noise term ϵ . In other terms, when we overfit a model, we produce an analysis that corresponds too closely (or even exactly) to a particular data set. This model may therefore fail to fit new observed data well, making the prediction of future observations unreliable. A common characteristic of an overfitted model is that it contains more parameters than what can be justified with the available data set. Overfitting is tempting to occur, since a highly complex model will fit the observed data very well and produces a small error in Equation 2.7. However, because the whole goal of statistical inference is on how to deal with unobserved data, i.e. how to generalize to data we have not yet seen, it is not sufficient to just consider minimizing the objective function. In Figure 2.5 is an illustration of a model fitting done right and a heavily overfitted model with a 20th order polynomial. The data points have been generated from a linear function with an added uniform random noise. The linear fit will generalize much better than the higher order fit in this example. Even though the higher order model fits the data exactly, it will not generalize as well as the simpler model since it has learned an additional random phenomenon ϵ , which can not be learned. This illustration is also an example of a principle known as *Occam's razor*, which states that the simpler explanation is usually better (see e.g., Vapnik, 1998). Dealing with overfitting is one of the key tasks in data analysis, and the ability to handle it is what separates professionals from amateurs (Abu-Mostafa et al., 2012).

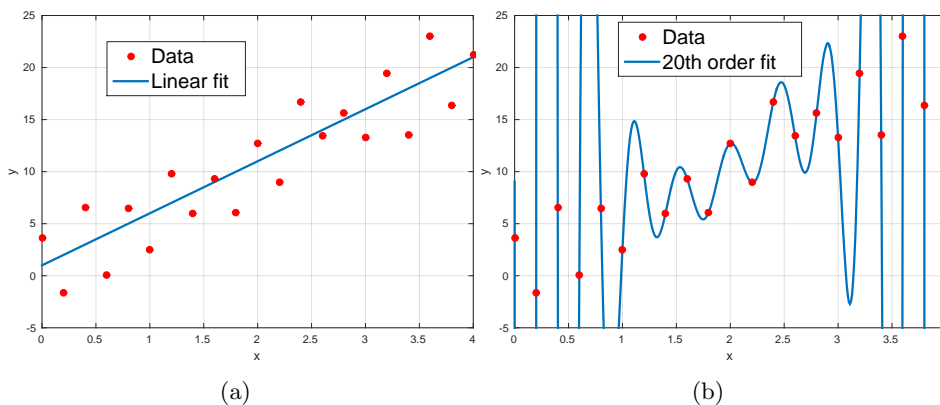


Figure 2.5: (a) A linear fit to data. (b) 20th order polynomial fit.

2.7.2 Bias-variance trade-off

In Figure 2.6 is shown the connection between complexity of the estimator model \hat{f} and its generalization error. In the beginning, the model is too simple (e.g. a constant) to capture the pattern in the data. Situation like this is called *underfitting*. This can be caused either by poor model fitting or a lack of sufficient data. In the other end, when we have fitted the model too well to the data, overfitting occurs. The optimal fit to the data lies somewhere between the two extremes, when the model is not too simple and not too complex. When a model is too simple, it is said to be biased by a priori knowledge with small variance. As we increase the model complexity to get a better fit to data we increase the model variance. Straight line for example has a large bias and small variance, whereas a complicated polynomial has a small bias but large variance. In other words, when we have a less complex hypothesis set \mathcal{H} we have better chances of generalization, whereas when we have a more complex hypothesis set \mathcal{H} we have better chances of approximating the true model g . It turns out that there is a trade-off between the generalization and approximation performance for the model \hat{f} , which is known as the *bias-variance trade-off* (see e.g., Marsland, 2014). To make this more explicit, we will first define the average estimator model \bar{f} as:

$$\bar{f}(\mathbf{x}) = E_D \left[\hat{f}^{(D)}(\mathbf{x}) \right]. \quad (2.13)$$

That is, $\bar{f}(\mathbf{x})$ is the average estimator model over all possible data sets D . The notation $\hat{f}^{(D)}$ here denotes the dependency of the model \hat{f} to the data D , since \hat{f} is learned using the data set D . Using the above definition it can

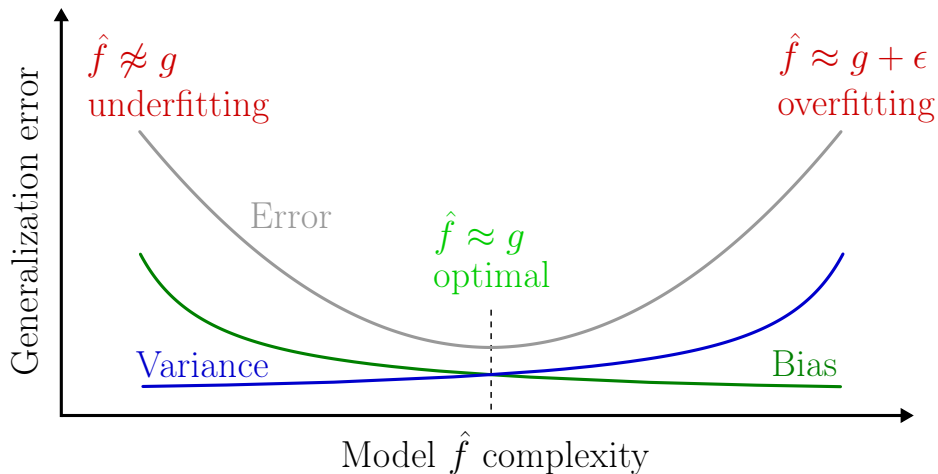


Figure 2.6: Illustration of the connection between model complexity and generalization performance.

be shown (Abu-Mostafa et al., 2012) that:

$$\begin{aligned}
E_D [E_{\mathbf{x}}(\text{SE}(\mathbf{x}))] &= E_D \left[E_{\mathbf{x}} \left[\left(\hat{f}^{(D)}(\mathbf{x}) - g(\mathbf{x}) \right)^2 \right] \right] & (2.14) \\
&= E_{\mathbf{x}} \left[E_D \left[\left(\hat{f}^{(D)}(\mathbf{x}) - g(\mathbf{x}) \right)^2 \right] \right] \\
&= E_{\mathbf{x}} \left[\left(\bar{f}(\mathbf{x}) - g(\mathbf{x}) \right)^2 \right] + E_{\mathbf{x}} \left[E_D \left[\left(\hat{f}^{(D)}(\mathbf{x}) - \bar{f}(\mathbf{x}) \right)^2 \right] \right] \\
&= E_{\mathbf{x}} [\text{Bias}(\mathbf{x})] + E_{\mathbf{x}} [\text{Variance}(\mathbf{x})] \\
&= \text{Bias} + \text{Variance},
\end{aligned}$$

where $\text{SE}(\mathbf{x})$ is the squared error function for input \mathbf{x} (cf. Equation 2.8). We see from equation 2.14 that the expected SE over all possible data sets D can be decomposed into the bias and variance components. As we increase model variance we decrease the bias and vice versa. This relationship is also illustrated in Figure 2.6.

2.7.3 Information criteria

Considering the discussions in previous sections, it is now interesting to ask how should one select the optimal complexity of a model? Multiple methodologies originating from statistics and mathematical optimization have been developed for answering this question. One widely used approach is based on measuring how far away a fitted statistical model is from the true probability distribution which generated the observed data. Let $x^{(n)} = \{x_1, x_2, \dots, x_n\}$ be a set of n realizations of a random variable X with a true unknown distribution $G(X)$. Furthermore, let $F(X)$ be a distribution fitted to the data $x^{(n)}$, which we use for approximating $G(X)$. The probability density or mass functions of $G(X)$ and $F(X)$ are denoted as $g(x)$ and $f(x)$ respectively. To measure the closeness of these distributions, Akaike proposed in his work (Akaike, 1973) to use the *Kullback-Leibler information* (K-L information, Kullback and Leibler, 1951):

$$I(G; F) = E_G \left[\log \left\{ \frac{G(X)}{F(X)} \right\} \right] = E_G [\log G(X)] - E_G [\log F(X)], \quad (2.15)$$

where E_G denotes expectation with respect to distribution $G(X)$. In information theory the K-L information is also known as *relative entropy*. The K-L information is a fundamental building block for the concept known as *information criteria*, which are used to give a bias-corrected goodness of fit measure on how well $F(X)$ approximates $G(X)$. We will discuss later on why a bias-correction is required. Important properties of the K-L information are that $I(G; F) \geq 0$ always, and $I(G; F) = 0 \iff G(X) = F(X)$.

Proofs for these properties can be found e.g. in (Konishi and Kitagawa, 2007).

The first expectation term $E_G [\log G(X)]$ in the right side of Equation 2.15 is a constant since $G(X)$ is the true unknown distribution and therefore does not change. With the properties of the K-L information in mind, we notice that the best approximating distribution $F(X)$ is the one which maximizes the value of $E_G [\log F(X)]$ known as the *expected log-likelihood function*. In order to make calculations possible (due to the fact that $G(X)$ is unknown), we replace $G(X)$ in Equation 2.15 with the empirical distribution $\hat{G}(X)$ with a uniform probability function $\hat{g}(x_i) = 1/n, i = 1, \dots, n$ where n was the number of observed realizations of X . We denote this *empirical expected log-likelihood function* as $E_{\hat{G}} [\log F(X)]$. Due to the law of large numbers we have that $E_{\hat{G}} [\log F(X)] \rightarrow E_G [\log F(X)]$ as $n \rightarrow \infty$:

$$\begin{aligned} E_{\hat{G}} [\log F(X)] &= \int \log f(x) d\hat{G}(x) \\ &= \sum_{i=1}^n \hat{g}(x_i) \log f(x_i) = \frac{1}{n} \sum_{i=1}^n \log f(x_i), \end{aligned} \quad (2.16)$$

so we have $\frac{1}{n} \sum_{i=1}^n \log f(x_i) \rightarrow E_G [\log F(X)]$ as $n \rightarrow \infty$. The empirical expected log-likelihood function multiplied by n is therefore:

$$nE_{\hat{G}} [\log F(X)] = \sum_{i=1}^n \log f(x_i) \stackrel{\text{def}}{=} \mathcal{L}(f(x^{(n)})), \quad (2.17)$$

where $\mathcal{L}(f(x^{(n)}))$ is known as the *log-likelihood function* of the distribution $F(X)$. We see now that the higher the value of $\mathcal{L}(f(x^{(n)}))$ is, the smaller the K-L information is, and the better the fitted distribution $F(X)$ is. The log-likelihood can therefore be used to approximate the K-L information of $F(X)$. Because the probability function $f(x)$ is always parametrized by some vector $\theta \in \Theta \subset \mathbb{R}^d$, we can write $\mathcal{L}(f(x^{(n)}))$ as a function of θ :

$$\ell(\theta) = \mathcal{L}(f(x^{(n)} | \theta)). \quad (2.18)$$

Regarding the above discussion, one would select an estimator distribution $F(X)$ with a probability function $f(x | \hat{\theta})$ such that:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \ell(\theta), \quad (2.19)$$

where $\hat{\theta}$ is the MLE estimator. Note that $\ell(\hat{\theta})$ is the estimator of $nE_G[\log f(X | \hat{\theta})]$ and $n^{-1}\ell(\hat{\theta})$ is the estimator of $E_G[\log f(X | \hat{\theta})]$. It now seems reasonable that we should compare different competing models $f_j(x | \hat{\theta}_j), j = 1, \dots, p$ based on the values of $\ell(\hat{\theta}_j)$, where $\hat{\theta}_j$ is the MLE

estimator for model $f_j(x|\boldsymbol{\theta})$. It turns out however that $\ell(\hat{\boldsymbol{\theta}})$ is a biased estimator of $nE_G[\log f(X|\hat{\boldsymbol{\theta}})]$ and hence a biased goodness of fit measure, which can lead to overfitting. This happens because the MLE favors more complex models over simpler ones. Lets take a closer look on why the bias occurs. Let $\boldsymbol{\theta}_0$ be the parameter vector of the true data generating probability function $g(x|\boldsymbol{\theta}_0)$. Since $g(x|\boldsymbol{\theta}_0)$ is the true model, we always have $E_G[\log f(X|\hat{\boldsymbol{\theta}})] \leq E_G[\log g(X|\boldsymbol{\theta}_0)]$. For the log-likelihood function we have the opposite, i.e. always $\ell(\boldsymbol{\theta}_0) \leq \ell(\hat{\boldsymbol{\theta}})$ since $\hat{\boldsymbol{\theta}}$ is by definition the value which maximizes $\ell(\boldsymbol{\theta})$. This result seems to be contradicting the fact that $\ell(\hat{\boldsymbol{\theta}})$ can be used as an estimator for $nE_G[\log f(X|\hat{\boldsymbol{\theta}})]$. One would think that if log-likelihood approximates the expected log-likelihood function, then we would have $E_G[\log f(X|\hat{\boldsymbol{\theta}})] \leq E_G[\log g(X|\boldsymbol{\theta}_0)]$ and $\ell(\hat{\boldsymbol{\theta}}) \leq \ell(\boldsymbol{\theta}_0)$, which as demonstrated, is not the case. So what is the problem here? The reason why this problem occurs is that the known data $x^{(n)} = \{x_1, x_2, \dots, x_n\}$ was used twice in the estimation process. Once for fitting the model $f(x|\boldsymbol{\theta})$, and then reusing it for estimating the goodness of this model with $\ell(\boldsymbol{\theta})$. This usage of same data twice is what gives rise to the bias in $\ell(\boldsymbol{\theta})$. Consequently, in order to have a fair estimate on the goodness of a model $f(x|\boldsymbol{\theta})$ we need to remove the bias from the corresponding log-likelihood value $\ell(\boldsymbol{\theta})$.

General form of information criterion

As we discussed in the previous section, the log-likelihood function $\ell(\boldsymbol{\theta})$ needs to be bias-corrected before it can be trusted as fair measure of the goodness of an approximating model $f(x|\boldsymbol{\theta})$. In other words, we need to subtract the bias from the value of $\ell(\boldsymbol{\theta})$. The bias term $b(G)$ of the log-likelihood as an estimator of the expected log-likelihood is defined by:

$$b(G) = E_{G(x^{(n)})}[\ell(\hat{\boldsymbol{\theta}}) - nE_G[\log f(X|\hat{\boldsymbol{\theta}})]], \quad (2.20)$$

where $G(x^{(n)}) = \prod_{i=1}^n G(x_i)$ is the joint distribution of the data $x^{(n)}$. The bias free measure of model goodness is therefore of the form:

$$\text{IC} = \ell(\hat{\boldsymbol{\theta}}) - b(G). \quad (2.21)$$

The value IC obtained after the removal of the bias is called *information criterion*. Since $G(X)$ is the unknown true distribution the bias term also needs to be approximated with the available data $x^{(n)}$. The specific form of the bias term $b(G)$ depends on the relationship between $g(x)$ and $f(x)$ and the method we use to fit $f(x)$. The IC value is usually multiplied by a constant factor of -2 for "historical reasons". It is well known for example that -2 times the logarithm of the ratio of two maximized likelihood values is asymptotically χ^2 -distributed under certain conditions and assumptions (Burnham and Anderson, 2002). The -2 constant also arises in other statistical contexts, such as in the *deviance statistic* (see e.g., Gelman et al.,

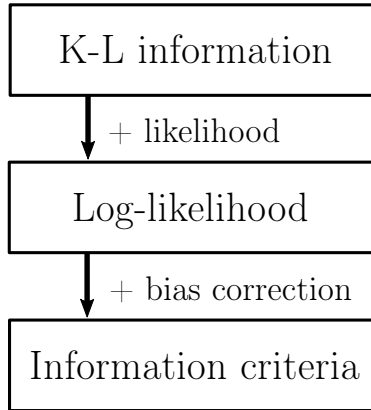


Figure 2.7: The connection of K-L information, log-likelihood function and information criteria.

2013). For further inquiry, the reader is encouraged also to study the connections of IC to the χ^2 -based hypothesis testing and likelihood ratio tests (see e.g., Akaike, 1973; Wilks, 1938). When the IC in Equation 2.21 is multiplied with -2 , we get the general form for information criterion (Konishi and Kitagawa, 2007):

$$\text{IC} \equiv -2(\ell(\hat{\boldsymbol{\theta}}) - b(G)) = -2\ell(\hat{\boldsymbol{\theta}}) + 2b(G). \quad (2.22)$$

Note that given the form of the IC in Equation 2.22, the better our fitted model $f(x | \hat{\boldsymbol{\theta}})$ is, the lower the value of IC. In Figure 2.7 is illustrated the summarized connection between K-L information, log-likelihood function and information criteria. In the next subsections we will briefly go through some examples of information criteria.

Takeuchi's information criterion

The information criterion with the bias term resulting from the estimation of the expected log-likelihood using the log-likelihood of the approximation model, has asymptotically the form:

$$\text{TIC} = -2 \sum_{i=1}^n \log f(x_i | \hat{\boldsymbol{\theta}}) + 2 \text{Tr} \left(I(\hat{\boldsymbol{\theta}}) J^{-1}(\hat{\boldsymbol{\theta}}) \right), \quad (2.23)$$

where $\text{Tr} \left(I(\hat{\boldsymbol{\theta}}) J^{-1}(\hat{\boldsymbol{\theta}}) \right)$ is the approximated bias term with $d \times d$ matrices $I(\hat{\boldsymbol{\theta}})$ and $J^{-1}(\hat{\boldsymbol{\theta}})$. The $\text{Tr}(\cdot)$ operator stands for the trace of matrix. Matrices $I(\hat{\boldsymbol{\theta}})$ and $J(\hat{\boldsymbol{\theta}})$ have the forms:

$$I(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f(x_i | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x_i | \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad (2.24)$$

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(x_i | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad (2.25)$$

with the corresponding $(j, k)^{\text{th}}$ elements:

$$I_{jk}(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f(x_i | \boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \log f(x_i | \boldsymbol{\theta})}{\partial \theta_k} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad (2.26)$$

$$J_{jk}(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log f(x_i | \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \quad (2.27)$$

This information criterion is known as *Takeuchi's information criterion* (TIC, Takeuchi, 1976; Stone, 1977). TIC is rarely used in practice but is a more general having less assumptions than other more used IC statistics. One of the reasons for low popularity of TIC is that it is not widely known and that it is much more complicated to compute than other commonly used IC (see e.g. Anderson, 2008). The complication arises from the fact that TIC involves the estimation of two $d \times d$ matrices of first and second partial derivatives, matrix inverse, and then matrix product. Unless a large data set is available, the bias term in TIC is often numerically unstable. However, it turns out that a very good estimate for the bias term can be calculated in other IC statistics, one of which we discuss next.

Akaike information criterion

The *Akaike information criterion* (AIC) is almost identical to TIC, but it is a more often used and famous model selection criterion. To be more precise, AIC is actually a special case of TIC with a stronger assumption. In AIC we assume that the hypothesis space $\mathcal{H} = \{f(x | \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\}$ contains the true model $g(x | \boldsymbol{\theta}_0)$, i.e. $\exists \boldsymbol{\theta}_0 \in \Theta$ so that $f(x | \boldsymbol{\theta}_0) = g(x | \boldsymbol{\theta}_0)$. Under this assumption, it follows that $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ and thus the bias term in Equation 2.23 asymptotically becomes:

$$b(G) = \text{Tr} \left(I(\hat{\boldsymbol{\theta}}) J^{-1}(\hat{\boldsymbol{\theta}}) \right) = \text{Tr} \left(I(\hat{\boldsymbol{\theta}}) I^{-1}(\hat{\boldsymbol{\theta}}) \right) = \text{Tr} (\mathbf{I}_{d \times d}) = d. \quad (2.28)$$

Plugging the corresponding bias term into Equation 2.23 we get:

$$\text{AIC} = -2 \sum_{i=1}^n \log f(x_i | \hat{\boldsymbol{\theta}}) + 2d. \quad (2.29)$$

Bayesian information criterion

A Bayesian approach for measuring the goodness of fit of a statistical model is based on approximating the *marginal likelihood* (also known as marginal

distribution) of data $x^{(n)}$ by using Laplace's approximation for integrals and Taylor expansion (Konishi and Kitagawa, 2007). The marginal likelihood is defined as:

$$p(x^{(n)}) = \int f(x^{(n)} | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \exp(\ell(\boldsymbol{\theta})) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (2.30)$$

where $f(x^{(n)} | \boldsymbol{\theta})$ is the likelihood of data $x^{(n)}$ and $\pi(\boldsymbol{\theta})$ is the prior distribution for model parameters $\boldsymbol{\theta}$. It can be shown, that using Laplace's approximation for integrals and Taylor expansion for $\ell(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ around the MLE estimator $\hat{\boldsymbol{\theta}}$, we get:

$$p(x^{(n)}) \approx \exp\{\ell(\hat{\boldsymbol{\theta}})\} \pi(\hat{\boldsymbol{\theta}}) (2\pi)^{d/2} n^{-d/2} |J(\hat{\boldsymbol{\theta}})|^{-1/2}, \quad (2.31)$$

where again d is the number of model parameters, n the number of samples, and

$$\begin{aligned} |J(\hat{\boldsymbol{\theta}})|^{-1/2} &= \det\left(-\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^{-1/2} \\ &= \det\left(-\frac{1}{n} \frac{\partial^2 \log f(x^{(n)} | \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)^{-1/2}. \end{aligned} \quad (2.32)$$

By taking the logarithm, multiplying by -2 , and ignoring terms with order less than $\mathcal{O}(1)$ with respect to n in Equation 2.31, we get the *Bayesian information criterion* (BIC):

$$\begin{aligned} \text{BIC} &= -2 \log f(x^{(n)} | \hat{\boldsymbol{\theta}}) + d \log n \\ &= -2 \log \prod_{i=1}^n f(x_i | \hat{\boldsymbol{\theta}}) + d \log n \\ &= -2 \sum_{i=1}^n \log f(x_i | \hat{\boldsymbol{\theta}}) + d \log n. \end{aligned} \quad (2.33)$$

Minimum description length

In 1978, Jorma Rissanen proposed in his work (Rissanen, 1978) the *minimum description length* (MDL) principle from an information-theoretic perspective for measuring the goodness of a statistical model. MDL is based on the idea that the best model to explain data $x^{(n)}$, is the simplest such model which compresses data $x^{(n)}$ the best. If data contains regularities, then we are able to generate a code C such that this code explains the data in the shortest possible way. To illustrate this idea more, consider two binary strings: a random string of bits $B_1 = 011001010111\dots$, and the string $B_2 = 001001001001\dots$, with a clear regular pattern. With B_1 , we can not compress this string at all, since it is a random bit string containing no

regularities. On the other hand with B_2 , we can compress it by writing a code with a `for`-loop which repeatedly prints the bit string 001. Motivated by this example, we now state the MDL principle. Given data $x^{(n)}$ and a hypothesis set \mathcal{H} , the MDL principle states that we should select a model \hat{f} such that:

$$\hat{f} = \arg \min_{f \in \mathcal{H}} L_C(x^{(n)} | f) + L_C(f), \quad (2.34)$$

where $L_C(x^{(n)} | f)$ stands for the code length of data $x^{(n)}$ given the model f and $L_C(f)$ is the code length of the model f itself. In other words, we select the model for which the sum of the data encoding length (given the model) and the model encoding length is the shortest. It can be shown that every *prefix code* C has a corresponding probability distribution $F(X)$ (see e.g., Rissanen, 1989; Grünwald, 2007). A prefix code is a code system, in which no whole code word is a prefix of any other code word in the system. In terms of probability distributions, the MDL can be stated as (considering terms up to order $\mathcal{O}(\log n)$, Konishi and Kitagawa, 2007):

$$\begin{aligned} \text{MDL} &= -\log f(x^{(n)} | \hat{\theta}) + \frac{d}{2} \log n \\ &= -\sum_{i=1}^n \log f(x_i | \hat{\theta}) + \frac{d}{2} \log n, \end{aligned} \quad (2.35)$$

which is equivalent with Equation 2.34. The first term in the right-hand side is the encoding length in sending the data $x^{(n)}$ by using the probability distribution $f(x^{(n)} | \hat{\theta})$, which is specified by the maximum likelihood estimator $\hat{\theta}$ as the encoding function. The second term is the encoding length for the maximum likelihood estimate $\hat{\theta}$ with accuracy $\delta = \mathcal{O}(n^{-1/2})$. It is interesting to note that the MDL coincides with the BIC ($\text{MDL} = \text{BIC}/2$) that was derived within the Bayesian framework. As with all the information criteria in the previous subsections, a model having the smallest MDL value is considered to be the best model to explain the data $x^{(n)}$.

2.7.4 Regularization

Model complexity selection using information criteria were based on selecting the model which minimized the IC value. The information criteria gave a fair comparison between competing models by calculating the unbiased log-likelihood value. As we discussed earlier, the best model is the one having enough expressive power to capture the relevant pattern in the data, but simple enough not to capture the irrelevant non-existing pattern produced by the noise in the data. Another approach to reach this same goal besides using information criteria, is to use the concept of *regularization*. Regularization is the means to constrain the model training process via penalization in order to prevent overfitting from happening. Constraining the

model training process using regularization helps us to reach the optimal balance in the bias-variance trade-off as we discussed in section 2.7.2 (see Figure 2.6). In the next two subsections, we will present two common ways of regularizing the model fitting process.

Penalty functions

Model fitting is fundamentally a constrained optimization problem which calls for the need of nonlinear programming. As we discussed in section 2.6.1, often the goal in model fitting is to minimize some objective function \mathcal{E} (see Equation 2.7). In section 2.6.1, Equation 2.8, we saw an example of the very common error function SE. We could find such a model which minimizes the SE value, but this would produce an overfitted model. To alleviate this problem we can use a method called *penalty functions* (Bazaraa, 2013), which is a way to regularize the training process by introducing a penalty term into the minimization problem of Equation 2.7. The penalty term makes sure that we do not minimize \mathcal{E} too much, but to sufficient extent. In effect, this causes too extreme models not to be chosen from \mathcal{H} . To make things more explicit, let $\boldsymbol{\theta} \in \mathbb{R}^d$ be an independent variable, $\mathcal{E}(\boldsymbol{\theta})$ an error function and $\Omega(\boldsymbol{\theta})$ a penalty function. A model fitting optimization problem can now be formulated as:

$$\begin{aligned} & \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{minimize}} && \mathcal{E}(\boldsymbol{\theta}) \\ & \text{subject to} && \Omega(\boldsymbol{\theta}) = 0. \end{aligned} \tag{2.36}$$

A vector $\boldsymbol{\theta}$ that satisfies the constraint of Problem 2.36 is called a *feasible solution* and the set of feasible solutions is called a *feasible region*. The minimization of $\mathcal{E}(\boldsymbol{\theta})$ is therefore limited to the feasible region of the parameter space. When considering overfitting, we can think the penalty function $\Omega(\boldsymbol{\theta})$ as limiting the solutions $\boldsymbol{\theta}$ to those which have a good balance in the bias-variance trade-off. In other words, $\Omega(\boldsymbol{\theta})$ reduces overfitting. The idea to regularize with penalty functions is to reformulate the optimization problem of Equation 2.36 in the following way:

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{minimize}} \quad \mathcal{E}(\boldsymbol{\theta}) + \eta (\Omega(\boldsymbol{\theta}))^2, \tag{2.37}$$

where $\eta > 0$ is a large positive number. Note that this optimization problem is now unconstrained for $\boldsymbol{\theta}$. However, the vector $\boldsymbol{\theta}$ is still constrained to the feasible region due to the penalty term, because otherwise a large penalty in $\eta (\Omega(\boldsymbol{\theta}))^2$ would occur. The learning algorithm consequently favors solutions $\boldsymbol{\theta}$ such that $\Omega(\boldsymbol{\theta}) = 0$. This reformulation of Problem 2.36 therefore has a built-in regularization. It must be emphasized, that a general ML problem does not require the condition $\Omega(\boldsymbol{\theta}) = 0$ to be fulfilled exactly. Examples of

regularizing the MSE are given below:

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 + \lambda \sum_{j=1}^d I(\theta_j \neq 0), \quad (2.38)$$

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 + \lambda \sum_{j=1}^d |\theta_j|, \quad (2.39)$$

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\theta})^2 + \lambda \sum_{j=1}^d \theta_j^2, \quad (2.40)$$

where $\lambda > 0$ is the regularization parameter controlling the extent of regularization. The regularization used in Equation 2.38 is called ℓ_0 -regularization. The function $I(a)$ maps to 1 if a is true and 0 otherwise. That is, the penalty term in Equation 2.38 is the number of nonzero elements of $\boldsymbol{\theta}$ multiplied by λ . Equations 2.39 and 2.40 are called ℓ_1 -, and ℓ_2 -regularization respectively. The latter two are also known as *least absolute shrinkage and selection operator* (LASSO) and *Tikhonov regularization*. A regression method which linearly combines both ℓ_1 -, and ℓ_2 -regularizations is known as *elastic net* (Zou and Hastie, 2005). The idea of regularization relates to the MLE procedure by a process called *penalized maximum likelihood estimation*, which can be shown to be equivalent with the regularization method (see e.g., Akaike, 1980; Konishi and Kitagawa, 2007; Wahba, 1978).

Early stopping

Some model fitting problems can be solved analytically (e.g. in Tikhonov regularization), but others must be solved numerically using iterative optimization methods. Iterative optimization methods include e.g. gradient-descent, golden-section search and Newton-Raphson method (Bazaraa, 2013). Iterative methods improve the model fit to training data in a step-by-step manner. In this case it is important to identify the optimal number of iteration steps because a too low number of iterations will result in underfitting and too large number of iterations will result in overfitting. A form of regularization called *early stopping* is used to avoid overfitting in cases where we need to train a model with iterative methods (Bishop, 1996). The early stopping regularization is implemented in the following way:

1. Split the data set D into a training set D_t and a validation set D_v .
2. Perform one iteration in the model learning process using set D_t for training and set D_v for evaluating model performance.

3. Repeat step 2 as long as the model performance improves and stop at iteration i , which is the first time when the model performance decreases.
4. Choose the model parameters from iteration $i - 1$ and stop the training process.

Figure 2.6 illustrating the bias-variance trade-off presents also the idea of early stopping. With increasing model fit (number of iterations) the performance improves until optimal point is reached after which performance starts to get worse.

2.7.5 Cross validation

In many cases one is interested in estimating the prediction performance of a fitted model f . With a classification model for example, we might want to give some estimated value, such as a classification accuracy $p\%$, on the model performance with new data. One way to do this estimation is to first partition the observed data set D into two subsets: *training data set* D_t and *validation data set* D_v (as we did in early stopping). The set D_t is then used for training the model f and set D_v for validating its performance. It is very important that the sets D_t and D_v do not overlap, i.e. $D_t \cap D_v = \emptyset$. Otherwise, we risk having biased performance estimates with the validation set D_v , because data points in D_v were used also to train the model f .

One can now ask, how should we partition the data set D ? If D_t contains most of the data points and D_v only small amount, then we get better model fitting, but not so reliable performance estimation. If on the other hand D_v contains most of the data points and D_t only small amount, then we get a poor fit, but more reliable performance estimation. A method called *cross validation* (CV) is a way to take the advantage from both of these approaches: use all the data for fitting and performance estimation. The CV procedure (see e.g., Gelman et al., 2013) is implemented in the following way:

1. Index the data points in D by using a set $\mathcal{I} = \{1, 2, \dots, n\}$.
2. Partition the index set \mathcal{I} into k disjoint random sets I_1, I_2, \dots, I_k . That is, each $I_j \subset \mathcal{I}$, $\cup_{i=1}^k I_i = \mathcal{I}$ and $I_j \cap I_i = \emptyset \quad \forall j \neq i$.
3. Denote next the set I_{-i} as the set containing all indexes without the ones included in set I_i , that is $I_{-i} = \bigcup_{\substack{j=1 \\ j \neq i}}^k I_j$.
4. For each I_j , $j \in \{1, 2, \dots, k\}$ train a model using data indexed by I_{-j} and use the inputs in \mathcal{X} indexed by I_j with the trained model

to calculate the estimated set $\hat{\mathcal{Y}}_j$ for the outputs in \mathcal{Y} indexed by I_j . Denote the set of estimated outputs as $\hat{\mathcal{Y}} = \cup_{j=1}^k \hat{\mathcal{Y}}_j$.

5. Report model goodness as $\mathcal{E}(\mathcal{Y}, \hat{\mathcal{Y}})$ where \mathcal{E} is some performance function.

This CV procedure of partitioning the data into k disjoint *hold-out sets* indexed by I_j is called more specifically *k-fold cross validation* (KCV). A special case of KCV where $k = n$, is known as *leave-one-out cross validation* (LOOCV). The partitioning of the folds can be done either systematically or randomly depending on the application. With geographical data sets for example, in order to reduce the effect of SAC in performance estimates one could design a fold partitioning so that data points in D_v do not have training data points of D_t close to them (Pohjankukka et al., 2017). Regarding section 2.7.3 it can be shown that CV offers an alternative approach to estimate the K-L information from a predictive point of view. In fact, it can be shown that LOOCV is asymptotically equivalent to AIC for linear regression models (Stone, 1977; Shibata, 1989; Konishi and Kitagawa, 2007).

Nested cross validation

Besides offering prediction performance estimation CV is also used for selecting model hyperparameters such as the λ regularization parameter in Equations 2.38, 2.39 and 2.40. In this situation however a problem arises. In a CV procedure where the data set D is partitioned into sets D_t and D_v biased performance estimation is bound to happen. The bias occurs due to a phenomenon known as *data snooping* (see e.g., Abu-Mostafa et al., 2012) which refers to the case where we modify our model after we have evaluated it using set D_v . In other words, we are using the validation data to guide the model selection process which therefore results in biased performance estimation. Depending on the application and CV procedure however, the bias can be smaller or larger. For example for LOOCV with a large data set the bias is usually negligible for the purposes of the application and can be ignored. In order to prevent data snooping from occurring in the model performance estimation we use a CV procedure called *nested cross validation* (NCV). NCV is almost the same as normal CV with the exception that now the data partitioning includes a third set D_e called a *test data set*. The set D_e is used to evaluate the model performance after it has been fully trained using sets D_t and D_v with hyperparameters selected. The selection of model parameters never includes the data in set D_e which therefore prevents data snooping from happening.

2.7.6 Spatial k-fold cross validation

When we are dealing with geographically distributed natural data it is not necessarily sufficient to consider only model performance criteria such as IC or standard CV, which rely on i.i.d. assumptions. These assumptions can cause the standard criteria to favor false models due to optimistic bias caused by SAC. In the included publication (Pohjankukka et al., 2017) a CV based method called *spatial k-fold cross validation* (SKCV) was proposed for measuring the prediction performance of spatial models, which takes into account the bias caused by SAC. In SKCV, optimistic bias in performance estimates is prevented by making sure that the training data set D_t only contains data points that are at least a certain spatial distance away from the prediction point.

We now introduce notation to give formal definition for the SKCV. Let $\mathbf{c}_i \in \mathbb{R}^2$ denote a geographical location (coordinate) vector of data point (\mathbf{x}_i, y_i) . A geographical data point is denoted as $\mathbf{d}_i = (\mathbf{x}_i, y_i, \mathbf{c}_i)$ and the corresponding data set of n geographical data points as $D_c = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_n\}$. A value $r_\delta \in \mathbb{R}^+$ is the so-called *dead zone* radius, which determines the data points removed from the training data at each SKCV iteration. Next, we denote the set $\mathcal{V} = \{\mathcal{V}_1, \dots, \mathcal{V}_k\}$ as the set of k distinct CV folds, where we have $\mathcal{V}_p \subset D_c \forall p$, $\mathcal{V}_p \cap \mathcal{V}_q = \emptyset$, when $p \neq q$ and $\cup_{i=1}^k \mathcal{V}_i = D_c$. Furthermore, we use the Euclidean distance function e for calculating spatial distances between data pairs $\mathbf{d}_i, \mathbf{d}_j$. In Algorithm 1 we show the pseudocode for the SKCV. When $k = n$, then the SKCV is called *spatial leave-one-out CV* (SLOO, Le Rest et al., 2014; Pohjankukka et al., 2017). An illustration of the SKCV method is also given in Figure 2.8. The SKCV can be used for estimating the spatial prediction performance of a model as a function of r_δ (and therefore for model complexity selection) and also for selecting suitable hexagonal data sampling grid for new research areas (see Pohjankukka et al., 2017). Naturally, as we increase the dead zone radius r_δ with data sets involving SAC the prediction performance decreases. The

Algorithm 1 Spatial k-fold cross validation

Require: $\mathcal{V}, D_c, \mathcal{A}, r_\delta$

Ensure: $\hat{\mathbf{y}}$

- 1: **for** $i \leftarrow 1$ to k **do**
 - 2: $D_r \leftarrow \bigcup_{\mathbf{d}_l \in \mathcal{V}_i} \{\mathbf{d}_j \in D_c \mid e(\mathbf{c}_j, \mathbf{c}_l) \leq r_\delta\}$ ▷ Remove close data points
 - 3: $\hat{f} \leftarrow \mathcal{A}(D_c \setminus D_r)$ ▷ Build model using reduced training set
 - 4: **for** $\mathbf{d}_l \in \mathcal{V}_i$ **do**
 - 5: $\hat{\mathbf{y}}[l] \leftarrow \hat{f}(\mathbf{x}_l, \mathbf{c}_l)$ ▷ Make prediction
 - 6: **return** $\hat{\mathbf{y}}$ ▷ The predicted $\hat{\mathbf{y}}$
-

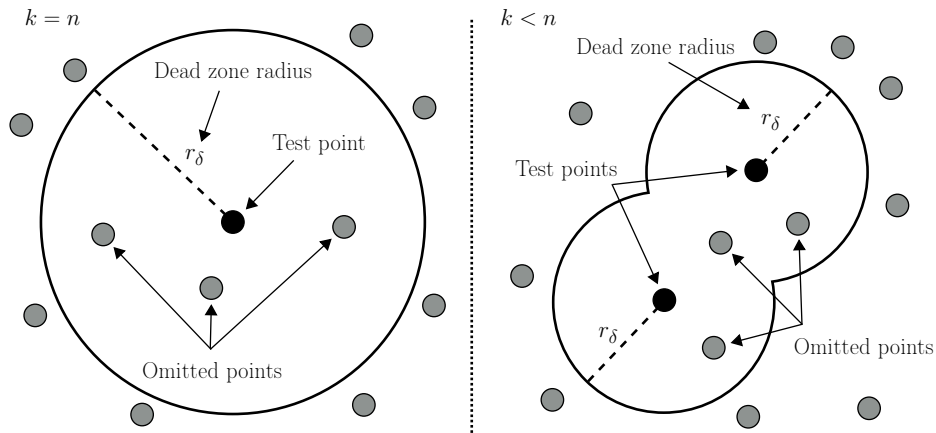


Figure 2.8: Reduction of the training set in the SKCV procedure. The black and gray points correspond to test and training data points respectively. The gray data points inside the perimeters of radius r_δ are omitted from the training data, after which the test points are predicted using the remaining training data (i.e. the gray data points outside the perimeters).

reason why the method can be used for data sampling density selection is because in the SKCV the test data is always at least r_δ distance units away from the training data. By then sampling data from a new research area in a hexagonal manner with radius r_δ the sampled data points are always at most r_δ distance units away from the training data. The SKCV therefore gives insight on how good generalization performance can be expected to achieve using a hexagonally sampled data set with radius r_δ .

One might also have few issues raised with the SKCV which we will address here. Firstly, since the SKCV involves the reduction of the training set one might argue that this obviously results in pessimistic bias. It is however shown in the publication that the bias caused by data reduction is negligible when compared with the bias caused by SAC. Secondly, it is important that one takes care when selecting the number of folds k in the SKCV. If this number is very small (say $k = 2$), then it could happen that most of the training data is removed due to large combined dead zone of the test data points. The selection of k is application-specific and must be chosen to suit the purposes of the corresponding application.

2.8 Feature selection

In the previous section we were dealing with model evaluation and selection. There was no further investigation on the predictor features in set \mathcal{X} and was used as such. In this section, we focus on measuring the goodness of these

features by implementing *feature selection*. In many practical applications it is important to implement feature selection on the observed data D . The data might contain irrelevant and even detrimental features, which should be discarded from the data set. We implement feature selection for reasons such as:

- To remove irrelevant information from the data. The irrelevant features can be considered as noise in the data. Only slightly useful features can also be discarded depending on the application.
- To reduce the dimensionality of the data. Dimensionality reduction decreases the problems caused by the well known concept of *the curse of dimensionality*, which states in simplified terms: the number of data points needed to achieve successful model approximation grows exponentially with respect to the dimensionality of the data.
- To reduce the model fit time. It is well known for example in statistical simulation that high dimensional data sets are generally very difficult and time consuming to fit in terms of convergence. Fortunately, many methods such as MCMC or the improved version of this the Hamiltonian (or hybrid) Monte Carlo (HMC, see e.g., Neal, 1994) methods have been developed for dealing with high dimensional data.

In the included publication (Pohjankukka et al., 2018) feature selection was implemented to recognize the optimal predictors needed for MS-NFI attribute estimation. Many methods have also been developed for feature selection and dimensionality reduction such as *principal component analysis* (PCA), *greedy selection*, genetic algorithm (GA) and *automatic relevance determination* (ARD). For more information about these methods see the work of (Theodoridis and Koutroumbas, 2008). In the next subsections we will go through few of these.

2.8.1 Greedy forward/backward selection

Greedy selection (see e.g., Pahikkala et al., 2010) is a depth-first type feature selection method which chooses features sequentially based on how much they increase or decrease the model performance. There are two greedy feature selection methods: *forward selection* and *backward selection*.

Greedy forward selection (GFS) proceeds by sequentially ordering the features based on how much they improve the model performance. To be more precise, the steps of the forward selection procedure are the following:

1. Denote $\mathcal{F} = \{w_1, w_2, \dots, w_d\}$ as the feature set and $\mathcal{F}^* = \emptyset$ as the set to be used for constructing an ordered version of \mathcal{F} .

2. For each feature $w_j \in \mathcal{F}$, form a candidate set $F_j = \mathcal{F}^* \cup \{w_j\}$, fit approximation model using features of F_j and calculate a model performance estimate p_j for this feature combination.
3. For index i with best value of p_i , set $\mathcal{F}^* = \mathcal{F}^* \cup \{w_i\}$ and $\mathcal{F} = \mathcal{F} \setminus \{w_i\}$.
4. Repeat the steps 2-3 until $\mathcal{F} = \emptyset$ and return the set \mathcal{F}^* .

The returned set \mathcal{F}^* after the steps shown above contain all the features of \mathcal{F} , which are ordered based on sequential maximum improvement of model performance. In other words, the first feature w_j to be selected to \mathcal{F}^* , is the best predictor feature to be used if we used only one feature in our model. The second feature w_i selected, is the one which has the maximum model performance improvement, when combined with the first selected feature w_j . The third, fourth, et cetera features are selected using this same logic.

Greedy backward selection (GBS) is the opposite of GFS. The steps of GBS are almost identical to those of GFS, but instead of including it removes features in a step-by-step manner. GBS starts with all the features in its disposal and then it sequentially removes features based on how much the model performance decreases. At each iteration step, a feature with the lowest decrement on the model performance is removed. GBS produces a flipped version of \mathcal{F}^* , first is the worst feature, then the second worst, et cetera.

2.8.2 Genetic algorithm

GA (see e.g., Goldberg, 1989) is a heuristic search method inspired by the theory of natural selection by Charles Darwin. GA is based on the idea that from a population of possible solutions, good solutions survive and produce offspring, whereas bad solutions do not survive and are discarded from the population. The GA method is a good alternative to problems involving non-convex optimization problems and can help in avoiding local optima. A central component in GA is the *chromosome*, which represents a single solution in the population space \mathbf{B} . In feature selection, the chromosome can be represented as a binary vector $\mathbf{b} \in \mathbf{B} \subset \{0, 1\}^d$, where the i th element (known as *gene*) of the vector \mathbf{b} corresponds to a yes/no decision on whether the i th feature should be included in that solution. Other central components of GA include *crossover* and *mutation* operators. The crossover operator refers to the case where two (or more) parent chromosomes \mathbf{b}_i and \mathbf{b}_j produce offspring chromosomes \mathbf{b}_k , which share genetic material with both parents. For example, if $\mathbf{b}_i = (0, 1, 0, 0)$ and $\mathbf{b}_j = (0, 0, 1, 1)$, then one possible offspring chromosome is $\mathbf{b}_k = (0, 1, 1, 1)$. In this example, the offspring \mathbf{b}_k shares half of the genes with parent \mathbf{b}_i and the other half with parent \mathbf{b}_j . The mutation operation refers to a random change in the genes of an individual chromosome \mathbf{b} . The genes of \mathbf{b} may swap places or

change values in mutation. Examples of gene swap and value changes are $(0, 0, 1, 1) \rightarrow (1, 0, 1, 0)$ and $(0, 0, 1, 1) \rightarrow (0, 0, 1, 0)$. In GA, crossover is set to occur with high probability and mutation with low probability. The decision on what chromosomes get to reproduce is evaluated using some *fitness function* \mathcal{E} suitable for the corresponding problem. The general procedure of the GA feature selection can be summarized into the following steps:

1. Denote $\mathcal{F} = \{w_1, w_2, \dots, w_d\}$ as the feature set and $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_k\}$ as the set of k different binary chromosomes.
2. Evaluate each chromosome in \mathbf{B} using the fitness function $\mathcal{E}(\mathcal{F}, \mathbf{b})$ and select the suitable parent chromosomes. Denote the set of these parents as \mathbf{B}_p .
3. Perform crossover and mutation operations with the chromosomes in \mathbf{B}_p and produce a new population set \mathbf{B}^* consisting from the offspring chromosomes. Evaluate the offspring chromosomes (feature combinations) of \mathbf{B}^* with $\mathcal{E}(\mathcal{F}, \mathbf{b})$.
4. If termination conditions are met (good enough feature combinations found), return the feature combination corresponding to the best chromosome in \mathbf{B}^* . If termination conditions are not met, set $\mathbf{B} = \mathbf{B}^*$ and repeat the steps 2-4 until conditions are met.

2.8.3 Automatic relevance determination

ARD is an elegant feature selection method provided by *relevance vector machines* (RVM, Tipping, 2000; Junttila et al., 2008), which is a Bayesian analogue to the well-known maximum margin based method *support vector machines* (SVM, see Vapnik, 1998). RVM starts from a problem where the goal is to find a parameter vector $\boldsymbol{\theta}$ such that:

$$t = \sum_{j=1}^d \theta_j \phi_j(\mathbf{x}) = \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}), \quad (2.41)$$

where $\mathbf{x} \in \mathbb{R}^m$ is an input vector, $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots, \phi_d(\mathbf{x}))$ is a vector of basis functions and $t \in \mathbb{R}$ is an unknown true output value. The observed output value $y_i \in \mathcal{Y}$, $i \in \{1, 2, \dots, n\}$, representative of t_i , includes a noise term $\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \forall i$ so that:

$$y_i = t_i + \varepsilon_i = \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_i) + \varepsilon_i. \quad (2.42)$$

By making multivariate Gaussian assumptions, denoting $\beta = \sigma^{-2}$ and having $\theta_j | \alpha_j \sim \mathcal{N}(0, \alpha_j^{-1}) \forall j$ with hyperparameter α_j , it can be shown (Bishop,

2006) that the posterior predictive distribution over y for a new input vector \mathbf{x} is:

$$\begin{aligned} p(y | \mathbf{x}, D, \boldsymbol{\alpha}, \beta) &= \int p(y | \mathbf{x}, \boldsymbol{\theta}, \beta) p(\boldsymbol{\theta} | D, \boldsymbol{\alpha}, \beta) d\boldsymbol{\theta} \\ &= \mathcal{N}(\mathbf{m}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2(\mathbf{x})), \end{aligned} \quad (2.43)$$

where we have:

$$\begin{aligned} p(y | \mathbf{x}, \boldsymbol{\theta}, \beta) &= \mathcal{N}(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}), \\ p(\boldsymbol{\theta} | D, \boldsymbol{\alpha}, \beta) &= \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}), \\ \mathbf{m} &= \beta \boldsymbol{\Sigma} \boldsymbol{\Phi}^T \mathbf{y}, \\ \boldsymbol{\Sigma} &= (\mathbf{A} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1}, \\ \sigma^2(\mathbf{x}) &= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\Sigma} \boldsymbol{\phi}(\mathbf{x}). \end{aligned} \quad (2.44)$$

In (2.44), $\boldsymbol{\Phi}$ is the $n \times d$ design matrix such that the i th row represent the vector $\boldsymbol{\phi}(\mathbf{x}_i)$, $\mathbf{y} \in \mathbb{R}^{n \times 1}$ is a column vector of the observed output values and \mathbf{A} is a $d \times d$ diagonal matrix with diagonal elements α_j , i.e. $\mathbf{A} = \text{diag}(\alpha_j)$. The optimal hyperparameters $\boldsymbol{\alpha}, \beta$ of Equation 2.43 are obtained by maximizing the *evidence*:

$$\hat{\boldsymbol{\alpha}}, \hat{\beta} = \arg \max_{\boldsymbol{\alpha} \in \mathbb{R}^d, \beta \in \mathbb{R}} P(\mathbf{y} | \boldsymbol{\alpha}, \beta). \quad (2.45)$$

The maximization of evidence in Equation 2.45 is implemented using e.g. *expectation-maximization* (EM) algorithm, making the RVM method sometimes computationally more complex than SVM. The details for Equation 2.45 have been discarded here to avoid laborious derivation and can be found e.g. in (Bishop, 2006; Fletcher, 2010).

The ARD feature selection comes into this when we are carrying out the evidence maximization procedure. Here, many of the α_j will tend to infinity causing the posterior distribution of parameter θ_j to have zero mean and variance. That is, when $\alpha_j \rightarrow \infty \Rightarrow \theta_j = 0$. The corresponding column with basis function ϕ_j in the design matrix $\boldsymbol{\Phi}$ is therefore effectively pruned out. The inputs \mathbf{x}_i corresponding to the remaining non-zero parameters θ_i after pruning are called *relevance vectors* and are analogous to the *support vectors* of SVM.

Chapter 3

Research studies and results

3.1 Research publications

In this section, we will briefly go through the research studies conducted in the publications included into this thesis. In each of the next subsections, a summary is first given to introduce and motivate the research, which is followed by a description of the used data sets and methods, and finally the results and contributions to the research questions are presented.

3.1.1 Publication I: Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data

Summary

The publication (Pohjankukka et al., 2014b), examines the predictability of soil type and its hydraulic conductivity using open natural resource data from Parkano, Finland. Prior information about soil conditions is an important factor in a multitude of applications. In forestry for example, the level of success of a forest harvest is dependent on a careful route planning. Well planned routing decisions minimize collateral damage to the forest and maximize safety of the harvest, which results in reduced costs for both the harvest operation and the forest owner. The hydraulic conductivity attribute of soil has greatest influence to its load bearing capacity. It is therefore essential to measure directly or estimate the hydraulic conditions of a soil prior to the harvest in order to fulfill the safety and efficiency requirements. Regression and classification analyses are implemented in the publication for assessing the prediction capability of soil hydraulic conductivity and soil type respectively.

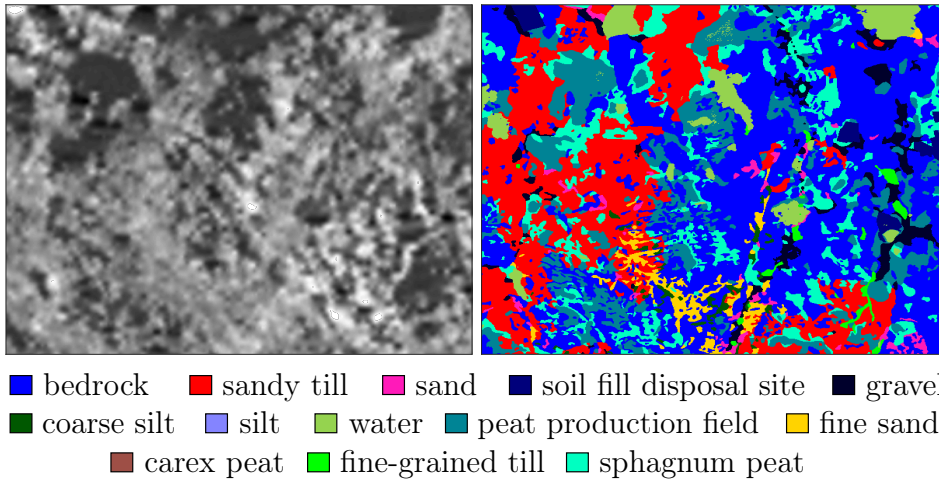


Figure 3.1: *Left*: gamma-ray spectroscopy imagery from Parkano. *Right*: corresponding ground soil type imagery with legend given below.

Methods and data

The regression part of the analysis is implemented using the well-known *ridge regression* (also known as regularized least squares, RLS) method. Target variable in regression case is the water conductivity exponent (or water permeability exponent) x_{wp} , which measures the vertical flow speed of water in soil. Different soil types tend to have distinct values of x_{wp} , making successful estimation of x_{wp} useful in route planning for selecting harvest tracks with optimal soil types and highest load bearing capacities. The classification analysis is implemented using *logistic regression* and kNN methods. The target variable in this case is the ground soil type class. The input predictor features consist from gamma-ray spectroscopy and various topographic feature (e.g. topographic height) raster data sets. A total of 342479 data points were used in the analyses. The data sets in the publication are provided by the GTK and the NLS. In Figure 3.1 is illustrated the gamma-ray spectroscopy and ground soil type data sets. In the gamma-ray image darker areas correspond to areas with greater humidity since humid soil absorbs more gamma radiation than dryer soil. The ground soil type image is a pre-classified (by the GTK) data set with pixel values corresponding to soil types.

Results and contribution to research questions

The regression results for water permeability exponent x_{wp} show that RLS achieves slightly better *concordance index* (C-index, see e.g., Pohjankukka et al., 2014b) of 0.63 than baseline method (random coin flip, C-index 0.5).

The classification results for soil type shows 44.5% prediction accuracy for logistic regression and 50.5% prediction accuracy for kNN.

The publication contributes to research question (**RQ1**) by giving quantitative results on the predictability of terrain conditions using the corresponding natural resource data sets with ML approach.

Author’s contribution

The author’s responsibilities in the publication consisted from: preprocessing and merging of the data, modeling and evaluation of the results, and writing of the article. Preprocessing involved cleaning the data by discarding irrelevant features (irrelevance determined by data providers) and data points with missing feature values. Merging of the data here refers to combining the data sources into a single data matrix with uniform resolution. In order to provide compatible data sets, this needed to be done since the data files had different resolutions and geographical covers. Modeling and evaluation included steps such as: selecting optimal kNN model using LOOCV, calculation of prediction accuracies et cetera. Implementation was conducted using Matlab and Python environments. Self-made and off-the-shelf (RLScore, scikit-learn, see Pahikkala and Airola, 2016) code libraries were used in the analyses.

3.1.2 Publication II: Predicting water permeability of the soil based on open data

Summary

The publication (Pohjankukka et al., 2014a) studies the same problem as Publication I, but with additional predictor data sets and at different geographical location. The predictability of water permeability exponent x_{wp} is examined using AEM data, gamma-ray spectroscopy, topographic feature data, MS-NFI data, and peat bog thickness data. The research area is located in the northern part of the municipality of Sodankylä, Pomokaira, Finland. Water permeability is a key factor when estimating soil load bearing capacity, mobility and infrastructure potential of a terrain. Soil with high levels of water permeability tend to stay dry and traversable, whereas soil with low permeability creates a risk for mobility. Northern sub-arctic areas have similar soil types so successful prediction in the region of Northern Finland can be generalizable to other similar regions. The study was motivated by technical and cost issues originating from forest industry.

Methods and data

Regression analysis was conducted using RLS and kNN methods. The target variable in the analyses is the water permeability exponent x_{wp} . The predictor features consisted from the MS-NFI data, aerial gamma-ray spectroscopy, AEM data, topographical feature data, and peat bog mask data. Additional features such as windowed mean, windowed standard deviation, Gabor (Weldon et al., 1996) and local binary pattern (LBP, Pietikäinen et al., 2011) were derived from gamma-ray and AEM data. The MS-NFI data set describes the state of Finnish forests with attributes such as tree volume per hectare, number of trees per hectare, amount of ground vegetation, et cetera. The gamma-ray data indicates many significant characteristics of the soil, e.g. the tendency to stay moist. Similar to the gamma-ray, the AEM data gives information on various kinds of soil conductors. Topographical data included features such as local height difference, flow accumulation area, confluence and inclination (Schwanghart and Kuhn, 2010). Peat bog mask data is a boolean 1/0 raster data set describing the thickness of peat in the research area. A value 1 in the peat bog data indicates an area with peat thickness greater than 60 cm. A total of 1788 data points were available from Pomokaira research area. The data sets were provided by the GTK, NLS and LUKE.

Results and contribution to research questions

In Figure 3.2 is shown the prediction results for x_{wp} . In practical harvesting operations it is of interest to know how well a model performs if the predicted new data point is located r meters away from the closest known data. For this reason, we have plotted the prediction performance as the function of distance of test data to the closest known training data. With the baseline C-index value being 0.5 we see from the prediction results that the regression models perform pretty well, especially up until 100 meters prediction distance. The 6-nearest neighbor model has slightly better results than RLS. It is curious to notice here that if only the data coordinate information is used as features for the prediction model, then the difference is almost negligible to prediction model where we use the features also. This indicates there is a strong SAC within the data. The sparser spatial distribution of the data set also affects this phenomenon in the results since when comparing to the data set used in Publication I, the data points in the Pomokaira analysis had much greater average geographical distance between them.

The publication contributes to research question **(RQ1)** by providing quantitative empirical evidence for the use of natural resource data in terrain condition prediction.

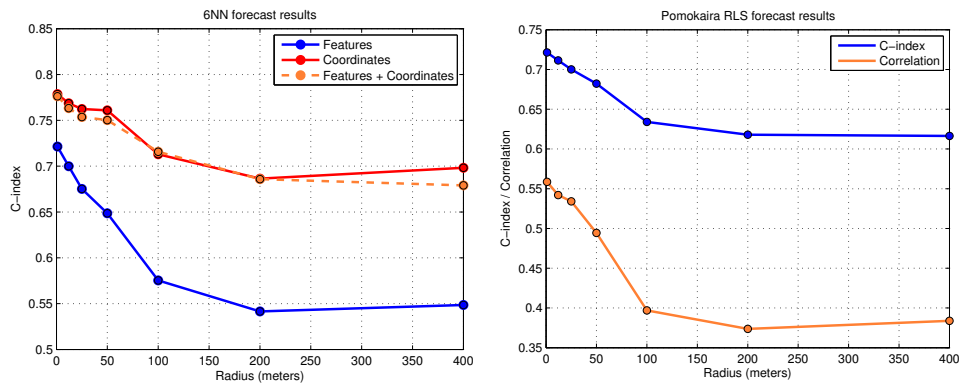


Figure 3.2: *Left*: prediction performance of 6-nearest neighbor for x_{wp} as the function distance of training data to test data. Different curves represent whether only features, coordinates or both were used in the prediction. *Right*: Same graph as in left, but for the RLS regression model and with only features used in prediction. In addition to the C-index, the RLS prediction performance is measured also using Pearson correlation.

Author's contribution

The author's responsibilities in the publication consisted from: preprocessing and merging of the data, modeling and evaluation of the results, and writing of the article. For more detailed explanations, see author's contribution in section 3.1.1 (same as in Publication I).

3.1.3 Publication III: Predictability of boreal forest soil bearing capacity by machine learning

Summary

In publication (Pohjankukka et al., 2016) is studied the prediction capability of soil penetration resistance and soil damage type by using a combination of open RS and manually on-site collected data. Terrain trafficability is a key factor to be considered in order to achieve successful route planning for forest harvest operations. Implementing the harvest at correct timing is crucial since badly timed operations have greater chances to cause large economical costs and excessive ecological damage. In poor soil conditions, besides having the risk of getting stuck to wet soil, the forest harvesters can damage trees unscheduled to be harvested. Damaging the roots of the trees can lead to fungal infections and in the worst case to tree decay. According to (Pennanen and Mäkelä, 2003), it is estimated that a yearly costs of 100 million euros in Finland alone result due to timber procurement causes. By

providing optimized route plans using ML approach to predict terrain soil conditions these costs could be significantly reduced. Linear and nonlinear prediction methods are used in the publication together with a validation method which we call *leave-one-out cross validation with a dead zone* (LOOCVDZ). The research area is located in the province of Eastern Finland, Pieksämäki.

Methods and data

Regression analysis was conducted using RLS, multilayer perceptron (MLP), MLP early-stopping committee (MLP-ESC) and kNN methods. LOOCVDZ was used for estimating the prediction performance as a function of geographical distance between the prediction point and closest known training data. The target variables in this analysis consist from soil damage type and penetration resistance. The soil damage type is specified by an integer value which defines the damage occurred to soil when being exposed to pressure from a forest harvester. The soil penetration resistance data was collected using a penetrometer (Muro and O'Brien, 2004) in the research area. A penetration resistance measurement is made by dropping a penetrometer on the soil point, pressing it against the soil, and then recording the depth of the resulting hole. The predictor data sets consisted from: MS-NFI data, DEM data, weather data, aerial gamma-ray data, peatland data, subsoil and topsoil data, and soil moisture data (for more information on MS-NFI and gamma-ray data see Publications I, II). The DEM data is constructed using airborne laser scanning (ALS) techniques. Several derived geomorphometric features from DEM were used: plan curvature, profile curvature, slope, topographic wetness index, flow area, aspect, diffuse insolation and direct insolation (Zevenbergen and Thorne, 1987; Wood, 1996, 2009; Beven and Kirkby, 1979; Seibert and McGlynn, 2007). The weather data consists from temperature and rainfall information from years 2011-2013. The peatland data is derived from NLS topographic database covering the whole of Finland. Subsoil and topsoil data are pre-classified raster data sets describing soil class types from Pieksämäki target area. The soil moisture data was measured by calculating the weight difference of soil samples before and after drying it. A total of 11795 data points were available in the analyses. The data sets were provided by LUKE, GTK, NLS and the FMI.

Results and contribution to research questions

The results for soil damage in Figure 3.3 show moderate prediction performance up to 20 meters. Approximately after 20 meters of prediction range the performance drops sharply to almost baseline levels. RLS and kNN were the best prediction methods in the soil damage case. For penetration resis-

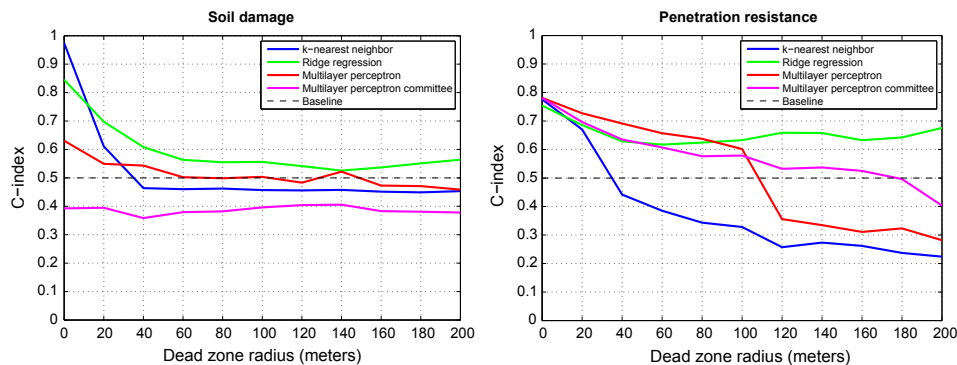


Figure 3.3: *Left*: prediction performance results for soil damage. *Right*: corresponding prediction results for penetration resistance.

tance MLP prediction models fared well against RLS and kNN. MLP had overall best results with up to 100 meter prediction range. Better results for the penetration resistance were to be expected since the data for it was more reliable than for the soil damage type. Soil damage data was based on expert assessment, whereas penetration resistance was based on on-site measurements. It was concluded in the included publication (Publication III), that a greater number of physical data samples with higher information value within the harvest machinery stand is needed, in order to produce sufficiently reliable ML-based prediction systems.

The publication contributes mostly to research question **(RQ1)** by providing quantitative results for the applicability of natural resource data to terrain condition prediction. Contributions are also made to research question **(RQ2)** since first steps are laid to the method discussed in Publication IV.

Author’s contribution

The author’s responsibilities in the publication consisted from: preprocessing and merging of the data, modeling and evaluation of the results using LOOCVDZ, and writing of the article. For more detailed explanations, see author’s contribution in section 3.1.1 (similar tasks as in Publication I).

3.1.4 Publication IV: Estimating the prediction performance of spatial models via spatial k-fold cross validation

Summary

The publication (Pohjankukka et al., 2017) introduces a novel CV method for spatial models, which is applied to the three research areas introduced

in Publications I-III. New predictor data sets are used together with the ones used in earlier publications and a more closer examination to the SAC inherent in the data sets is conducted. In forestry, harvest route selection via an ML based recommendation system would involve making soil point predictions on the corresponding route, and the goodness of this route would then be evaluated using these predictions. In practice, this recommendation system would need to make predictions from the forest harvester’s current geographical location (using known close by data) when doing online evaluation of the route alternatives. Due to SAC, an optimistic bias is involved into the prediction points which are far away from the harvester’s current location. This can cause the recommendation system to suggest potentially unsafe route possibilities. The publication seeks to tackle this problem by reducing the effect of SAC in point predictions by eliminating training data points in a suitable way, making the route evaluations more realistic.

Methods and data

The publication uses kNN as the prediction method in all the analyses. Since the prediction method itself has no effect on the presence of SAC in the data kNN was a natural selection due to its simplicity. The data sets in this study consist from the data used in Publications I-III with additional data included into Parkano and Pieksämäki cases. The supplementary data sets are the MS-NFI attributes and stoniness data for Parkano and Pieksämäki cases respectively. The stoniness data describes the approximated amount of stones in a soil point by using steel-rod sounding (Tamminen, 1991). In steel-rod

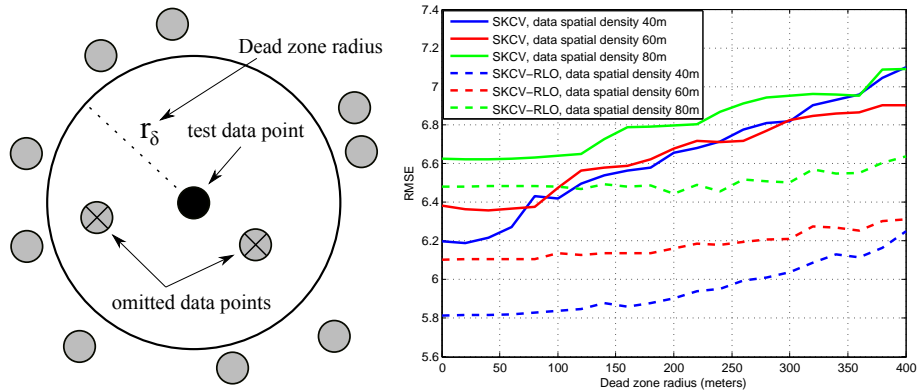


Figure 3.4: *Left:* Illustration of the SKCV procedure. *Right:* SKCV results (solid lines) for Parkano data with *root mean squared error* (RMSE) measure. The spatial density refers to the sparsity of data. The dashed lines represent analogous results for SKCV-RLO which confirms that SAC indeed affects the results.

sounding, a rod is pushed into the soil after which the penetration depth and the number of stone hits is recorded. Semivariograms and Moran's I statistics were used for confirming the presence of SAC in the data sets. In the publication, the SKCV method was introduced for measuring the prediction performance of a model dealing with a spatially autocorrelated data. The idea of SKCV is shown in the left side image of Figure 3.4. Training data points within a geographical radius of r_δ are omitted from the training data in order to simulate a scenario, where a prediction needs to be made to a geographical location so that the distance of training data to the prediction point is at least r_δ meters. In Figure 3.5 is shown a hypothesized practical scenario of this where a forest harvester makes routing decisions based on point predictions. In the presence of SAC, the prediction performance decreases as the distance to the prediction point increases. Furthermore, to confirm that the decrement to prediction performance as distance r_δ increases is truly because of SAC, and not simply because of omitting data points, a modified version of the SKCV called *spatial k-fold cross validation random-leave-out* (SKCV-RLO) is implemented. SKCV-RLO is identical to SKCV with the exception that instead of removing training data points within r_δ meters from the prediction point, we remove them randomly the same amount as we would with the SKCV. If SAC is present in the data, then SKCV-RLO should perform better. The publication also discusses on how the SKCV can be applied as a heuristic for data sampling density selection.

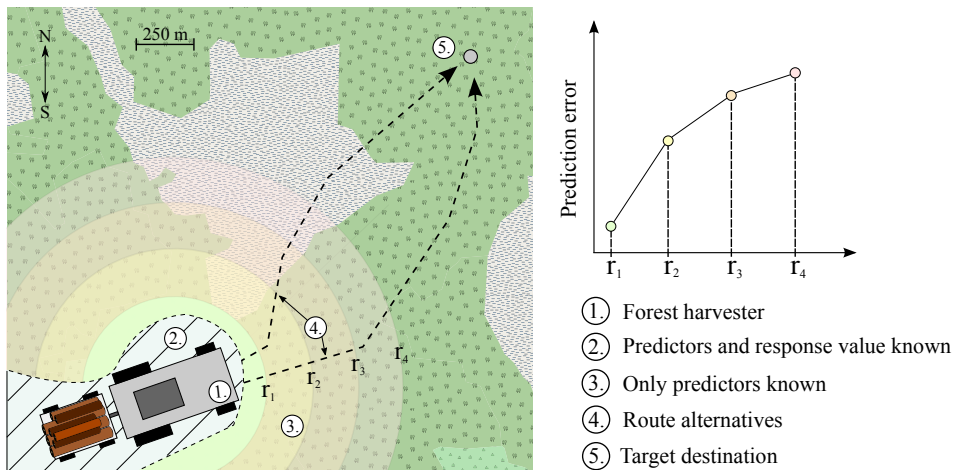


Figure 3.5: The forest harvesting example. The harvester driver needs to select an optimal route to target destination. Because of SAC, the prediction error increases the further away we make point predictions.

Results and contribution to research questions

In the right side of Figure 3.4 we see the SKCV and SKCV-RLO results for Parkano research area. The prediction performance clearly decreases as we increase the prediction distance r_δ for all spatial densities of the data. The spatial density refers to the sparsity of the data set when a hexagonally sampled data is used (for details, see Pohjankukka et al., 2017). When geographically sparser data is used the effect of prediction distance r_δ decreases (as can be seen in the results of Figure 3.4). This is due to the reason that as we decrease the spatial density of the data (resulting in a sparser data set), pairwise geographical distances between data points increases. As data points are farther away from each other geographically, small values of r_δ do not result in large removal of data points in the SKCV. This is especially true when the prediction distance r_δ is smaller than the average pairwise geographical distance between data points. On the other hand, if the data set is spatially very dense, then the effect of r_δ is much larger because more data is removed in the SKCV. The behavior of the results was similar in all three research areas and the SKCV-RLO results confirmed that SAC indeed creates an optimistic bias into the prediction results, making SKCV a relevant tool for spatial models.

The publication contributes to research question **(RQ2)** by proposing a novel SKCV method for measuring the prediction performance of spatial models.

Author's contribution

The author's responsibilities in the publication consisted from: modeling and evaluation of the results, inspection of SAC in the data, applying the SKCV method to data, and writing of the article. The SAC in the data was investigated using semivariogram and Moran's I statistics. Matlab and Python environments were used for implementing the modeling and evaluation parts. For more details, see author's contribution in section 3.1.1 (similar tasks as in Publication I).

3.1.5 Publication V: Comparison of estimators and feature selection procedures in forest inventory based on airborne laser scanning and digital aerial imagery

Summary

In publication (Pohjankukka et al., 2018), ALS and digital aerial imagery (DAI) data are used for predicting forest inventory attributes. Feature selection is also implemented to find the most relevant ALS and DAI features needed for predicting forest inventory attributes. In order to manage forest

resources efficiently, it is required to have accurate information on forest attributes in the form of thematic maps. Creating these maps manually throughout large forests is obviously both highly expensive and laborious. Instead of collecting samples manually, a RS based approach is preferred by using an estimator model for the forest attributes, where features derived from ALS and DAI data are used as predictors. ALS is currently considered being the most accurate RS data for estimating forest attributes according to (Næsset, 2002, 2004; Maltamo et al., 2006). By combining ALS and DAI data sets it is possible to accurately estimate forest inventory attributes. GA (see e.g. Van Coillie et al., 2005) and GFS methods are used in the feature selection procedure together with linear and nonlinear prediction methods used as estimators. The study was conducted with data from Åland province, Finland. Related study about the minimum number of covariates and optimal number of field data in Bayesian estimation context of predicting forest inventory variables is studied in the work of (Junttila et al., 2015).

Methods and data

The prediction methods used in the publication include kNN, RLS and MLP-ESC. In addition to GA and GFS feature selectors, a nested version of GFS called *nested greedy forward selection* (NGFS) is used. NGFS is almost identical to GFS with the exception that an extra outer loop is created similarly as in NCV. In NGFS, feature selection is first implemented using training and validation data sets $\mathcal{D}_t, \mathcal{D}_v$, and the corresponding selected features are then evaluated using a test data set \mathcal{D}_e . NGFS provides the means to study the stability of the feature selection process because of the extra outer loop. Data balancing prior to feature selection is also tested due to highly skewed distributions in some of the target variables. A total of seven forest attributes were subject to prediction: tree diameter, all trees; tree height, all trees; tree basal area, all trees; tree volume, all trees; tree volume, pine trees; tree volume, spruce trees and tree volume, broadleaf trees. The predictor features consisted from 154 ALS and DAI variables such as height above ground, Haralick texture features (Haralick et al., 1973), rgb- and color-infrared images. A total of 10 different analyses were implemented with different prediction method, CV, feature selector and data balancing combinations.

Results and contribution to research questions

The results of the study indicated that around a maximum of 20-40 features were sufficient to reach optimal prediction capability in all cases. Most reliable prediction results were obtained for tree height (all trees) and least

reliable for tree volume (spruce and broadleaf trees). The low prediction performance for the two last mentioned target variables is partly explained by the low number of such trees in Åland research area (i.e. low number of representative data points). The low prediction performance for tree volume attributes was also noted as instability in the NGFS feature selection. That is, for tree volume there seems to be no clear best features to be identified when looking at the NGFS results. In the top plots of Figure 3.6 we can see the GFS and NGFS results for the target variable tree height (h). In this case, we note that MLP-ESC is the best estimator with optimum prediction performance achieved with less than 10 features. In the bottom plots of Figure 3.6 we see the corresponding results for tree volume, spruce

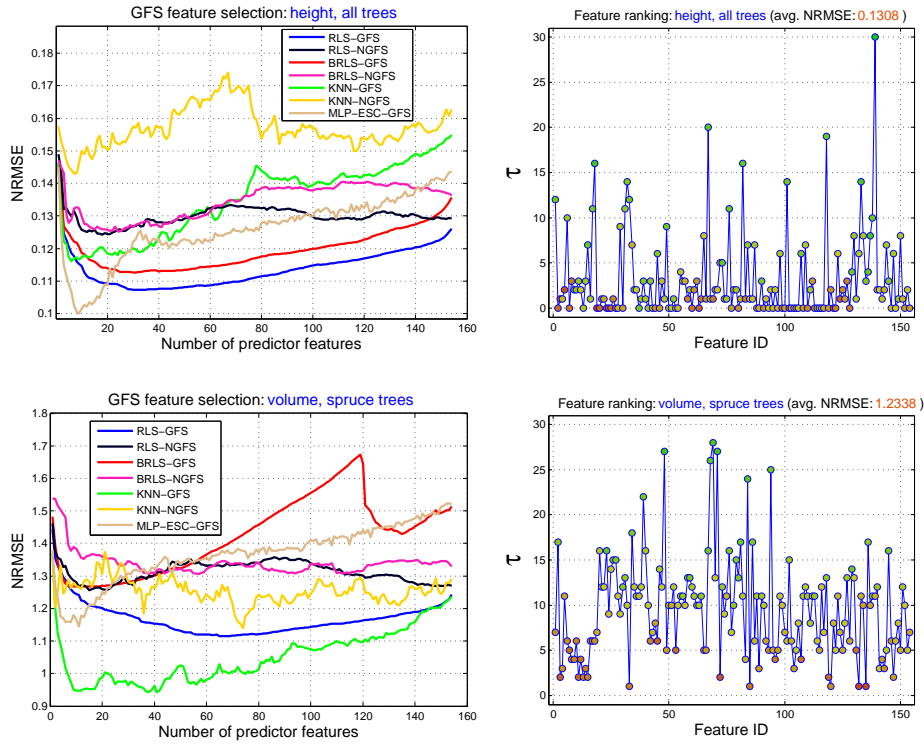


Figure 3.6: *Top-left:* the GFS results for tree height. The prediction performance measured with *normalized root mean squared error* (NRMSE) is plotted as the function of predictor features. *Top-right:* the corresponding feature selection results for NGFS. Here $\tau =$ "the number of times the feature was selected to the feature set which produces optimal prediction performance". The Feature ID refers to the identifier number of the specific feature. *Bottom- left and right:* The same graphs as above but for tree volume (spruce).

(v_s). Firstly, we notice that tree volume has a much higher prediction error than height target variable. Secondly, by comparing the feature selection results of NGFS we notice much more variability with tree volume than with tree height. One can see that the feature selection results are more definite for height variable and more random for volume variable. This observation together with the prediction error results suggests the absence of clear pattern in the data for tree volume target variable. According to the figures, features 139 (h_DE, selected 30 out of 30 times), 67 (i60f, selected 20 out of 30 times) and 118 (nir_ASM, selected 19 out of 30 times) were the top three features for h . The corresponding features for v_s were 69 (i20l, selected 28 out of 30 times), 48 (d3l, selected 27 out of 30 times) and 71 (i60l, selected 27 out of 30 times). For complete list of the features with descriptions see the Publication V included into this thesis. It must lastly also be mentioned, that while the prediction of volume target variable was difficult in this study, it is widely known in the literature (see works e.g. in Maltamo et al., 2014) that species-wise prediction is a more challenging problem than prediction over all tree species when using ALS data. This is due to the fact that laser scanning does not easily differentiate tree species from one another.

The publication contributes to research question **(RQ1)** by providing quantitative results on the predictability of MS-NFI attributes using ALS and DAI predictor features. Furthermore, feature selection results are given for obtaining optimal MS-NFI estimation.

Author's contribution

The author's responsibilities in the publication consisted from: preprocessing and merging of the data, modeling and evaluation of the results, feature selection, and writing of the article. Most of the tasks here were similar to those in Publication I (section 3.1.1). Feature selection was implemented using self-made and LUKE provided code libraries.

3.1.6 Publication VI: Reliable AUC estimation of spatial classifiers, with application to mineral prospectivity mapping

Summary

In the publication (Airola et al., 2018), a study was conducted for estimating orogenic gold mineral occurrences using RS data from Central Lapland, Finland. A model performance estimation method is proposed utilizing the CV methods proposed in the works of (Airola et al., 2009, 2011; Pohjankukka et al., 2017). Mineral prospectivity mapping (MPM) techniques are used to delineate areas favorable for mineral exploration. By combining information

from geospatial, geophysical and geochemical data sets the MPM can be used for estimating the likelihood of mineral presence within a research area. In practical applications such as this, the assumption of i.i.d. data samples is usually not valid which causes model performance estimation methods to produce either negatively or positively biased statistics. The pooling procedure, performed by methods such as LOOCV can introduce a substantial negative bias to these statistics. SAC on the other hand can cause a positive bias as we have seen in (Pohjankukka et al., 2017). The publication proposes the *leave-pair-out spatial cross validation* (LPO-SCV) method that corrects both of these biases in the performance estimates.

Methods and data

Multiple linear and nonlinear classifiers are used in the publication. SVM, logistic regression and RLS methods are used as linear classifiers, and for nonlinear classifiers kNN and random forest methods (see e.g. Hastie et al., 2001; Breiman, 2001a) are applied. Prediction performance is estimated using the proposed LPO-SCV method for removing the bias caused by both data pooling and SAC. In the left side of Figure 3.7 is presented the main idea behind LPO-SCV. On each CV round, both positive and negative test instances corresponding to data instances with opposite label values are left out (for more details see also Pahikkala et al., 2008), as well as the data instances within r_δ distance away from them. This procedure is repeated for all possible positive-negative pairs. The LPO-SCV therefore simulates a scenario where the left out test pair is at least r_δ distance away from the training data. The target variable for prediction in this analysis was the orogenic gold occurrence, a real value indicating the presence of gold mineral in a soil point. In the included publication (Airola et al., 2018), positive-negative pair in LPO-SCV corresponds to positive and negative gold occurrence label values. As predictor features, raster images derived from airborne and ground based geophysics, till geochemistry and geological interpretations are used. The predictor features consist from the same features as generated in the work of (Nykänen, 2008). These features consists from mineral exploration related geoscientific spatial data sets that are derived from airborne geophysics (magnetic and EM), regional till geochemistry, ground geophysics (gravity) and a scale digital geological map. The data sets are provided by GTK and the FGI.

Results and contribution to research questions

In the right of Figure 3.7 is shown a comparison of LPO-SCV and *leave-one-out spatial cross-validation* (LOO-SCV) results. LOO-SCV is simply the leave-one-out case of SKCV. From the prediction performance results it was

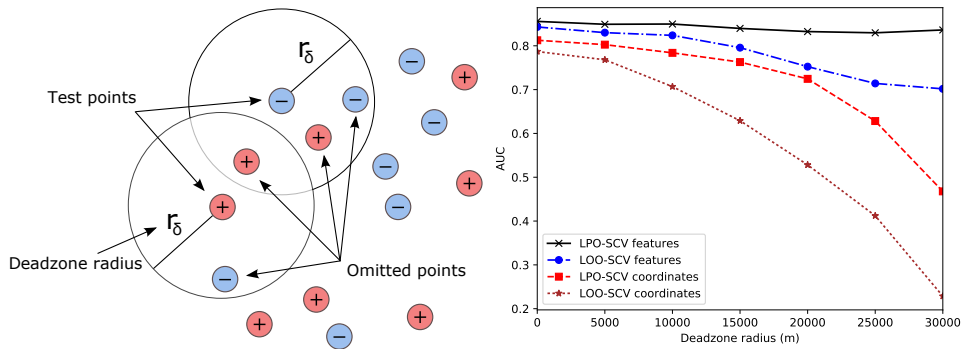


Figure 3.7: *Left:* Illustration of the LPO-SCV. On a single CV round, a positive-negative pair as well as the data instances within r_δ distance away from them is left out. *Right:* Comparison of LPO-SCV and LOO-SCV results for kNN ($k = 250$) using the regular feature set, and only coordinates as features.

clearly noted the presence of both the pooling and spatial biases. It is shown in the publication that high prediction performance can be obtained even with classifiers completely failing in generalizing outside the training data if the spatial dependencies are not taken into account. Simple linear models turned out to work very well in this analysis, which is likely to be result of a highly imbalanced data set (small amount of positive gold data instances) that was in disposal. Nevertheless, the results demonstrated that prediction performance estimation is highly dependent on the validation strategy. The research community involved with spatial data analysis is encouraged to provide thorough spatial CV evaluations that reflect the characteristics of the data and the model performance under the corresponding application, which the classical model validation methods are not able to provide.

The publication contributes to research question **(RQ2)** by providing quantitative results on the effects of pooling and SAC biases in model evaluation. The LPO-SCV method is presented to counter these biases.

Author's contribution

The author's responsibilities in the publication consisted from: modeling and evaluation of the results, inspection of SAC in the data, applying the SKCV method to data, and writing of the article. Author's tasks were similar to those in Publication IV (section 3.1.4).

3.2 Research results

This section summarizes the main research questions and evaluates the corresponding research results. In what follows, each research question is introduced and briefly motivated, which is then succeeded by a discussion of the research results respectively.

3.2.1 (RQ1): Are the provided open natural resource data sets applicable in predicting terrain conditions and forest attributes in Finland?

The first motivating question regarding this thesis focuses on the quality of the provided open natural resource data for estimating forest and terrain conditions. The data sets range from satellite and airborne imagery to manually collected samples. Since the collection of these data sets result in yearly costs for many institutions, it is of interest to study whether additional value for these data sets could be gained by utilizing them in other than their main intended applications, and hence increasing the benefit-cost ratio of the data. With respect to the corresponding application, this information can also be used to better focus the data collection processes, i.e. to concentrate only on relevant data when irrelevant data has been recognized. Furthermore, the answer to **(RQ1)** also gives insight if the data sets need to be modified either by changing the amount of collected data, or by collecting data of higher information value.

Research results for (RQ1)

The empirical results of the included publications (Pohjankukka et al., 2014a,b, 2016, 2017, 2018; Airola et al., 2018) indicate a moderate prediction performance for estimating terrain conditions and forest attributes with the provided data sets. In order to achieve successful and safe applications using the ML approach, natural resource data with higher information value should be investigated. Moderate prediction performance is not sufficient especially for applications which require high safety standards, such as forest harvesting in peatland areas. Peatland areas pose a threat for heavy machinery (see e.g., Pohjankukka et al., 2017) so accurate prediction performance is critical in these applications. Prediction performance could be improved by using larger and higher resolution data sets with samples gathered more densely and from a wider set of geographical areas. Having e.g. data sets with uniform two meter resolution would increase the information utility value of the data. At this resolution the effect of individual trees and their roots can be detected. Collecting RS and field measurement data also from a wider set of geographical areas will better guarantee that

the data represents large area phenomena and not just a local phenomenon. This clearly allows higher probability for generalization. Feature selection for estimating forest attributes showed good results in the Åland case (Pohjankukka et al., 2018), revealing around 74-87% of the predictor features unnecessary for accurate estimation. The feature selection results differed of course for distinct forest attributes, but nevertheless showing overall that the amount of predictor features can be reduced with no loss in prediction performance. In fact, it was seen that the addition of too many features results in decreased prediction performance. Low performance for predicting tree species-specific volumes in the Åland case was also partly explained by the fact that Åland area has a relatively small distribution of trees, resulting in low number of representative data points. Therefore, the data information value is also low.

3.2.2 (RQ2): How to evaluate the prediction performance of a model involving spatially dependent data?

The second research question was motivated by the application of natural resource data in forest harvesting operations. As it was discussed earlier, spatially distributed natural data sets always contain SAC. This means that models which apply these natural data sets should take into account the inherent SAC in the data in order to prevent optimistic prediction performance estimation. Considering practical applications, it is therefore not enough to use model goodness of fit measures which heavily rely on the i.i.d. assumption of the data. A model performance evaluation method taking into account the SAC is therefore needed, and the model validation should reflect the intended application. In the forest harvesting example this is especially important since safety and efficiency are crucial factors and biased estimations are not wanted.

Research results for (RQ2)

The included publications (Pohjankukka et al., 2016, 2017; Airola et al., 2018) tackled the problem of SAC in model evaluation by proposing the SKCV method. This method was based on adjusting the training data set in CV procedure to simulate the scenario where prediction point is always within certain spatial distance from the training data. The research results revealed that indeed geospatial data sets contain SAC, which was shown both in the SAC measures and in the prediction results. These results confirmed that goodness of fit measures accounting for SAC should be considered in order to give realistic model performance estimates. Standard methods for evaluation models such as IC and CV should therefore not be used in their basic form, but rather should be modified slightly to suit the

requirements and nature of the corresponding application and data respectively. Disregarding the inherent spatial dependencies in the data can result in too optimistic models, a situation not wished for especially in applications with high potential damages and expenses. Even though simple, the SKCV method provides insight to the limit a spatial model can be trusted, after which further data should be collected in order to extend the spatial prediction range.

Chapter 4

Conclusions

4.1 Summary of the thesis

In Chapter 1 we started by going through the introduction, motivation and the resulting research questions of this thesis. It was discussed that in the era of big data available from many domains, there is a rising need for data-driven applications using ML approach. Big data comes from a large variety of sources ranging from satellite imagery to manual data collection. Various institutions collect data for many different purposes such as large-area strategic forest planning (MS-NFI) or for forest harvest planning. Since the data collection processes inherently produce costs, it is of value to know the usefulness of these data sets. By providing quantitative measures indicating the applicability of these data sets, we can enhance both the data collection processes and the applications exploiting these data sets. In Chapter 2 the theoretical background for this thesis was provided. An introduction to GIS systems, RS techniques and data representation formats were covered. The SAC dependency issue involved in spatial data analysis was explained, which is not taken into account by many classical model fitness criteria. Lastly, a coverage to ML paradigm and methods was given. Chapter 3 presented the summaries and research results of the corresponding publications included into this thesis. For each publication, a motivation for the research was given together with a compact description of analysis details and research findings. Finally, the main research questions of this thesis were revisited and the corresponding results to these questions were summed up. The purpose of this chapter is to provide a concluding discussion and present the main outcomes of the research.

4.2 Discussion and outcomes

Data-driven decision making is on its way to become a common tool to be used in the future. As data sources continue to increase, there is no doubt that data processing and inference via artificial intelligence systems is needed, both now and in the future. This thesis was motivated by the use and availability of open natural resource data in forestry applications (e.g. resource management and forest harvesting). We discussed in earlier chapters how the utilization of natural resource data with ML approaches can help gain significant benefits in practical applications. These benefits include e.g. saves in operation costs and increments in safety. The applicability of the provided data sets was inspected in the included publications showing moderate prediction performances in the corresponding applications. The moderate performance can be explained partly by the low information value and amount of the data sets. The performance could be improved by providing larger amounts of higher quality data sets (e.g. more information value with higher resolution). Improvement could be obtained in this way since larger and more accurate data sets are more likely to contain useful information. Furthermore, the publications also revealed that SAC is an important factor to be considered in applications involving spatial models. The SKCV method taking into account the effects of SAC was used for estimating the performance of these models. To summarize, the research findings of this work can be listed as the following main outcomes:

1. Prediction of terrain conditions showed moderate performance (Publications I-III). Larger and higher quality data sets from a wider set of research areas should be investigated in order to improve the prediction performance.
2. Feature selection should be implemented on all the currently available natural resource data sets in order to recognize relevant and irrelevant features. Research results showed that in the Åland case (Publication V) at most 87% of the predictor features are unnecessary for obtaining optimal prediction performance.
3. Model evaluation methods should be designed to reflect the corresponding application. As it was shown in earlier chapters, SAC dependency is inherently present in spatial data sets, which is not accounted for in many classical model goodness of fit measures. The SKCV method was proposed for estimating the prediction performance of spatial models (Publications IV, VI). Results indicated that SAC is indeed present in the data and can cause significant bias into the prediction estimates.

As a final note, some mentions on the limitations and future work are in order. The research studies conducted in the attached publications can be improved by testing yet more methodologies and evaluation measures. Due to resource limitations however, this was not possible to implement in this work and remains to be conducted in future research. The research results of this thesis can also be improved as additional and higher quality data sets are obtained in the future with the advancements in RS imaging technologies.

References

- Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning From Data*. AMLBook.
- Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. (2009). A comparison of AUC estimators in small-sample studies. In Džeroski, S., Geurts, P., and Rousu, J., editors, *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pages 3–13, Ljubljana, Slovenia. PMLR.
- Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., and Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics and Data Analysis*, 55(4):1828–1844.
- Airola, A., Pohjankukka, J., Torppa, J., Middleton, M., Nykänen, V., Heikkonen, J., and Pahikkala, T. (2018). Reliable AUC estimation of spatial classifiers, with application to mineral prospectivity mapping. *Data Mining and Knowledge Discovery*. Accepted.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*. Academiai Kiado.
- Akaike, H. (1980). Likelihood and the bayes procedure. *Trabajos de Estadística Y de Investigacion Operativa*, 31(1):143–166.
- AMDAR (2018). The WMO AMDAR Observing System. <https://www.wmo.int/pages/prog/www/GOS/ABO/AMDAR/>. Online; accessed 15 January 2018.
- Anderson, D. R. (2008). *Model Based Inference in the Life Sciences*. Springer.
- Bakker, W. H., Feringa, W., Gieske, A. S. M., Gorte, B. G. H., Grabmaier, K. A., Hecker, C. A., Horn, J. A., Huurneman, G. C., Janssen, L. L. F., Kerle, N., van der Meer, F. D., Parodi, G. N., Pohl, C., Reeves, C. V., van Ruitenbeek, F. J., Schetselaar, E. M., Tempfli, K., Weir, M. J. C., Westinga, E., and Woldai, T. (2009). *Principles of Remote Sensing: an introductory textbook*. ITC Educational Textbook Series 2, 4th edition. University of Twente, Faculty of Geo-Information and Earth Observation (ITC).
- Bazaraa, M. S. (2013). *Nonlinear Programming: Theory and Algorithms*. Wiley Publishing, 3rd edition.

- Beven, K. J. and Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24(1):43–69.
- Bishop, C. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.
- Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.
- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference*. Springer-Verlag New York, 2nd edition.
- Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209.
- Cressie, N. (2015). *Geostatistics*. John Wiley & Sons, Inc.
- Earth observation portal (2018). Satellite Missions Database. <https://directory.eoportal.org/web/eoportal/satellite-missions>. Online; accessed 17 January 2018.
- European space agency, earth portal (2018). ESA Earth Observation Missions. <https://earth.esa.int/web/guest/missions>. Online; accessed 17 January 2018.
- Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A*, 222:309–368.
- Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725.
- Fletcher, T. (2010). Relevance vector machines explained.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

- Grünwald, P. D. (2007). *The Minimum Description Length Principle (Adaptive Computation and Machine Learning)*. The MIT Press.
- Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.
- Hyvönen, E., Päänttjä, M., Sutinen, M.-L., and Sutinen, R. (2003). Assessing site suitability for scots pine using airborne and terrestrial gamma-ray measurements in finnish lapland. *Canadian Journal of Forest Research*, 33(5):796–806.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.
- Junttila, V., Kauranne, T., Finley, A. O., and Bradford, J. B. (2015). Linear models for airborne-laser-scanning-based operational forest inventory with small field sample size and highly correlated LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5600–5612.
- Junttila, V., Maltamo, M., and Kauranne, T. (2008). Sparse bayesian estimation of forest stand characteristics from airborne laser scanning. *Forest Science*, 54(5):543–552.
- Konishi, S. and Kitagawa, G. (2007). *Information Criteria and Statistical Modeling*. Springer Publishing Company, Incorporated, 1st edition.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- Landsat (2018). The Landsat Program. <https://landsat.gsfc.nasa.gov/>. Online; accessed 11 January 2018.
- Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., and Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23(7):811–820.
- Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer Verlag.

- Longley, P., Goodchild, M., Maguire, D., and Rhind, D. (2005). *Geographic Information Systems and Science*. Wiley, 2nd edition.
- Mäkisara, K., Katila, M., Peräsaari, J., and Tomppo, E. (2016). The multi-source national forest inventory of finland methods and results 2013. Technical report, Natural Resource Institute Finland.
- Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., and Kangas, J. (2006). Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Canadian Journal of Forest Research*, 36(2):426–436.
- Maltamo, M., Næsset, E., and Vauhkonen, J. (2014). *Forestry Applications of Airborne Laser Scanning*, volume 27 of *Managing Forest Ecosystems*. Springer Netherlands, 1st edition.
- Marsland, S. (2014). *Machine Learning: An Algorithmic Perspective, Second Edition*. Chapman & Hall/CRC, 2nd edition.
- Mukhopadhyay, N. (2000). *Probability and Statistical Inference*. CRC Press.
- Muro, T. and O’Brien, J. (2004). *Terramechanics: Land Locomotion Mechanics*. CRC Press.
- Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*, 80(1):88–99.
- Næsset, E. (2004). Accuracy of forest inventory using airborne laser scanning: evaluating the first nordic full-scale operational project. *Scandinavian Journal of Forest Research*, 19(6):554–557.
- Neal, R. M. (1994). *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, Canada.
- NLS (2014). *NLS Topographic database*. National Land Survey of Finland.
- Nykänen, V. (2008). Radial basis functional link nets used as a prospectivity mapping tool for orogenic gold deposits within the central Lapland greenstone belt, northern Fennoscandian shield. *Natural Resources Research*, 17(1):29–48.
- Pahikkala, T. and Airola, A. (2016). Rlscore: Regularized least-squares learners. *Journal of Machine Learning Research*, 17(221):1–5.
- Pahikkala, T., Airola, A., Boberg, J., and Salakoski, T. (2008). Exact and efficient leave-pair-out cross-validation for ranking RLS. In Honkela, T.,

- Pöllä, M., Paukkeri, M.-S., and Simula, O., editors, *Proceedings of the 2nd International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'08)*, pages 1–8, Espoo, Finland. Helsinki University of Technology.
- Pahikkala, T., Airola, A., and Salakoski, T. (2010). Speeding up greedy forward selection for regularized least-squares. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 325–330.
- Pennanen, O. and Mäkelä, O. (2003). Raakapuukuljetusten kelirikkoheittojen vähentäminen, metsätehon raportti. Technical report, Metsäteho Ltd.
- Pietikäinen, M., Hadid, A., Zhao, G., and Ahonen, T. (2011). *Computer Vision Using Local Binary Patterns*. Springer Publishing Company, Incorporated, 1st edition.
- Pohjankukka, J., Nevalainen, P., Pahikkala, T., Hyvönen, E., Middleton, M., Hänninen, P., Ala-Ilomäki, J., and Heikkonen, J. (2014a). Predicting water permeability of the soil based on open data. In Lazaros, I., Ilias, M., and Harris, P., editors, *Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014)*, volume 436 of *IFIP Advances in Information and Communication Technology*, pages 436–446. Springer.
- Pohjankukka, J., Nevalainen, P., Pahikkala, T., Hyvönen, E., Sutinen, R., Hänninen, P., and Heikkonen, J. (2014b). Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. In Borga, M., Heyden, A., Laurendeau, D., Felsberg, M., and Boyer, K., editors, *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014)*, pages 1822–1827. IEEE.
- Pohjankukka, J., Pahikkala, T., Nevalainen, P., and Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019.
- Pohjankukka, J., Riihimäki, H., Nevalainen, P., Pahikkala, T., Ala-Ilomäki, J., Hyvönen, E., Varjo, J., and Heikkonen, J. (2016). Predictability of boreal forest soil bearing capacity by machine learning. *Journal of Terramechanics*, 68:1–8.
- Pohjankukka, J., Tuominen, S., Pitkänen, J., Pahikkala, T., and Heikkonen, J. (2018). Comparison of estimators and feature selection procedures in forest inventory based on airborne laser scanning and digital aerial imagery. *Scandinavian Journal of Forest Research*. Accepted.

- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.
- Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry Theory*. World Scientific Publishing Co., Inc., River Edge, NJ, USA.
- Satellite imaging corporation (2018). Satellite Sensors. <https://www.satimagingcorp.com/satellite-sensors/>. Online; accessed 17 January 2018.
- Schwanghart, W. and Kuhn, N. J. (2010). Topotoolbox: A set of matlab functions for topographic analysis. *Environmental Modelling and Software*, 25(6):770–781.
- Seibert, J. and McGlynn, B. L. (2007). A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models. *Water Resources Research*, 43(4).
- Sentinel-1 team (2013). *Sentinel-1 User Handbook*. European Space Agency.
- Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4).
- Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 4th edition.
- Shibata, R. (1989). *Statistical Aspects of Model Selection*, pages 215–240. Springer Berlin Heidelberg.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.
- Shumway, R. H. and Stoffer, D. S. (2005). *Time Series Analysis and Its Applications (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47.
- Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models (in japanese). *Mathematical science*, (153):12–18.
- Tamminen, P. (1991). Kangasmaan ravinnetunnusten ilmaiseminen ja viljavuuden alueellinen vaihtelu etelä-suomessa: expression of soil nutrient status and regional variation in soil fertility of forested sites in southern Finland. Technical Report 777.

- Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*. Academic Press, 4th edition.
- Tipping, M. E. (2000). The relevance vector machine. In Solla, S. A., Leen, T. K., and Müller, K., editors, *Advances in Neural Information Processing Systems 12*, pages 652–658. MIT Press.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240.
- Tomppo, E., Katila, M., Mäkisara, K., and Peräsaari, J. (2008). *Multi-source national forest inventory - methods and applications*, volume 18 of *Managing Forest Ecosystems*. Springer.
- Van Coillie, F., Verbeke, L., and De Wulf, R. (2005). GA-driven feature selection in object-based classification for forest mapping with IKONOS imagery in Flanders, Belgium. In Olsson, H., editor, *Proceedings of Forest-Sat 2005 in Borås May 31-June 3 : Skogsstyrelsen report 8b*, pages 11–15. National Board of Forestry.
- Vapnik, V. N. (1998). *Statistical Learning Theory*, volume 1. Wiley-Interscience.
- Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):364–372.
- Weitkamp, C. (2005). *Lidar, Range-Resolved Optical Remote Sensing of the Atmosphere*. Springer.
- Weldon, T. P., Higgins, W. E., and Dunn, D. F. (1996). Efficient gabor filter design for texture segmentation. *Pattern Recognition*, 29:2005–2015.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.
- Wood, J. (1996). *The geomorphological characterisation of digital elevation models*. PhD thesis, University of Leicester, England.
- Wood, J. (2009). Geomorphometry in landsurf. In Hengl, T. and Reuter, H. I., editors, *Geomorphometry*, volume 33 of *Developments in Soil Science*, pages 333–349. Elsevier.
- Zevenbergen, L. W. and Thorne, C. R. (1987). Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12(1):47–56.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

Publication I

Arctic Soil Hydraulic Conductivity and Soil Type Recognition Based on Aerial Gamma-Ray Spectroscopy and Topographical Data

Jonne Pohjankukka, Paavo Nevalainen, Tapio Pahikkala, Eija Hyvönen, Raimo Sutinen, Pekka Hänninen and Jukka Heikkonen. In Magnus Borga, Anders Heyden, Denis Laurendeau, Michael Felsberg, and Kim Boyer, editors, Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014), pages 1822–1827. IEEE, 2014.

Copyright © 2014 IEEE. Reprinted with permissions from respective publisher and authors.

Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data

Jonne Pohjankukka¹, Paavo Nevalainen¹, Tapio Pahikkala¹,
Pekka Hänninen², Eija Hyvönen², Raimo Sutinen² and Jukka Heikkonen¹

¹Department of Information Technology, University of Turku, FI-20014 Turku, Finland

²Geological Survey of Finland, FI-02151 Espoo, Finland

{Jonne.Pohjankukka, ptneva, Tapio.Pahikkala, Jukka.Heikkonen}@utu.fi

{Pekka.Hanninen, Eija.Hyvonen, Raimo.Sutinen}@gtk.fi

Abstract—A central characteristic of soil in the arctic is its load bearing capacity since that property influences forest harvester mobility, flooding dynamics and infrastructure potential. The hydraulic conductivity has the greatest dynamical influence to bearing capacity and hence is essential to measure or estimate. In addition, the arctic soil type information is needed in many cases, e.g. in roads and railways building planning. In this paper we propose a method for hydraulic conductivity estimation via linear regression on aerial gamma-ray spectroscopy and publicly available topographical data with derived elevation based features. The same data is also utilized for the arctic soil type recognition; both logistics regression and nearest neighbor classification results are reported. The classification results for logistic regression resulted in 44.5 % prediction performance and 50.5 % for 8-nearest neighbor classifier respectively. Linear regression results for estimating the hydraulic conductivity of the soil resulted in C-index value of 0.63. The hydraulic conductivity and soil type estimation results are promising and the proposed topographic elevation features are apparently new for remote sensing community and should also have a wider general interest.

I. INTRODUCTION

This paper is about predicting the soil type and its hydraulic conductivity by regression analysis using publicly available multi-source data. Soil type and the level of its granularity in arctic areas is of great interest to many different parties. The interested parties range from heavy industry to single consumers. For instance mining industry is interested on the type and granularity of the soil, in order to select the best strategy of placement for the mining machinery. Forest industry is interested on the load capacity of the soil, when placing forest harvest machinery on the areas of interest. Great caution is needed with the heavy machinery and accurate predictions for the soil type are required in order to avoid any accidents and minimize the moving costs. Swamp areas for example are a high risk for heavy machinery and predetermined knowledge of their locations is required.

We conduct a research on the usefulness of aerial gamma-ray spectroscopy data (referred later as gamma-ray data) combined with topographical height data when predicting the

qualities and characteristics of the soil, namely its type and hydraulic conductivity. 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.

We base our analysis on the gamma-ray data and topographical data provided by the Geological Survey of Finland (GTK) and the National Land Survey of Finland (NLS). Hydraulic conductivity of the soil can be estimated using the water permeability exponent provided by the GTK, which describes the rate of water flow speed in different soil types. Related studies have been conducted in the paper of A. Azzalini and J. Diggle [1] where they predict soil respiration rates from temperature, moisture content and soil type. Another related research was published in the paper of P. Scull, J. Franklin and O.A. Chadwick [12]. In their paper they use classification tree analysis for predicting the soil type in desert landscapes. Related work has been done also by P. Nevalainen et al. [10], R.P.O. Schulte et al. [13], H. Gao et al. [3] and Hyvönen et al. [5]. Other related research was conducted by R.A. Chapuis [2], R. Kiss [6], H.S. Mahmood et al. [7], N.J. McKenzie [8], I.D. Moore et al. [9], A.T. Ramli et al. [11] and J.V.A. Zachary [14]. The main novelty of this paper related to the previous studies is the proposed elevation features derived from topographical data and the use of aerial gamma-ray data.

II. THEORY

In our analysis we are going to use regression analysis for predicting the characteristics of the soil. Because regression analysis is a well-known approach to model the relationship between the explanatory variables x_1, \dots, x_p and dependent variable y , we are going to describe the used methods only briefly, namely regularized linear regression and logistic regression.

A. Regularized linear regression

For predicting the water conductivity of the soil we are using regularized least squares estimation. As it is well known, when

doing regularized linear regression we want to find a vector $\mathbf{w} \in \mathbb{R}^p$ such that the error function:

$$E(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \mathbf{w}^T \mathbf{x}_i \right)^2 + \frac{\lambda}{n} \mathbf{w}^T \mathbf{w}, \quad (1)$$

is minimized. In equation (1), $\mathbf{x}_i \in \mathbb{R}^p$ is the input data for i th observation, $y_i \in \mathbb{R}$ is the response value for i th observation, $n \in \mathbb{N}^+$ is the number of observations and $\lambda \in \mathbb{R}$ is the regularization parameter.

B. Logistic regression

Logistic regression is used for classifying the dependent categorical variable based on one or more predictor variables. The basic idea is to use the logistic function:

$$F(t) = \frac{1}{1 + e^{-t}}, \quad (2)$$

where $t = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ is a linear combination of the explanatory variables. The output of the logistic function is interpreted as a probability. The logistic function (2) is used in the case of binary classification, which is why in our study we use the multinomial version of logistic regression. Multinomial logistic regression is the generalization of logistic regression to allow arbitrary number of classes. In multinomial logistic regression the activation function is the softmax logit function:

$$P(y_i = j | \mathbf{x}_i) = \begin{cases} \frac{1}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{x}_i}}, & k = K \\ \frac{e^{\beta_j \cdot \mathbf{x}_i}}{1 + \sum_{k=1}^{K-1} e^{\beta_k \cdot \mathbf{x}_i}}, & 1 \leq k \leq K - 1 \end{cases}$$

where class $y_i = K$ is selected as the "pivot" element, β_j is the vector of regression coefficients associated with class j , \mathbf{x}_i is the set of explanatory variables associated with observation i . Observation \mathbf{x}_i is classified into class j^* such that:

$$j^* = \underset{j}{\operatorname{argmax}} P(y_i = j | \mathbf{x}_i).$$

The values for the sets of parameters β_1, \dots, β_K are solved using maximum a posteriori (MAP) estimation with quasi-Newton optimization algorithms.

III. TEST AREA AND DATA SETS

The research area is located in the municipality of Parkano, which is a part of the Pirkanmaa region. Pirkanmaa is located in the province of Western Finland. The size of the target area is $144,4804 \text{ km}^2$. The target area is located in ETRS-TM35FIN coordinates at 278 kmE, 6882 kmN, zone 35. Our data sets consist of aerial gamma-ray data and topographical height data. The data are 601×601 pixel images, with one pixel corresponding to a $20m \times 20m$ area. When considering all the derived features used in the analysis we get a total of nine images for input data and one image for output data. We now present our data sets, firstly gamma-ray data and its

derived features and secondly topographical height data and its derived features.

A. Aerial gamma-ray data

We conduct soil type predictions by using aerial gamma-ray spectroscopy and topographical height features. The aerial gamma-ray data is provided by the Geological Survey of Finland (GTK). This data is well suited for forest harvest applications, especially for applications in arctic areas such as Northern part of Finland. The naturally occurring chemical elements kalium (K) and thorium (Th) emit electromagnetic gamma-ray radiation of extremely high frequency. By using the gamma-ray radiation of kalium, we can infer many valuable characteristics of the soil. For instance, we can infer based on the gamma-ray radiation of kalium the level of humidity, roughness and frost heaving of the corresponding soil.

The intensity of gamma radiation is affected by the density, porosity, grain size and humidity of the soil. The amount of water in the soil affects inversely to the gamma radiation of kalium. From this we can infer that, the less gamma radiation is emitted from the area under study, the more there is water in the area. This is useful information for we can use it to classify whether the area under study is a suitable environment for pine trees or we can conduct forecasts for the load capacity and frost heaving of the soil. We can also use the gamma radiation for classifying the soil type.

The soil in the Parkano target area is classified into 10 different soil type classes by the GTK. We use five different features derived from the gamma-ray data: Gamma radiation intensity itself and windowed 3×3 mean, 5×5 mean, 3×3 standard deviation and 5×5 standard deviation. In Figure 1 we present the gamma-ray data from Parkano target area and the soil type classification. The brighter the area in the gamma-ray picture is, the more gamma radiation is emitted in the area and hence less water there is in the area. In Figure 2 we present the statistical features derived from gamma-ray data. As a side note one can notice that standard deviation detects well edges in an image, which is also intuitive. The areas near the edges have more variance than the uniform areas.

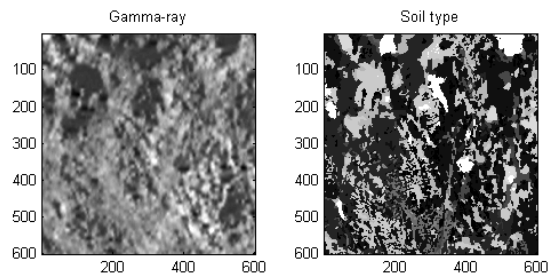


Fig. 1: Parkano aerial gamma-ray data (left) and soil type classification (right) of Parkano target area. In the gamma-ray picture one can note two larger dark areas. These correspond to swamp and lake areas.

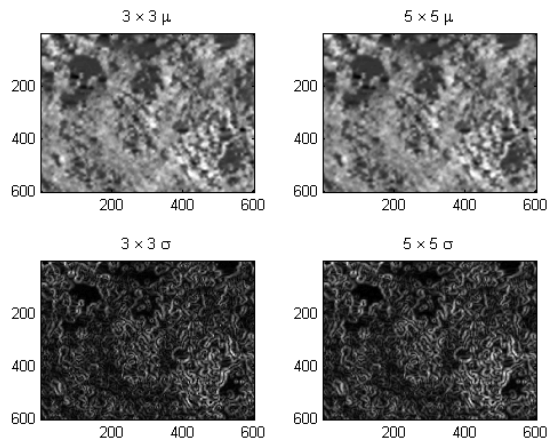


Fig. 2: Gamma-ray derived statistical features. 3×3 mean (upper-left), 5×5 mean (upper-right), 3×3 standard deviation (lower-left) and 5×5 standard deviation (lower-right). Standard deviation detects well the soil type boundaries.

B. Topographical height data

In addition to the gamma-ray data presented above we also used topographical data provided by the National Land Survey of Finland (NLS) in the analysis. There are several alternative attributes possible to derive from topographical height data. In our analysis we use ground inclination, convergence index and flow accumulation defined in the paper of R. Kiss [6]. Figure 3 depicts topographical height data and its three derived features.

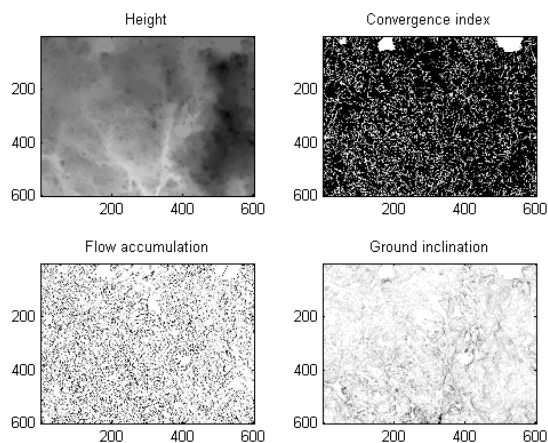


Fig. 3: Topographical height and its features. Height (upper-left), convergence index (upper-right), flow accumulation (lower-left) and ground inclination (lower-right).

C. Soil types

The GTK provided the analysis with two different types of soil data from Parkano target area: ground soil classification data and topsoil classification data. Both soil type data sets had the same soil type classification in approximately 92 % of the data. For this reason we used only ground soil classification data. The soil type data is represented by positive integer values, which indicate the pre-classified soil types. The total data points used for analysis consisted of total of 361201 data points, but because some of the soil types in the target area consisted of uninteresting types, such as peat production field, soil fill disposal site and water the overall number of used data points was 342479.

For the regularized regression we are using the so called water conductivity exponent, which is denoted as x_{wp} . Water conductivity exponent is represented by positive real values. This value is used to estimate the vertical flow speed of water in different soil types. According to the GTK the conductivity speed of water is determined by the formula:

$$V(x_{wp}) = 10^{-x_{wp}} \times 3600 \times 24 \frac{m}{s},$$

where m/s stands for meters per second. This formula estimates the speed of water flow through the soil. Different soil types affect the rate at which water is penetrating through the soil. In Figure 4 we have plotted the soil types based on their water conductivity exponent ranges. We can see that many of the soil types overlap each other based on their x_{wp} ranges. For example bedrock and carex peat could be considered as a cluster of their own. If we cluster the soil types based on x_{wp} ranges we could arrive into the following clustering: {sandy till, fine-grained till}, {gravel, sand}, {fine sand, coarse silt}, {silt} and {bedrock, carex peat}. We left out sphagnum peat because it has the widest range and overlaps most of the clusters.

In Table I we can see the soil types, their averaged x_{wp} values (x_{wp} has a lower and upper bound for each soil type) and relative percentages of the data. For instance, bedrock and carex peat have a water flow of approximately $0 \frac{m}{s}$, whereas for gravel the speed is approximately $0.09 \frac{m}{s}$ and for sandy till it is $27 \frac{m}{s}$.

TABLE I: List of soil types, averaged values x_{wp} and relative percentages from the target area.

Id	Type	x_{wp}	%
1	Bedrock	11	7
2	Sandy till	3,5	34
3	Fine-grained till	4,5	22
4	Gravel	6	1
5	Sand	5,75	4
6	Fine sand	7	1
7	Coarse silt	7,25	2
8	Silt	8,75	1
9	Carex peat	10,5	13
10	Sphagnum peat	6	15

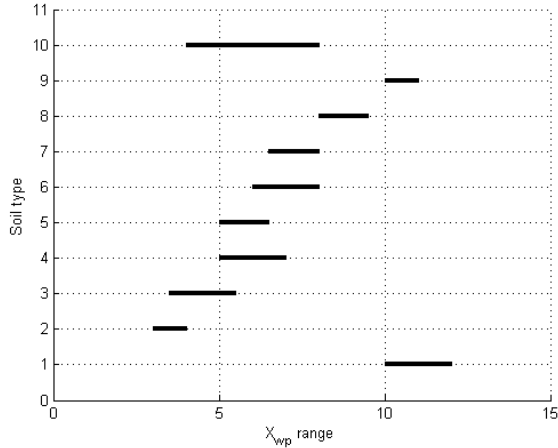


Fig. 4: The ranges for water conductivity exponents illustrated. The Y-axis describes the soil type id value, which can be used to reference Table I. X-axis describes the values for the x_{wp} exponent.

IV. ANALYSIS AND RESULTS

We now explain our analysis and the corresponding results in two parts. First we predict the soil type using multinomial logistic regression and then we predict the rate of water conductivity using regularized linear regression based on the data presented before. We also used k-nearest neighbor classifier for comparison purposes with the logistic regression. Optimal value 8 for k was selected by 10-fold cross validation approach. By predicting the rate of water conductivity we can infer the soil type by using the water conductivity exponents presented in Table I. Water conductivity exponent can also be used to predict the carrying capacity of the soil, which is an important factor e.g. when we are making strategic decisions on what routes should heavy forest machinery use. We use 10-fold cross-validation for estimating the prediction and classification accuracies of our models.

A. Soil type classification with multinomial logistic regression

Multinomial logistic regression was used in MATLAB-environment with output classes for the soil types ranging from 1 to 10 respectively. We did the same analysis with three different feature settings:

$$\begin{aligned} Z_i^{(s)} &= (f_i, \mu_{i3}, \mu_{i5}, \sigma_{i3}, \sigma_{i5}), \\ Z_i^{(t)} &= (t_{i1}, t_{i2}, t_{i3}, t_{i4}), \\ Z_i^{(st)} &= (f_i, \mu_{i3}, \mu_{i5}, \sigma_{i3}, \sigma_{i5}, t_{i1}, t_{i2}, t_{i3}, t_{i4}), \end{aligned}$$

where s stands for gamma-ray statistical features, t stands for topographical features and st is the combination of both. The symbol i refers to the features of observation i . The features are the following: f_i is the intensity of gamma-ray radiation, μ_{i3} is the 3×3 windowed mean of gamma intensity, where f_i is the center point, μ_{i5} is the same as μ_{i3} but with a 5×5 window.

Similarly σ_{i3} and σ_{i5} are the windowed standard deviations. The topographical features are t_{i1}, \dots, t_{i4} , that is the height, ground inclination, convergence index and flow accumulation.

Positive integers from 1 to 10 were used to differentiate between the soil type classes. Because the soil types weren't defined to have any ordinality, the classes were converted into 10-bit binary vectors, where only one bit had the value 1 specifying the class of the observation.

The best results were received by using both the derived gamma-ray and topographical features. Prediction performance was approximately 5-6 % lower in the case of logistic regression when either gamma-ray or topographical features were used alone. Similar results were noticed with 8-nearest neighbor where the prediction performance was more than 11 % lower by using only either gamma-ray or topographical results. The confusion matrices of the achieved results (in percentages) using both gamma-ray and topographical features for logistic regression and 8-nearest neighbor can be seen in Table II and III correspondingly.

The results indicate that in most cases especially the soil types sandy till and sphagnum peat are detected well from the data. It was noticed that statistical gamma-ray features give better results for the classification when compared with the classification using topographical results. Combining both statistical and topographical features we get the best results.

The results were compared with baseline performance, where the idea is to replace the forecasts with a constant value such that the used error measure is minimized. The baseline predictor (predicting the mode of soil type labels) achieved prediction performance of 34,8 %, almost 10 % lower than logistic regression and more than 15 % lower than 8-nearest neighbor.

The results are promising in that sense that the best predictions are received with soil types where the water conductivity is fastest. This result is useful especially for heavy machinery in industry.

Nevertheless of the good prediction accuracy of sandy till, we must also take into account the available data we used. The training data for the classifier consisted 34 % of sandy till, 22 % fine-grained till and 15 % of sphagnum peat. This inevitably affects the prediction results and the difficulty to detect the low frequency classes e.g. gravel, fine sand and silt.

B. Predicting water conductivity of the soil with regularized linear regression

As mentioned before, water conductivity exponent x_{wp} refers to the rate of speed, at which water is penetrating through the soil. We use regularized linear regression for predicting the value of x_{wp} at unknown regions. The values used for regularization parameter λ ranged from $2^{-20}, \dots, 2^{20}$. Three different error measures were used for estimating prediction performance: mean absolute deviation percentage error (MADPE), mean interval absolute deviation (MIAD) and concordance index (CI), see [4]. Explicitly, the error measures are:

TABLE II: Confusion matrix demonstrating the results (in percentages) of 10-fold cross-validation for soil type classification using both statistical gamma-ray and topographical features with multinomial logistic regression. The amount of data points in the diagonal of the matrix consist 44.5 % of the classifications. Y-axis denotes the class label prediction and X-axis denotes the real class label.

1	0.4	0.2	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0
2	5.4	26.5	12.6	0.7	2.9	0.7	0.2	0.1	3.4	3.2
3	0.5	4.1	5.8	0.0	0.3	0.5	0.8	0.4	1.6	1.1
4	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.1	0.5	0.1	0.0	0.1	0.2	1.2	0.3	0.1	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.1	1.0	0.9	0.0	0.2	0.1	0.0	0.0	2.3	2.0
10	0.2	2.5	2.6	0.0	0.5	0.0	0.0	0.0	5.0	8.2
	1	2	3	4	5	6	7	8	9	10

TABLE III: Confusion matrix of same classification task as with logistic regression, but with the classifier being 8-nearest neighbor. The amount of data points in the diagonal of the matrix consist 50.5 % of the classifications. The axes are the same as in Table II.

1	1.7	1.5	0.6	0.1	0.1	0.0	0.0	0.0	0.1	0.1
2	3.5	23.7	7.8	0.4	1.9	0.6	0.7	0.3	3.1	2.8
3	0.9	5.6	10.7	0.1	0.6	0.5	0.2	0.1	2.0	1.7
4	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
5	0.1	0.4	0.1	0.1	0.8	0.0	0.0	0.0	0.1	0.2
6	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
7	0.0	0.4	0.2	0.0	0.0	0.1	1.1	0.2	0.1	0.0
8	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0
9	0.2	1.5	1.3	0.0	0.3	0.1	0.1	0.0	4.5	2.3
10	0.2	1.6	1.3	0.0	0.4	0.1	0.0	0.0	2.5	7.6
	1	2	3	4	5	6	7	8	9	10

$$MADPE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i|}, \quad (3)$$

$$CI = \frac{1}{N} \sum_{y_i < y_j} h(\hat{y}_i - \hat{y}_j), \quad (4)$$

$$MIAD = \frac{1}{n} \sum_{i=1}^n e(\hat{y}_i), \quad \text{where} \quad (5)$$

$$e(\hat{y}_i) = \begin{cases} |y_i^u - \hat{y}_i|, & \text{if } \hat{y}_i > y_i^u \\ |y_i^l - \hat{y}_i|, & \text{if } \hat{y}_i < y_i^l \\ 0, & \text{otherwise.} \end{cases}$$

In the above equations we denote y_i as the i th response value and \hat{y}_i as the corresponding forecast value. In equation (4) we

denote $N = |\{(i, j) \mid y_i > y_j\}|$ 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 y_i^u and y_i^l in equation (5) represent the upper and lower bounds of the water conductivity exponent of i th observation. 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. Similarly as in the case of logistic regression, we used 10-fold cross-validation for approximating the prediction accuracy of the regularized linear regression. The optimal value for regularization parameter was found to be $\lambda = 2^{15}$. As a baseline prediction for comparing the different error measures of the linear model we used

median of the training data labels (i.e. median of x_{wp} values of the training set). For C-index we used the value 0.5 as baseline, because it is invariant for the distribution of the data labels. In regression we used both statistical and topographical features, that is the i th observation corresponded to a 1×10 row vector $\mathbf{x}_i = (1, f_i, \mu_{i3}, \mu_{i5}, \sigma_{i3}, \sigma_{i5}, t_{i1}, t_{i2}, t_{i3}, t_{i4})$, where the 1 is the constant for bias term. The corresponding output value is the real value $y_i \in \mathbb{R}^+$. The results for regularized linear regression can be seen in Table IV.

If we look at the concordance index (CI) we notice that regression was able to detect signal from the data, considering the amount of data (342479 data points). MIAD and MADPE also show lower error than the corresponding baselines.

TABLE IV: Table demonstrating the results for different error measures of the regularized linear regression using 10-fold cross-validation. The second column represents the value of the corresponding error measure and the third column is the baseline performance used for comparison.

Error measure	value	baseline
MADPE	0.33	0.35
MIAD	1.4	2.2
CI	0.63	0.5

V. CONCLUSIONS AND FUTURE WORK

The results indicate that gamma-ray and topographical data can be used to detect soil types up to approximately 44.5 % accuracy using multinomial logistic regression. An increase in the accuracy was achieved using 8-nearest neighbor classifier, which achieves approximately 50.5 % accuracy. We note that especially sandy till and sphagnum peat have good prediction accuracies when compared with other soil types. When considering low water conductivity soil types, the prediction accuracy was higher.

We also noted that regularized linear regression was able to detect signal from the data by having lower error rates in all the error measures when comparing to baseline error.

Given the hydraulic conduction prediction, additional expert rules or machine learning methods can be used to select the soil type from several possible indicated by the soil water conductivity. Such expert rules exist but are not yet implemented.

There is also a possibility to use aerial Light Detection and Ranging (LiDAR) data instead of the topographical height data. This would enable computation of several surface texture features. Additional features could give better capability to predict the actual soil types directly. This remains subject of further study.

ACKNOWLEDGEMENTS

This work is done as a part of ULJATH project, which is funded by the *Finnish Funding Agency for Technology and Innovation* (TEKES).

REFERENCES

- [1] A. Azzalini, P.J. Diggle. *Prediction of soil respiration rates from temperature, moisture and soil type*. Journal of the Royal Statistical Society - Series C: Applied Statistics, 43:505–526, 1994.
- [2] R.A. Chapuis. *Predicting the saturated hydraulic conductivity of soils: a review*. Bulletin of Engineering Geology and the Environment, 71: 401–434, 2012.
- [3] H. Gao, Q. Tang, X. Shi, C. Zhu, T.J. Bohn, F. Su, J. Sheffield, M. Pan, D.P. Lettenmaier, E.F. Wood. *Water budget record from variable infiltration capacity (vic) model. Algorithm Theoretical Basis Document for Terrestrial Water Cycle Data Records*, 2010.
- [4] M. Gönen, G. Heller. *Concordance probability and discriminatory power in proportional*. Biometrika, 92:965–970, 2005.
- [5] E. Hyvönen, M. Päänttjä, M-L. Sutinen, R. Sutinen. *Assessing site suitability for Scots pine using airborne and terrestrial gamma-ray measurements in Finnish Lapland*. Canadian Journal of Forest Research. 33(5), 796–806, 2003.
- [6] R. Kiss. *Determination of drainage network in digital elevation models, utilities and limitations*. Journal of Hungarian Geomathematics, 2:16–29, 2004.
- [7] H.S. Mahmood, W.B. Hoogmoed, E.J. van Henten. *Proximal Gamma-Ray Spectroscopy to Predict Soil Properties Using Windows and Full-Spectrum Analysis Methods*. Sensors, 13:16263–16280, 2013.
- [8] N.J. McKenzie, P.J. Ryan *Spatial prediction of soil properties using environmental correlation*. Geoderma, 89:67–94, 1999.
- [9] I.D. Moore, P.E. Gessler, G.A. Nielsen, G.A. Peterson *Soil Attribute Prediction Using Terrain Analysis*, Soil Science Society of America Journal, 57:443–452, 1993.
- [10] P. Nevalainen, J. Pohjankukka, T. Pahikkala, R. Sutinen, J. Varjo, J. Heikkonen *Open Natural Resource Data in Forecasting the Harvester Mobility*. Proceedings of The Federated Computer Science Event 2014 (YTP 2014) to appear.
- [11] A.T. Ramli, A.T. Rahman, M.H. Lee. *Statistical prediction of terrestrial gamma radiation dose rate based on geological features and soil types in Kota Tinggi district, Malaysia*. Applied Radiation and Isotopes, 59:393–405, 2003.
- [12] P. Sculla, J. Franklin, O.A. Chadwick. *The application of classification tree analysis to soil type prediction in a desert landscape*. Ecological Modelling, 181:1–15, 2005.
- [13] R.P.O. Schulte, J. Diamond, K. Finkle, N.M. Holden, A.J. Brereton. *Predicting the soil moisture conditions of Irish grasslands*. Irish Journal of Agricultural and Food Research, 44: 95–110, 2005.
- [14] J.V.A. Zachary. *Using topographic and soils data to understand and predict field scale moisture patterns*. Graduate thesis, Iowa State University, 2012.

Publication II

Predicting Water Permeability of the Soil Based on Open Data

Jonne Pohjankukka, Paavo Nevalainen, Tapio Pahikkala, Eija Hyvönen, Pekka Hänninen, Raimo Sutinen, Jari Ala-Ilomäki and Jukka Heikkonen. In Lazaros Iliadis, Ilias Maglogiannis, and Harris Papadopoulos, editors, Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014), volume 436 of IFIP Advances in Information and Communication Technology, pages 436–446. Springer, 2014.

Copyright © 2014 Springer Nature. Reprinted with permissions from respective publisher and authors.

Predicting Water Permeability of the Soil Based on Open Data

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

University of Turku, Computer Science Dept.

{Jonne.Pohjankukka,ptneva,Tapio.Pahikkala}@utu.fi,
{Eija.Hyvonen,Pekka.Hanninen,Raimo.Sutinen}@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. The results were measured using leave-one-out cross-validation with a dead zone between the training and testing data sets.

Keywords: 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.

Our input data set consists of 44 features which are publicly available. The data is in 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. R.P.O. Schulte et al.[3] focuses on soil moisture deficit, which is a related concept but not of concern in sub-polar areas, H. Gao et al. [4] and R.A. Chapuis [5] focus on water budget modeling, which was not yet attempted in our study. H.S. Mahmood et al. [6] uses on-site gamma-ray measurements for analysis of the farming soil. N.J. McKenzie [7] combines gamma-ray and digital elevation model to predict the chemical composition of the farming land. Closest to our paper is [8], where several data sources (topographic and remote sensing) are combined with 85 soil samples to assess the usability of the soil within and outside the sampled area. One can try to by-pass the water permeability estimation by directly learning the dynamic coupling between the precipitation and remotely observed soil moisture. This approach must include the digital terrain model (DTM) to estimate the water catchment. An example of this approach is [9].

The main novelty of this paper related to the previous studies is that the prediction is based on wide-area public data on a subpolar region. 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, \dots, 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:

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

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, \dots, 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 So-dankylä, 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 (VMI¹) 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 sub-arctic areas (e.g. Russia, Canada), several studies are underway to predict the main characteristics of the forest by remote measurement methods [10]. These methods include LiDAR and various satellite measurements.

The data providers are:

Table 1. Data providers, data and the grid size

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

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 electro-chemical properties of the soil were measured, see [11]. 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.

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

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 spot-wise verification and calibration methods are applied to it before publishing the data [12]. 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.

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.

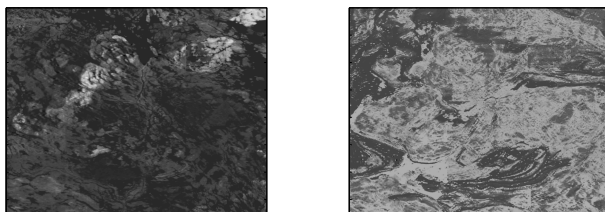


Fig. 1. Aerial data: gamma-ray (left) and electromagnetic data (right). Air-borne electromagnetic data is sensitive to geological properties to depth of hundreds of meters, but it also indicates some features of the top soil.

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 [13]. 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 [14]. 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 [15]
- Local Binary Pattern (LBP) with pixel radii $r \in \{1, 2\}$, see [16]

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 [17]).

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:

$$x_{wp} = -\log_{10} v, [v] = \frac{m}{sec}, \quad (2)$$

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 omitted 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. 443. 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.

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 optimistic 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 data set 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 in Sec. 3.7 before, the prediction subject is the water permeability exponent x_{wp} defined by Eq. 2. 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

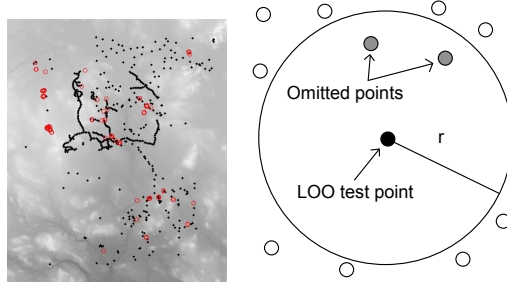


Fig. 2. *Left:* 1788 sample points. Set A (1187 points, marked with red circles, distance to the nearest neighbor $d_{NN} \leq 86m$) is tightly packed and set B is very sparse (601 points, marked with black dots, aver. $d_{NN} \approx 1.1 km$). *Right:* the dead zone (with radius r) around the leave-one-out 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.

error measures were used for estimating prediction performance: mean absolute error (MAE) and concordance index (CI) [18]. Explicitly, the error measures are:

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

MAE prediction baseline \tilde{y} is the best possible prediction under the assumption that the prediction will be constant, thus constraining all values $\hat{y}_j = \tilde{y}, j = 1..n$ in the error minimization process. MAE baseline becomes thus:

$$MAE_b = \arg \min_{\tilde{y}} \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \tilde{y}}{y_i} \right|$$

The prediction performance should be better than this to be useful. The corresponding percentage values (MAPE and MAPE_b) have been used in the rest of the text.

Concordance index counts the occurrences when the prediction fails to be monotonical. In equation (3) we denote $N = | \{(i, j) \mid y_i > y_j \} |$ as the normalization constant which equals to the number of data pairs with different label values. $h(\cdot)$ is Heaviside step function.

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 case where prediction is monotonically correct.

4.2 Results

The results for regression analysis can be seen in Fig. 3 on p. 443.

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.

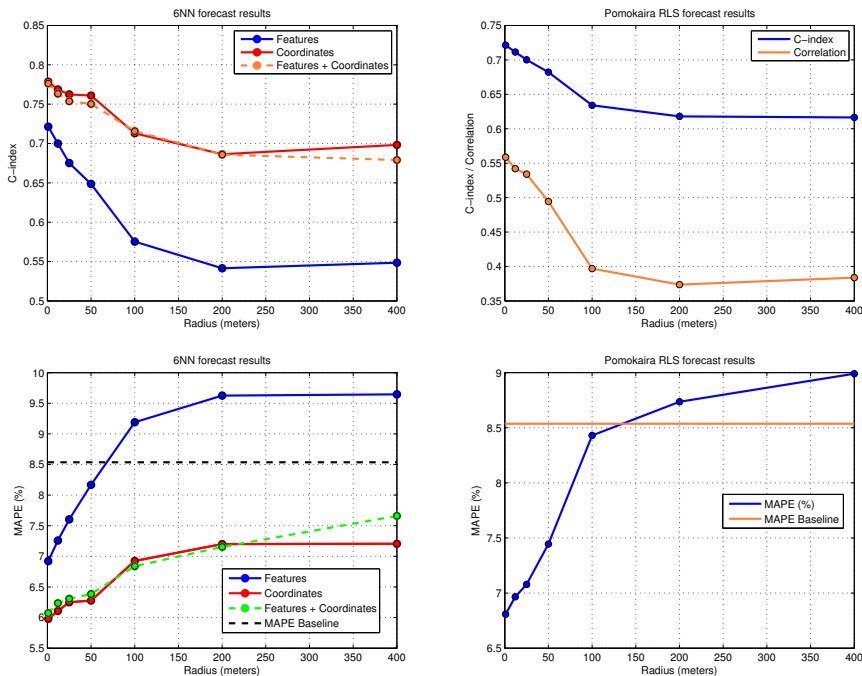


Fig. 3. *Left:* k-NN results with $k = 6$ and 3 different feature sets. *Right:* RLS results on features-only case. C-index and Pearson correlation at top and MAPE below. The prediction performance is adequate below 120 meters.

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

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 in the sense of interpolation and extrapolation performance, and general prediction is usually analyzed in the terms of Machine Learning performance.

The feature selection was not attempted. There were two reasons for this:

- the number of features (96) remained modest.
- the forest inventory features are unique to Finland. They can be largely substituted by various remote measurements [12], which would extend the application scope of the method to whole sub-polar area.

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.

There are several possible improvements. Since the mapping from water permeability to soil types is not unique, see [19], 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.

The Aerial Light Detection and Ranging (LiDAR) data can substitute most of the topographical and forest inventory data features. This would extend the scope of the prediction to any location at the arctic zone, where only aerial and satellite measurements are economical. LiDAR has also potential for derived features like geological morphology [20] and soil water budget modeling [10].

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 and online learning based on measurements from the harvester fleet. The potential applications aim to wide-area routing and location planning. In this regard, even a modest prediction power of features-only prediction could yield a cumulative effect on route decisions.

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*.

The authors would like to thank the anonymous reviewers for valuable comments and suggestions.

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, 505–526 (1994)
2. Sculla, P., Franklin, J., Chadwick, O.: The application of classification tree analysis to soil type prediction in a desert landscape. *Ecological Modelling* 181, 1–15 (2005)
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, 95–110 (2005)
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, 401–434 (2012)
6. Mahmood, H., Hoogmoed, W., van Henten, E.J.: Proximal gamma-ray spectroscopy to predict soil properties. *Sensors* 13, 16263–16280 (2013)
7. McKenzie, N., Ryan, P.: Spatial prediction of soil properties using environmental correlation. *Geoderma* 89, 67–94 (1999)
8. Emadi, M., Baghernejad, M., Pakparvar, M., Kowsar, S.: An approach for land suitability evaluation using geostatistics, remote sensing, and geographic information system in arid and semiarid ecosystems. *Environ Monit Assess* 164, 501–511 (2010)
9. J.P.W.P.R.H.: Soil moisture estimation using remote sensing. In: *Proceedings of the 27th Hydrology and Water Resources Symposium*, The Institute of Engineers Australia, Melbourne (2002)
10. 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)
11. Hyvönen, E., Päänttjä, 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(11), 796–806 (2003)
12. Tomppo, E., Katila, M., Mäkisara, K., Peräsaari, J.: *Multi-source national forest inventory – methods and applications*. *Managing Forest Ecosystems*, vol. 18. Springer (2008)
13. Hautaniemi, H., Kurimo, M., Multala, J., Leväniemi, H., Vironmäki, J.: The “three in one” aerogeophysical concept of gtk in 2004. *Geological Survey of Finland, Special Paper* 39, 21–74 (2005)
14. Schwanghart, W., Kuhn, N.: Topotoolbox: a set of matlab functions for topographic analysis. *Environmental Modelling & Software* 25, 770–781 (2010)
15. Weldon, T.P., Higgins, W.E., Dunn, D.F.: Efficient gabor filter design for texture segmentation. *Pattern Recognition*, 2005–2015 (1996)
16. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 971–987 (2002)
17. Kiss, R.: Determination of drainage network in digital elevation models, utilities and limitations. *Journal of Hungarian Geomathematics* 2, 16–29 (2004)

18. Gönen, M., Heller, G.: Concordance probability and discriminatory power in proportional. *Biometrika* 92, 965–970 (2005)
19. 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. In: *Proceedings of the 22nd International Conference on, ICPR 2014* (to appear, 2014)
20. Sutinen, R., Hyvönen, E., Middleton, M., Ruskeenieni, T.: Airborne lidar detection of postglacial faults and pulju moraine. *Global and Planetary Change*, 24–32 (2014)

Publication III

Predictability of boreal forest soil bearing capacity by machine learning

Jonne Pohjankukka, Henri Riihimäki, Paavo Nevalainen, Tapio Pahikkala, Jari Ala-Ilomäki, Eija Hyvönen, Jari Varjo and Jukka Heikkonen. *Journal of Terramechanics*, 68:1–8. Elsevier, 2016.

Copyright © 2016 Elsevier. Reprinted with permissions from respective publisher and authors.



Predictability of boreal forest soil bearing capacity by machine learning

J. Pohjankukka^{a,*}, H. Riihimäki^b, P. Nevalainen^a, T. Pahikkala^a, J. Ala-Ilomäki^b,
E. Hyvönen^c, J. Varjo^b, J. Heikkonen^a

^a Department of Information Technology, University of Turku, FI-20014 Turku, Finland

^b Natural Resources Institute Finland, FI-00790 Helsinki, Finland

^c Geological Survey of Finland, FI-02151 Espoo, Finland

Received 28 May 2015; received in revised form 14 September 2016; accepted 15 September 2016

Available online 30 September 2016

Abstract

In forest harvesting, terrain trafficability is the key parameter needed for route planning. Advance knowledge of the soil bearing capacity is crucial for heavy machinery operations. Especially peatland areas can cause severe problems for harvesting operations and can result in increased costs. In addition to avoiding potential damage to the soil, route planning must also take into consideration the root damage to the remaining trees. In this paper we study the predictability of boreal soil load bearing capacity by using remote sensing data and field measurement data. We conduct our research by using both linear and nonlinear methods of machine learning. With the best prediction method, ridge regression, the results are promising with a C-index value higher than 0.68 up to 200 m prediction range from the closest point with known bearing capacity, the baseline value being 0.5. The load bearing classification of the soil resulted in 76% accuracy up to 60 m by using a multilayer perceptron method. The results indicate that there is a potential for production applications and that there is a great need for automatic real-time sensing in order to produce applicable predictions.

© 2016 ISTVS. Published by Elsevier Ltd. All rights reserved.

Keywords: Terrain trafficability; Soil bearing capacity prediction; Forest harvesting; Machine learning; Open data

1. Introduction

Terrain trafficability in forests is currently one of the most important issues in boreal timber harvesting. Conducting harvesting operations during good soil bearing conditions is crucial since improperly timed operations can cause serious economical and ecological damage. Vehicular loading exceeding soil strength causes not only soil damage, but also damage to trees, mostly to the tree roots, but sometimes to tree stem as well due to increasing uncontrolled motion of the forwarder.

Damage to roots and stems can lead to fungal infection which eventually causes wood discoloration and in the

worst case decay. In addition, the water and nutrition conditions of the forest soil can change as a result of soil settling (Ring et al., 2006). The operation of forest machines is therefore avoided during the period of high soil failure risk and the harvesting is postponed to the winter when soil is normally frozen. It is estimated that the seasonal variation in timber procurement causes approximately 100 M € costs in Finland alone (Pennanen and Mäkelä, 2003). In addition, operations in poorly bearing conditions increase time and fuel consumption and decrease the efficiency of harvesting operations (Sirén et al., 2013).

Furthermore, deep ruts caused by forwarding affect the general acceptability of the forest operations. The costs caused by challenging trafficability conditions could be decreased by additional information on soil conditions, especially soil bearing capacity. The load bearing capacity

* Corresponding author.

E-mail address: jonne.pohjankukka@utu.fi (J. Pohjankukka).

of soil is often described by its penetration resistance. Accordingly, forest operations could be planned to be performed during adequate bearing capacity or routed to avoid sections of poor bearing capacity, thus minimizing the damage and maximizing the efficiency of harvesting.

In this study we conduct a research on the prediction of soil bearing capacity by using remote sensing and field measurement data. We have analyzed two cases, firstly visual soil damage classification and secondly soil penetration resistance prediction. The data sets are provided by Natural Resources Institute Finland (LUKE), Metsäteho Ltd., the Geological Survey of Finland (GTK), National Land Survey of Finland (NLS) and Finnish Meteorological Institute (FMI). Similar studies have been conducted in Pohjankukka et al. (2014a,b) where soil properties such as type and water permeability was estimated in order to have predictions on the soil bearing capacity using public data. Related studies have been conducted in Azzalini and Diggle (1994) where soil respiration rates are predicted from temperature, moisture content and soil type and (Schulte et al., 2005), where the soil type in desert landscapes was predicted using classification tree analysis.

2. Background

Timber harvesting systems vary across the world. In Finland, the mechanized cut-to-length harvesting system is utilized almost exclusively (Uusitalo, 2010). Harvesting operations in Finland are typically commercial thinnings or clear cuttings. In a traditional thinning operation only a part of the trees, on average 30%, are cut, leaving most of the trees standing (Äijälä et al., 2014). Depending on stand properties thinnings are typically done one to three times during the rotation of a stand (Äijälä et al., 2014).

The rotation period of a stand usually ends to a final felling, where all trees of commercial value are cut. Some individual tree clusters are left standing for example to retain biodiversity (Gustafsson and Perhans, 2010). The structure of private forest ownership in Finland has changed, which is causing pressure to change the forestry practices, as many forest owners are no more dependent of forest income and emphasize multiple values in management decisions. The commercial aspect of harvesting has become less pronounced, while environmental standpoint has gained more attention. More than a half of the forest owners are satisfied with the current forest management practices, where every sixth forest owner feels unsatisfied especially with clear cuttings, lack of management alternatives, soil preparation and damage caused by heavy machinery (Hänninen and Karppinen, 2010). So far the use of alternative forest management methods including selection cuttings has been marginal concentrating on urban forests, landscape protection areas, valuable habitats, riparian and other buffer zones. If uneven-aged forest management becomes more popular in future, it increases the amount of thinnings. Uneven-aged thinnings place even

more challenges to harvesting machinery in respect to avoiding damages and risk of root rot.

3. Research area and data sets

3.1. Research area

The data sets were collected from various locations around the area of Pieksämäki, a municipality located in the province of Eastern Finland 62°18'N 27°08'E. The research areas were divided into two cases based on the response variable. The predictor data sets varied between the two cases as illustrated in Tables 1 and 2.

3.2. Multi-source national forest inventory data

The Multi-Source National Forest Inventory (MS-NFI) holds the state of Finnish forests in high spatial resolution (20 m). The data is updated every second year. The parameters are derived by generalizing the field measured sample plot data applying mainly Landsat imagery and KNN method as well as digital map information. 43 numerical features include information regarding, for example, biomass and volume of growing stock and site type. These multi-source features exhibit built-in dependencies, thus the final number of useful features is lower. An excellent, detailed description regarding the MS-NFI is given by Tomppo et al. (2008).

3.3. Digital elevation model data

We downloaded digital elevation model (DEM) data from the file service for open data by the National Land Survey of Finland. The DEM was made from airborne laser scanning data with the resolution of at least 0.5 samples/m², which is equivalent to approximately 1.4 m distance between samples. The grid size of the DEM data set was 2 m. Several geomorphometric variables were derived from the NLS DEM in SAGA GIS environment. In our analysis we used the geomorphometric features: plan curvature, profile curvature, slope, topographic wetness index, flow area, aspect, diffuse insolation and direct insolation (Zevenbergen and Thorne, 1987; Wood, 2009, 1996; Beven and Kirkby, 1979; Seibert and McGlynn, 2007). These derived features are more efficient for prediction than raw height data alone.

Table 1
Predictor data sets used in prediction of soil damage response variable. RS stands for remote sensing data and FM stands for field measurement data.

Data set	Type	Grid size
Digital Elevation Model data	RS	2 m
Multi-source National Forest Inventory data	RS	20 m
Soil type data	RS	20 m
Peatland data	RS	20 m
Gamma-ray spectroscopy data	RS	50 m
Weather data	RS	10 km

Table 2
Predictor data sets used in prediction of penetration resistance response variable.

Data set	Type	Grid size
Stoniness data	FM	2 m
Peatland data	FM	2 m
Soil moisture data	FM	2 m
Digital Elevation Model data	RS	2 m
Multi-source National Forest Inventory data	RS	20 m
Weather data	RS	10 km

3.4. Weather data

Weather data consisting of temperature (°C) and rainfall (mm) for years 2011–2013 was provided by the Finnish Meteorological Institute. The grid size of the data set was 10 km. In our analyses we used the mean temperature and rainfall of the last 30 days as predictor features for each observation of the response value. For example if an observation of the response value was measured June 15, 2013 the mean temperature and rainfall predictor features for the response value observation were calculated from the time interval May 16 - June 14, 2013.

3.5. Aerial gamma-ray spectroscopy data

The aerial gamma-ray data with grid size of 50 m 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 (Hyvönen et al., 2003). Also the soil type, especially density, porosity, grain size and humidity of the soil have an effect on gamma-ray radiation. Areas with high gamma-radiation tend to have lower soil moisture and vice versa. We derived several statistical and textural features from Gamma-ray data such as: 3 × 3 windowed mean, 3 × 3 windowed standard deviation, Gabor filter features (see e.g. Feichtinger and Strohmer, 1997) and Local Binary Pattern features (Pietikäinen et al., 2011).

3.6. Peatland data

The peatland data was compiled by LUKE using the open geographic information data derived from NLS Topographic database (NLS, 2014) depicting the terrain and covering the whole of Finland. The positional accuracy of the NLS Topographic database corresponds to that of scales 1:5000–1:10,000 (NLS, 2014). The peatland mask consist of four different NLS Topographic database elements depicting different type of peatlands. These elements were first combined and then rasterized to 20 m grid using ArcMap software (ArcMap, 2014). The definitions for peatlands in the NLS Topographic database are: (1) area is mostly covered by peatland vegetation and (2) a mini-

imum of 0.3 m peat thickness (NLS, 2014). A minimum criteria for area is 1000 m². Area with peat thickness less than 0.3 m can also be classified as peatland if it is covered by peatland vegetation.

3.7. Subsoil and topsoil data

GTK provided the analysis of subsoil classification data and topsoil classification data from Piekäsämäki target area. The soil type data is represented by positive integer values, which indicate the pre-classified soil types. Both of the soil type data sets consisted of twelve distinct soil types e.g. bedrock, *Sphagnum* peat, *Carex* peat and sandy till. The grid size of these data sets were also 20 m.

3.8. Soil moisture data

Gravimetric soil water content was measured from the samples by drying the soil samples and calculating the weight difference of dry and wet soil sample (ASTM D2216-10, 2010).

3.9. Soil damage data

Approximately 36 km of strip roads were walked through and visually assessed into damage classes by a forest operations expert. The data was kindly provided to us by Metsäteho Ltd. in 2013. The soil damage data was classified into three main ordinal classes based on the rut depth caused by forest harvesting machinery. The three classes were: (1) No damage; (2) Slight damage; and (3) Damage.¹

The original dataset required preprocessing since the field recorded GPS-tracks included locational errors (zig-zag -motion). After the data was preprocessed to produce a smooth line form, we converted strip road lines into points and extracted values from selected features, e.g. MS-NFI and topographic variables.

3.10. Soil penetration resistance, stoniness and shear modulus

The total of 50 penetration resistance measurements were taken on two different locations in Kumpunen, Piekäsämäki, Finland (N 6921354, E 501297 in ETRS-TM35FIN coordinates). The study was conducted during a commercial thinning operation. The plot locations were selected based on expert judgment to cover the gradient between dry mineral soil with high bearing capacity and wet organic soil with low bearing capacity. 15 plots were measured from dry site and 35 plots were measured from wetter, partly paludificated site. Sites were located roughly 490 m apart from each other. Depth of organic soil varied from 0 to almost 90 cm. We measured the soil penetration

¹ Includes strip road sections covered by brush mat, originally classified as “potential damage”, since without brush mat they likely would have been damaged.

resistance using a penetrometer (Muro, 2004) at five different locations around and between the wheel tracks to avoid the random effect caused by e.g. hitting a tree root in a single measurement. This method is illustrated in Fig. 1. Shear modulus was measured at the same locations with a spiked shear vane (Ala-Ilomäki, 2013). The accumulation of logging residue significantly hindered measuring, and it was not always possible to place measurements systematically.

3.11. Rut depth measurement

Depth of both wheel ruts was measured using an inversed U-shaped frame with its feet resting on the undeformed soil surface outside wheel rut, which formed the reference level. Individual observations were averaged to plot level. First measurements were taken after the harvester and the rut formation was measured again after each pass of the forwarder collecting the timber from the cutting area. The extraction road was cleared of logging residue after the harvester pass in order to observe the effect of soil properties on forwarder rut formation without the reinforcing effect of brash (Sirén et al., 2013). The accumulated mass traversed over each measuring location was defined as the sum of net vehicle mass plus the mass of load for all the passes (Sirén et al., 2013).

4. Methods

The possibilities to predict the response variables were estimated using both linear and nonlinear methods. Next we will describe the used prediction methods including leave-one-out cross-validation with a dead zone approach for the model performance estimation given by the concordance index (C-index).

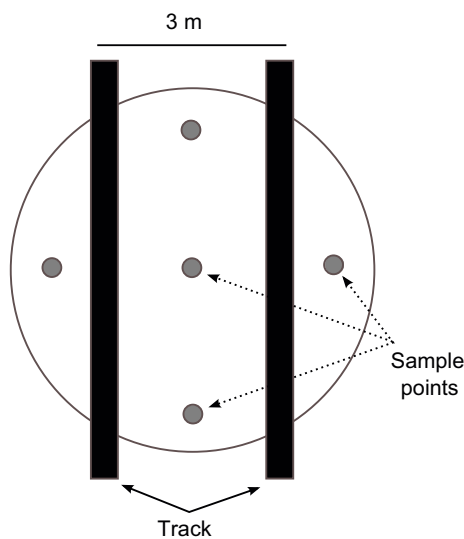


Fig. 1. Illustration of how the field measurements were made. Black rectangles represent the wheel tracks and gray points represent the measurement points. The width of the track was approximately 0.4 m and distance from the center of the left track to the center of the right track was 2.8 m.

4.1. Leave-one-out cross-validation with a dead zone

In geographical applications there is bound to be some sort of spatial autocorrelation between the data points. Data points very close to each other geographically have intuitively larger spatial autocorrelation than data points far apart. Accordingly, using the traditional cross-validation approach (see e.g. Abu-Mostafa et al., 2012) that assumes the mutual independence of the data points, is not suitable here, as it only estimates the prediction capability of individual test data points, regardless of their distance from the training data. We need a way of simulating the predictions in a practical situation which is why we use the so-called *leave-one-out cross-validation with a dead zone* (LOOCVDZ), (Pohjankukka et al., 2014b).

The idea of the LOOCVDZ method is to simulate the prediction capability of the model in such a situation, where the point for which the prediction is to be made is at least n meters away from the closest training point. For each data point at a time, we create a perimeter (dead zone) of radius δ around the point and remove from the training data all the points falling inside the perimeter including the test point itself. A model is trained with the reduced training data set and a prediction is performed for the test point with the learned model. This process is repeated over the whole data set, just like an ordinary leave-one-out cross-validation. The LOOCVDZ method gives us a way of simulating a harvester or a forwarder predicting soil bearing capacity, when the closest known measurements are at least n meters away. In Fig. 2 we have illustrated the LOOCVDZ method.

4.2. Concordance index

Concordance index (C-index) was the main performance measure used in the analyses (Gönen and Heller, 2005).

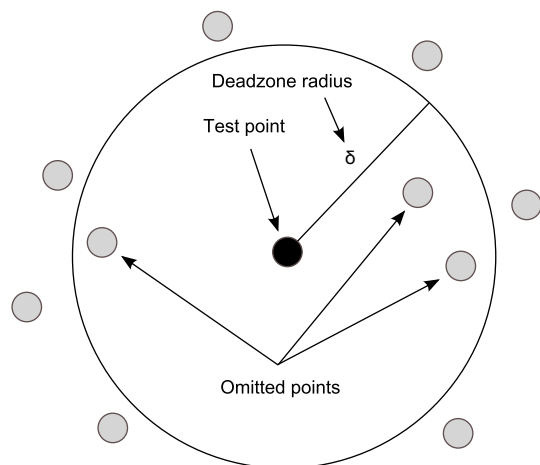


Fig. 2. Illustration of the dead zone with perimeter determined by δ . The black point is the one whose label we aim to predict. The data inside the dead zone will be omitted from training the model used to predict the label of the test point.

Concordance index measures the relative ranking of paired data points in the sets $V = \{y_1, \dots, y_n\}$ and $P = \{\hat{y}_1, \dots, \hat{y}_n\}$, where V is the set of observed labels and P is the corresponding set of predictions. The C-index measures how well the prediction model was able to rank the predictions into correct order. It is a particularly useful measure in situations where we are not especially interested in the absolute accuracy of the prediction value, but rather where we need to make a choice between a set of alternatives. In our application we are interested in selecting the most supporting area or route from a set of alternatives for the forest machine. Explicitly concordance index is defined as:

$$\text{C-index} = \frac{1}{N} \sum_{y_i < y_j} h(\hat{y}_i - \hat{y}_j), \quad (1)$$

where $N = |\{(i, j) | y_i < y_j\}|$ is 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 further apart from 0.5 C-index is, the better the model was able to capture the pattern in the data.

4.3. Ridge regression

Ridge regression, also known as Tikhonov regularization (Vapnik, 1998) is the regularized version of the standard linear regression. Let $\mathbf{x}_i \in \mathbb{R}^p$ be the feature vector of the i th sample point, $\mathbf{w} \in \mathbb{R}^p$ is a vector of weights and $y_i \in \mathbb{R}$ is the response value of i th sample. In ridge regression our task is to find the set of weights \mathbf{w} , such that the objective function:

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

is minimized. In (2), $n \in \mathbb{N}$ is the number of data points and $\lambda > 0$ is the regularization parameter.

4.4. Multilayer perceptron

Multilayer perceptron (MLP) is a feedforward neural network (Bishop, 1996; Nabney, 2004), where we try to minimize the objective function:

$$E(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (y_i - a_i(\mathbf{x}_i, \mathbf{w}))^2, \quad (3)$$

where \mathbf{w} is the set of weights of the network, a_i is the i th activation of the output node given input $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ is the corresponding response value. The set of weights \mathbf{w} is defined as:

$$\mathbf{w} := \left\{ w_{ij}^{(l)} \mid 1 \leq l \leq L, 0 \leq i \leq d^{(l-1)}, 1 \leq j \leq d^{(l)} \right\},$$

where L is the number of hidden layers and $d^{(l)}$ is the number of hidden nodes on layer l . The activation functions in the hidden nodes are $\tanh(\mathbf{x})$ functions and the output activation function was selected to be a linear function.

A popular regularization approach for MLPs is to construct a committee of MLP networks trained with early stop training (Bishop, 1996) in which training data are divided into two parts. The first part is used to train the MLP and the other part is used to monitor the validation error. Training is stopped when the validation error begins to increase. This random splitting scenario is repeated for all committee members and the final output of the MLP committee is obtained by counting the average output of the committee members. Early stop is an *ad hoc* method for regularization, but it is simple, fast and in many cases gives good results. We used a MLP early stop committee (MLP-ESC) of 10 networks with 10 hidden units.

4.5. k -nearest neighbor

k -nearest neighbor (Cover and Hart, 1967) is the simplest of the used methods, but still a powerful nonlinear method for many applications. In k -nearest neighbor we predict the label \hat{y}_i of a test point $\mathbf{x}_i \in \mathbb{R}^p$ by taking the average value of the labels of its k nearest neighbors, i.e. we use the formula:

$$\hat{y}_i = \frac{1}{k} \sum_{j=1}^k y_j,$$

where the values $y_j \in \mathbb{R}$, $i = 1, \dots, k$ are the labels of the points \mathbf{x}_j that are closest to the test point \mathbf{x}_i . Euclidian distance is the standard metric used in this method for finding the closest neighbors.

5. Analysis and results

We have separated our analysis into two cases based on the response variables:

- Case 1: Soil damage prediction.
- Case 2: Soil penetration resistance prediction.

Both of these variables can be used as indicators for soil load bearing capacity.

5.1. Case 1: Soil damage prediction

In soil damage prediction our data set consisted of 11,795 points from harvesting operations including both thinning and clearcutting. The predictor variables consisted from various remote sensing data sets and their derived features, totaling to 83 variables used in the analysis. The target variable for prediction was soil damage class consisting of three ordinal damage classes (no damage, slight damage, damage). The used data sets in this case are listed in Table 1.

We tried two different approaches in predicting soil damage class, firstly predicting the soil damage variable without any modifications to the label values and, secondly, combining the damage classes ‘slight damage’ and

‘damage’ into one class so that we could get a binary prediction problem (no damage - damage). The purpose of this second approach was to detect whether the prediction model is able to distinguish between non-damaged and damaged soil points. The multilayer perceptron model was trained with 11,295 data points and tested with a sample of 500 data points, because the overall calculations using LOOCVDZ would have taken far too much time using the entire data set. We have illustrated the prediction results using LOOCVDZ for these two approaches in Fig. 3.

In both cases the results indicate that a moderate prediction performance to a 20–30 m range is reached especially with k -nearest neighbor and ridge regression. Ridge regression stays above baseline up to 200 m but has nonetheless poor performance. Low prediction performance in case 1 was expected due to low quality of the provided response

variable. Poor results of the analysis based on visually classified data implicated the need for physical measurements. It was concluded that more accurate measurements were needed in order to improve the performance of the models. This insight motivated the collection of new data, i.e. penetration resistance as response variable.

5.2. Case 2: Soil penetration resistance prediction

Due to high noisiness of the response variable, the soil damage prediction resulted in maximum of 20–30 m moderate prediction performance. This problem was tackled in the case of soil penetration resistance prediction, where an accurate data set, measured with an electrically driven and recording penetrometer, was provided by LUKE. The penetrometer proved superior over the spiked shear vane in varying mineral soil peatland conditions. The results of this

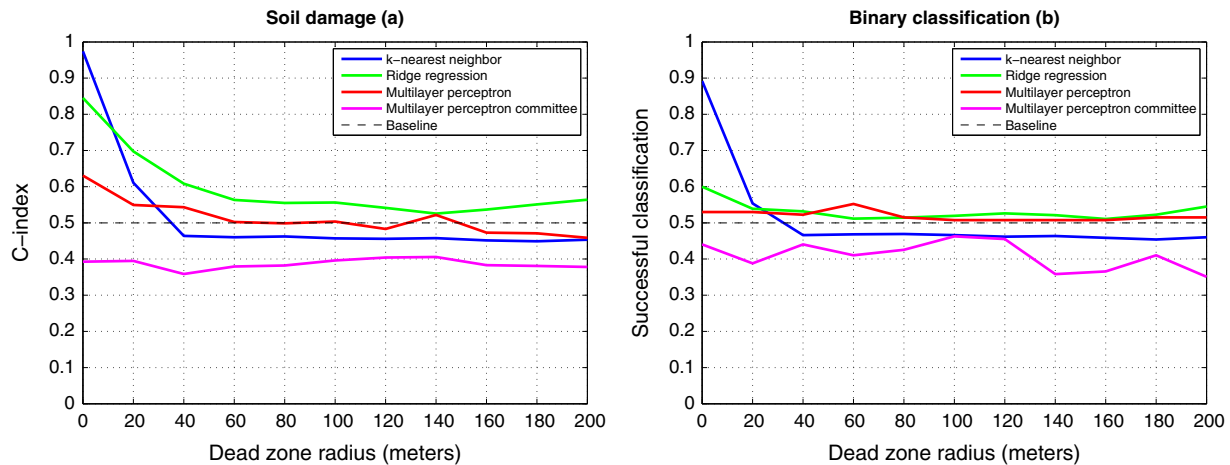


Fig. 3. Prediction results for case 1: soil damage class (a) and binary classification (b). The results show that there is a moderate prediction accuracy up to 20 m in both regression and classification. k -nearest neighbor achieves the highest performance to 20 m range. Ridge regression gives highest results after 20 m.

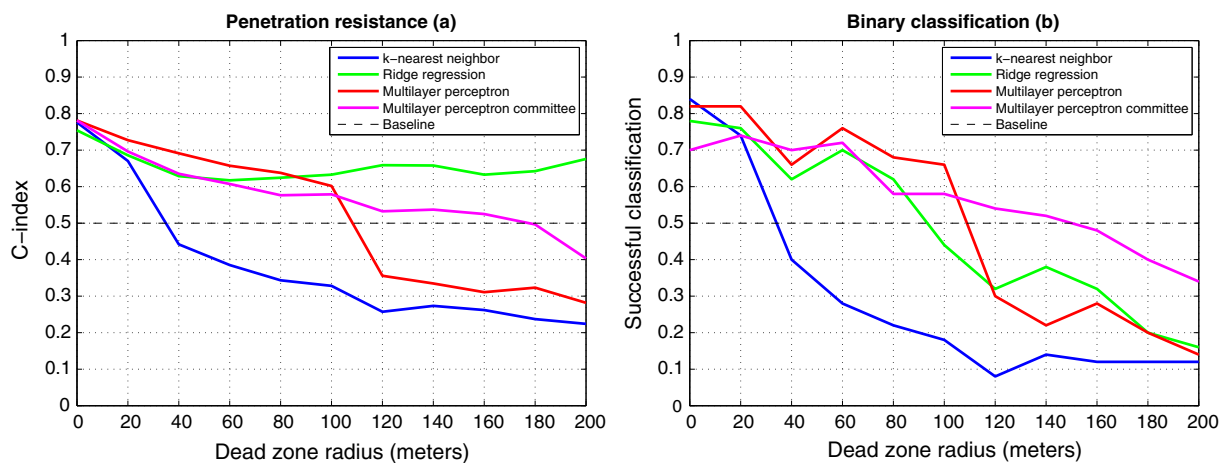


Fig. 4. Prediction results for case 2: soil penetration resistance (a) and soil bearing capacity binary classification (b). Ridge regression and multilayer perceptron achieve the highest results up to 100 m range. For multilayer perceptron (MLP) we have more than 70% classification accuracy up to 40 m range.

experiment indicated the need for real-time automatic sensors for harvesters in order to produce sufficiently accurate input data to produce applicable prediction rates. A total of 50 penetration resistance profiles were collected from two test sites. The test sites were selected to have differing properties in terms of penetration resistance. The first site was located on mineral soil with good bearing capacity, whereas the second was partly covered by a layer of peat. Based on physical measurements, the quality of the data was much higher than that of the soil damage data. We have listed the used data sets for this case in Table 2. In the analysis, the predictor data sets consisted of 50 variables. Also in this case we divided the prediction of soil penetration resistance into two approaches. In the first approach we implemented a regression model for the penetration resistance variable. In the second approach we divided the data into two classes based on the following criterion:

$$\text{Class of data point } \mathbf{x}_i = \begin{cases} 1 & \text{if } y_i \geq 5000 \text{ kPa} \\ 0 & \text{if } y_i < 5000 \text{ kPa} \end{cases}$$

where \mathbf{x}_i is the i th data point with elements corresponding to feature values and y_i is the corresponding value of penetration resistance for that point. In this case the problem was to classify a data point \mathbf{x}_i either into class 1 or 0. We have illustrated the results for both of these approaches in Fig. 4. We can notice that the results are much better than in it was in case 1. C-index stays above 0.6 up to 100 m for multilayer perceptron and ridge regression. Ridge regression stays above 0.6 up to 200 m. In classification case we got 66% classification accuracy up to 100 m by MLP and almost 70% accuracy up to 40 m by MLP-ESC.

6. Conclusion

The results indicated moderate prediction rates up to 20 m for the soil damage regression case. After 20 m the prediction performance drops dramatically and a random yes/no guess produces better results. C-index value stays just above the baseline 0.5 up to 200 m for the ridge regression model. In the soil damage classification case we achieved more than 60% classification rate up to 20 m as well. We therefore conclude that more precise measurements are needed for modelling purposes. In the case of penetration resistance prediction we achieved a C-index higher than 0.6 up to 200 m for ridge regression model. With soil bearing capacity classification we achieved more than 66% successful classification up to 100 m by MLP. Up to 20 m we achieved classification accuracy of more than 80% also by MLP. The better results in the case of penetration resistance data is explained by the higher quality of the used data sets because the data samples were based on physical measurements.

As a summary of the results, in case 1; the soil damage prediction remains moderate up to 20 m after which the result drop close to baseline value. In case 2; the soil pen-

etration resistance prediction the results remain very good up to 20 m, good up to 100 m after which the results start to drop below baseline. It was evident that the used data sets in penetration resistance case was more reliable and contained less noise than the data sets used in soil damage case. This points out the necessity of accurate and real-time measurements in order to produce applicable forecast models for harvesting operations. If the data quality is not high enough the prediction performance deteriorates rapidly. However the measured data sets were rather small, which suggests further analysis in more varied environments in the future.

We conclude, that more detailed field data is required, i.e. physical measurements and detailed information about the motions of the machinery within the stand, since for example the accumulation of traversed mass over each location is one of the main variables explaining soil damages (Sirén et al., 2013). These can be achieved through online learning based on trafficability data accumulated by harvesters or other field studies. Vertical distance to drainage network should also be tested (Murphy et al., 2009). Further validation with weather data and water budget models should be continued in the future as it is one of the key variables affecting trafficability of fine grained mineral and organic soils.

Acknowledgments

The study was carried out in *New Computational Methods for Effective Utilization of Public Data* (ULJATH)-project funded by the Finnish Funding Agency for Innovation (TEKES).

References

- Abu-Mostafa, Y., Magdon-Ismael, M., Lin, H., 2012. Learning from Data, vol. 1. AMLBook.
- Äijälä, O., Koistinen, A., Sved, J., Vanhatalo, K., Väisänen, P., 2014. Tapio - Hyvän metsänhoidon suositukset: Forest Management Practice Recommendations, vol. 1. Metsäkustannus Ltd.
- Ala-Ilomäki, J., 2013. Spiked shear vane – a new tool for measuring peatland top layer strength. *Mires Peat* 64 (2–3), 113–118.
- ArcMap, 2014. ArcGIS Platform, ESRI. <<http://www.esri.com/software/arcgis>>.
- ASTM D2216-10, 2010. Standard Test Methods for Laboratory Determination of Water (Moisture) Content of Soil and Rock by Mass. ASTM International, West Conshohocken, PA.
- Azzalini, A., Diggle, P., 1994. Prediction of soil respiration rates from temperature, moisture and soil type. *J. Roy. Stat. Soc. Ser. C: Appl. Stat.* 43 (3), 505–526.
- Beven, K., Kirkby, M., 1979. A physically based, variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* 24 (1), 43–69.
- Bishop, C., 1996. *Neural Networks for Pattern Recognition*. Oxford University Press.
- Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13 (1), 21–27.
- Feichtinger, H., Strohmer, T., 1997. *Gabor Analysis and Algorithms: Theory and Applications*. Birkhäuser.
- Gönen, M., Heller, G., 2005. Concordance probability and discriminatory power in proportional hazards regression. *Biometrika* 92 (4), 965–970.

- Gustafsson, L., Perhans, K., 2010. Biodiversity conservation in swedish forests: ways forward for a 30-year-old multi-scaled approach. *AMBIO: J. Hum. Environ.* 39 (8), 546–554.
- Hänninen, H., Karppinen, H., 2010. Metsänomistusrakenteen muutos ja puuntarjonta: change in forest ownership and wood supply, finnish forest sector economic outlook. Tech. rep., Finnish Forest Research Institute.
- Hyvönen, E., Päänttjä, M., Sutinen, M.-L., Sutinen, R., 2003. Assessing site suitability for scots pine using airborne and terrestrial gamma-ray measurements in finnish lapland. *Can. J. For. Res.* 33 (5), 796–806.
- Muro, T., O'Brien, 2004. *Terramechanics: Land Locomotion Mechanics*. CRC Press.
- Murphy, P., Ogilvie, J., Arp, P., 2009. Topographic modelling of soil moisture conditions: a comparison and verification of two models. *Eur. J. Soil Sci.* 60 (1), 94–109.
- Nabney, I., 2004. *NETLAB: Algorithms for Pattern Recognition*, vol. 1. Springer.
- NLS, 2014. NLS Topographic database, National Land Survey of Finland. <<http://www.maanmittauslaitos.fi/en/opendata>>.
- Pennanen, O., Mäkelä, O., 2003. Raakapuukuljetusten kelirikkohaittojen vähentäminen, Metsätehon raportti. Tech. Rep. 153, Metsäteho Ltd.
- Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T., 2011. *Computer Vision Using Local Binary Patterns, Computational Imaging and Vision*. Springer.
- Pohjankukka, J., Nevalainen, P., Pahikkala, T., Hyvönen, E., Sutinen, R., Hänninen, P., Heikkonen, J., 2014a. Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. In: Borga, M., Heyden, A., Laurendeau, D., Felsberg, M., Boyer, K. (Eds.), *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014)*. IEEE, pp. 1822–1827.
- Pohjankukka, J., Nevalainen, P., Pahikkala, T., Hyvönen, E., Middleton, M., Hänninen, P., Ala-Ilomäki, J., Heikkonen, J., 2014b. Predicting water permeability of the soil based on open data. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (Eds.), *Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014), IFIP Advances in Information and Communication Technology*, vol. 436. Springer, pp. 436–446.
- Ring, E., Löfgren, S., Bergkvist, I., Högbom, L., 2006. Många bäckar små.. Tech. Rep. 2, Skogforsk.
- Schulte, R., Diamond, J., Finkele, K., Holden, N., Brereton, A., 2005. Predicting the soil moisture conditions of Irish grasslands. *Irish J. Agric. Food Res.* 44, 95–110.
- Seibert, J., McGlynn, B., 2007. A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models. *Water Resour. Res.* 43 (4). na–na, w04501.
- Sirén, M., Ala-Ilomäki, J., Mäkinen, H., Lamminen, S., Mikkola, T., 2013. Harvesting damage caused by thinning of norway spruce in unfrozen soil. *Int. J. Forest Eng.* 24 (1), 60–75.
- Tomppo, E., Katila, M., Mäkisara, K., Peräsaari, J., 2008. *Multi-source national forest inventory methods and applications. Managing Forest Ecosystems*, vol. 18. Springer.
- Uusitalo, J., 2010. *Introduction to Forest Operations and Technology*, vol. 1. Metsäkustannus Ltd.
- Vapnik, V., 1998. *Statistical Learning Theory*, vol. 1. Wiley-Interscience.
- Wood, J., 1996. *The geomorphological characterisation of digital elevation models (Ph.D. thesis)*. University of Leicester.
- Wood, J., 2009. *Geomorphometry in landserf*. In: Hengl, T., Reuter, H. (Eds.), *Developments in Soil Scie*, vol. 33, pp. 333–349. 12.
- Zevenbergen, L., Thorne, C., 1987. Quantitative analysis of land surface topography. *Earth Surf. Proc. Land.* 12 (1), 47–56.

Publication IV

Estimating the prediction performance of spatial models via spatial k-fold cross validation

Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen and Jukka Heikkonen. *International Journal of Geographical Information Science*, 31(10):2001–2019. Taylor & Francis, 2017.

Copyright © 2017 Taylor & Francis. Reprinted with permissions from respective publisher and authors.

RESEARCH ARTICLE



Estimating the prediction performance of spatial models via spatial k -fold cross validation

Jonne Pohjankukka , Tapio Pahikkala , Paavo Nevalainen and Jukka Heikkonen

Department of Future Technologies, University of Turku, Turku, Finland

ABSTRACT

In machine learning, one often assumes the data are independent when evaluating model performance. However, this rarely holds in practice. Geographic information datasets are an example where the data points have stronger dependencies among each other the closer they are geographically. This phenomenon known as spatial autocorrelation (SAC) causes the standard cross validation (CV) methods to produce optimistically biased prediction performance estimates for spatial models, which can result in increased costs and accidents in practical applications. To overcome this problem, we propose a modified version of the CV method called *spatial k -fold cross validation* (SKCV), which provides a useful estimate for model prediction performance without optimistic bias due to SAC. We test SKCV with three real-world cases involving open natural data showing that the estimates produced by the ordinary CV are up to 40% more optimistic than those of SKCV. Both regression and classification cases are considered in our experiments. In addition, we will show how the SKCV method can be applied as a criterion for selecting data sampling density for new research area.

ARTICLE HISTORY


Received 27 June 2016
Accepted 20 June 2017


KEYWORDS

Spatio-temporal data modelling; spatial data mining; geographic information systems; geographic information science

1. Introduction

An important step in machine-learning applications is the evaluation of the prediction performance of a model in the task under consideration. For this one can use the k -fold cross validation (CV), which assumes that the data are independent. Geographic information system (GIS) datasets represent an example where the independence assumption naturally does not hold due to the temporal or spatial autocorrelation (SAC). SAC and its effects on spatial data analysis have been extensively studied in spatial statistics literature (Legendre 1993, Koenig 1999). For example, it has been shown that the failure to not account the effect of SAC in spatial data modeling can lead to over-complex model selection (Hoeting *et al.* 2006, Rest *et al.* 2014). Generally speaking, natural data exhibits SAC because of the first law of geography and fundamental principle in geostatistical analysis according to Waldo Tobler (Tobler 1970): ‘Everything is related to everything else, but near things are more related than distant things’. In spatial statistics, the degree of SAC of a dataset

CONTACT Jonne Pohjankukka  jjepoh@utu.fi

 Supplemental data for this article can be accessed [here](#).

© 2017 Informa UK Limited, trading as Taylor & Francis Group

can be measured using e.g. a semivariogram (Cressie 2015b), Moran's I (Moran 1950), Geary's C (Geary 1954) or Getis's G (Getis and Ord 1992).

There are numerous applications involving spatial data which have problems caused by SAC in the datasets such as natural resource detection, route selection, construction placement, natural disaster recognition, tree species detection, environmental monitoring, etc. (Ala-Illomäki *et al.* 2015). Consider the example of harvesting operations in forestry where optimal route selections are of key importance. In order to minimize the risk of harvester sinking into the soil, a route with the optimal carrying capacity is required. The route selection is based on predictions of soil types along the route which gives the harvester an estimate on the carrying capacity of the route. If the effect of SAC is not considered in the soil-type predictions while estimating the model performance, we might end up selecting a hazardous route. The reason for this is that the spatial model we are using gives overoptimistic prediction performance for soil types farther away from the harvester's current location. The model implicitly assumes that we have known soil types close to the predicted soil types which is not always the case. This fact must be taken into account in the model-prediction performance evaluation in order to avoid overoptimistic estimation. An illustration of the considered example is shown in Figure 1.

To counter the problems caused by SAC in spatial modeling, one usually tries to incorporate SAC as an autocovariate factor into the prediction models themselves, e.g. autocovariate models, spatial eigenvector mapping, autoregressive models (Lichstein *et al.* 2002, Diniz-Filho *et al.* 2003, Brenning 2005, Bahn *et al.* 2006, Dormann *et al.* 2007, Betts *et al.* 2009, Beale *et al.* 2010, Zhang and Wang 2010). A review of such methods is well presented by Dormann *et al.* (2007). Other methods include spatial clustering and resampling techniques for countering SAC (Ruß and Kruse 2010, Brenning 2012, Hijmans 2012). Despite the vast literature of techniques for spatial prediction, little attention is given for assessing the spatial prediction performance of a model via CV techniques. In

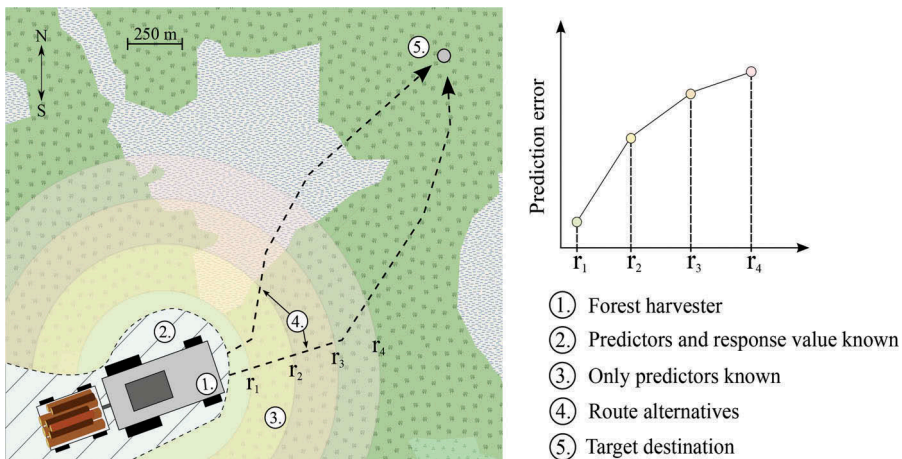


Figure 1. The forest harvesting example. The harvester driver needs to select an optimal route to target destination. Due to SAC, it is to be expected that the prediction error increases the further away we make point predictions. The background map in the image (also in Figure 8(a)) is by the courtesy of OpenStreetMap.

the work by Cressie (2015a), the author does not advocate CV for confirmatory data analysis because the independence assumption in the data samples is inherently not valid in geostatistical context.

In this article, we propose a novel CV method called *spatial k-fold cross validation* (SKCV) for estimating prediction performance under SAC-based independence violations in the data. SKCV is also applicable for selecting grid sampling density for new research areas. More specifically, SKCV attempts to answer the following two questions:

- (1) What is the prediction performance of a model at a certain geographical location when the closest data measurements used to train it lie at a given geographical distance?
- (2) Conversely, if the prediction performance is required to be at least a given level, how dense data sampling grid should be used in the experiment area to achieve it?

The question (2) is about the trade-off between the prediction performance and data collection costs. The SKCV method provides the model prediction performance as the function of geographical distance between the in-sample and out-of-sample data, and hence it indicates how close geographically training data has to be to the prediction area in order to achieve a required prediction performance. The idea in SKCV is to remove the optimistic bias due to SAC by omitting data samples from the training set, which are geographically too close to the test data.

To evaluate how well SKCV answers the above questions, it is tested with three real world applications using public GIS-based datasets. The applications involve assessing the predictability of water permeability of soil and forest harvest track damage. Both regression and classification models were used in these experiments. The usability of the SKCV method for determining the needed sampling grid density is tested by measuring the difference between the performance of model constructed with a given grid density and the result predicted with SKCV. We will explain this comparison in more detail in [Section 4.1](#).

We wish to emphasize that we use SKCV in this manuscript for assessing the spatial prediction performance of a model and not for model complexity selection even though model complexity selection can also be applied with SKCV. In the work by Rest *et al.* (2014), the authors used a similar spatial CV method as SKCV for model variable selection. In their work, they compared a special case of SKCV method, the spatial leave-one-out (SLOO) method with Akaike information criterion (AIC) (Akaike 1998) as a criterion for model variable selection. It turned out that SAC caused the AIC to select biased variables, whereas SLOO prevented this. In the work by Pohjankukka *et al.* (2014a, 2014b, 2016), the SKCV method was called *cross validation with a dead zone* method. Related studies on spatial data analysis can also be found in the works of Azzalini and Diggle (1994), Schulte *et al.* (2005) and Brenning (2012).

In what follows, a formal description of the SKCV method will be given in [Section 2](#), followed by description of used datasets in [Section 3](#) and experimental analyses with three sample cases in [Section 4](#), and finally [Section 5](#) includes conclusions.

2. Spatial k-fold cross validation

SKCV is a modification of the standard CV to overcome the biased prediction performance estimates of the model due to SAC of the data. The overoptimistic bias in the performance estimates is prevented by making sure that the training dataset only contains data points that are *at least* a certain spatial or temporal distance away from the test dataset.

We will denote our data point as $\mathbf{d}_i = (\mathbf{x}_i, y_i, \mathbf{c}_i)$, where $\mathbf{x}_i \in \mathbb{R}^n$ is a feature vector, $y_i \in \mathbb{R}$ a response value and $\mathbf{c}_i \in \mathbb{R}^2$ the geographical coordinate vector of i th data point. The dataset is denoted as $\mathcal{D} = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_M\}$. The value $r_\delta \in \mathbb{R}^+$ is the so-called *dead zone* radius, which determines the data points to be eliminated from the training dataset at each SKCV iteration. The set $\mathcal{V} = \{\mathcal{V}_1, \dots, \mathcal{V}_K\}$ is the set of CV folds, where each $\mathcal{V}_p \subset \mathcal{D}$ and $\mathcal{V}_p \cap \mathcal{V}_q = \emptyset$, when $p \neq q$ and $\cup_{p=1}^K \mathcal{V}_p = \mathcal{D}$. The training of the model is performed by a learning algorithm \mathcal{A} . The vector $\hat{\mathbf{y}} \in \mathbb{R}^M$ denotes the predicted response values by a prediction model \mathcal{F} . Note that the choice of \mathcal{F} does not affect the functionality of SKCV. We use the standard Euclidean distance e to calculate the *spatial distance* between two data points \mathbf{d}_i and \mathbf{d}_j . A formal presentation of the SKCV method is given in Algorithm 1. When the number of folds K equals the number of data points M , SKCV becomes SLOO. The SKCV algorithm is almost identical to normal CV with the exception of the reduction of the training set depicted in Figure 2 and in line 2 of Algorithm 1. In particular, when $r_\delta = 0$, SKCV reduces to normal CV.

Algorithm 1. Spatial k-fold cross validation

Require: $\mathcal{V}, \mathcal{D}, \mathcal{A}, r_\delta$

Ensure: $\hat{\mathbf{y}}$

1: **for** $i \leftarrow 1$ to K **do**

2: $\mathcal{H} \leftarrow \cup_{\mathbf{d}_k \in \mathcal{V}_i} \{\mathbf{d}_j \in \mathcal{D} \mid e(\mathbf{c}_j, \mathbf{c}_k) \leq r_\delta\}$ ▷ Remove data points too close

3: $\mathcal{F} \leftarrow \mathcal{A}(\mathcal{D} \setminus \mathcal{H})$ ▷ Build model using reduced training set

4: **for** $\mathbf{d}_k \in \mathcal{V}_i$ **do**

5: $\hat{\mathbf{y}}[k] \leftarrow \mathcal{F}(\mathbf{x}_k, \mathbf{c}_k)$ ▷ Make prediction

6: **end for**

7: **end for**

8: **return** $\hat{\mathbf{y}}$ ▷ The predicted $\hat{\mathbf{y}}$

There are three issues one might consider with SKCV which we will address here. First, since SKCV may involve removal of a large number of training data, this may introduce an extra pessimistic bias on the prediction performance not related to SAC. The size of this bias can be estimated via experiment in which one removes the same amount of randomly selected data from the training set on each CV round. Our experimental results in later sections confirm that the performance decrease observed by doing this is negligible compared to the one caused by SAC removal.

Second, the above considered issue becomes far more severe if the number of SKCV folds K is very small (say $K = 2$). It could happen that most of the training data is removed because the combined dead zones of the test data points will have a large *effective radius*. This concern is application-specific and the selection of the SKCV folds must be designed to suit the purposes of the application. For example, with a sparse

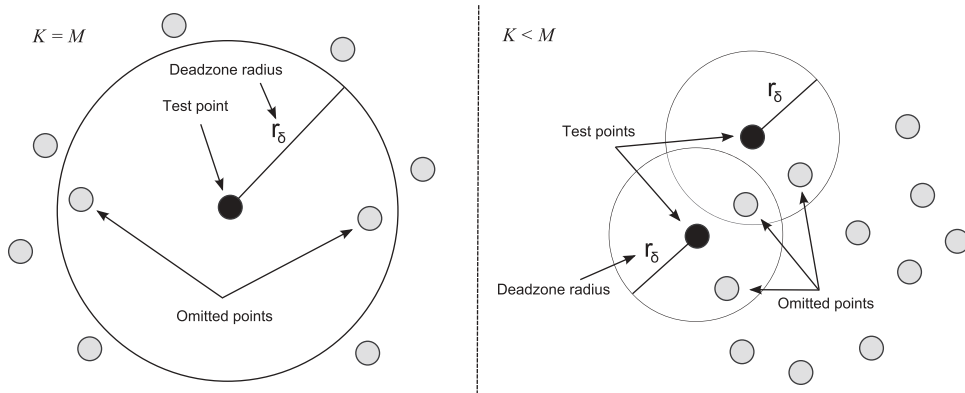


Figure 2. Reduction of the training set in the SKCV procedure. The black and gray points correspond to test and training data points. The gray data points inside the perimeters of radius r_δ are omitted from the training data, after which the test points are predicted using the remaining training data (i.e. the gray data points outside the perimeters).

dataset it would make a little practical sense to select the SKCV folds in such a way that all the training data is removed. For these reasons it is best to have $K = M$, which corresponds to the SLOO case of SKCV if computational resources allow it.

Third, one could ask whether the prediction performance for a given r_δ could be estimated by analyzing the prediction error obtained with, say, leave-one-out cross-validation simply as a function of the average distance to closest neighbors. While, this could be doable with datasets having both densely and sparsely measured areas, the data points in many available datasets tend to be much closer to each other than in the case we intend to simulate. For example, with a dense dataset with a maximum distance of 3 m between a data point and its nearest neighbors, one cannot simulate performing prediction for a data point having the closest measurements at least 25 m away.

Finally, let us consider the difference between spatial interpolation and regression. In the former, the only extra information available about the training data are their coordinates c , while in the latter one also has access to an additional information in the form of feature representation \mathbf{x} . However, the SKCV algorithm works in a similar way in both cases, as it is independent on the type of information the learning algorithms use for training a model or what the model uses for predicting the responses for new points.

3. Datasets

The three experimental cases differ on the availability and resolution of the datasets. In case 1 related to water permeability, prediction data were available throughout the research area, with the exception of areas where there were obstacles (e.g. buildings or lakes). In case 2 also related to water permeability prediction, there were scattered field measurement data and in case 3 related to harvester track damage prediction, the dataset was clustered into several areas. These cases are typical of common types of spatial prediction applications. The availability of the datasets in three cases is illustrated in Figure 3. The data range from remote sensing datasets such as satellite and airborne

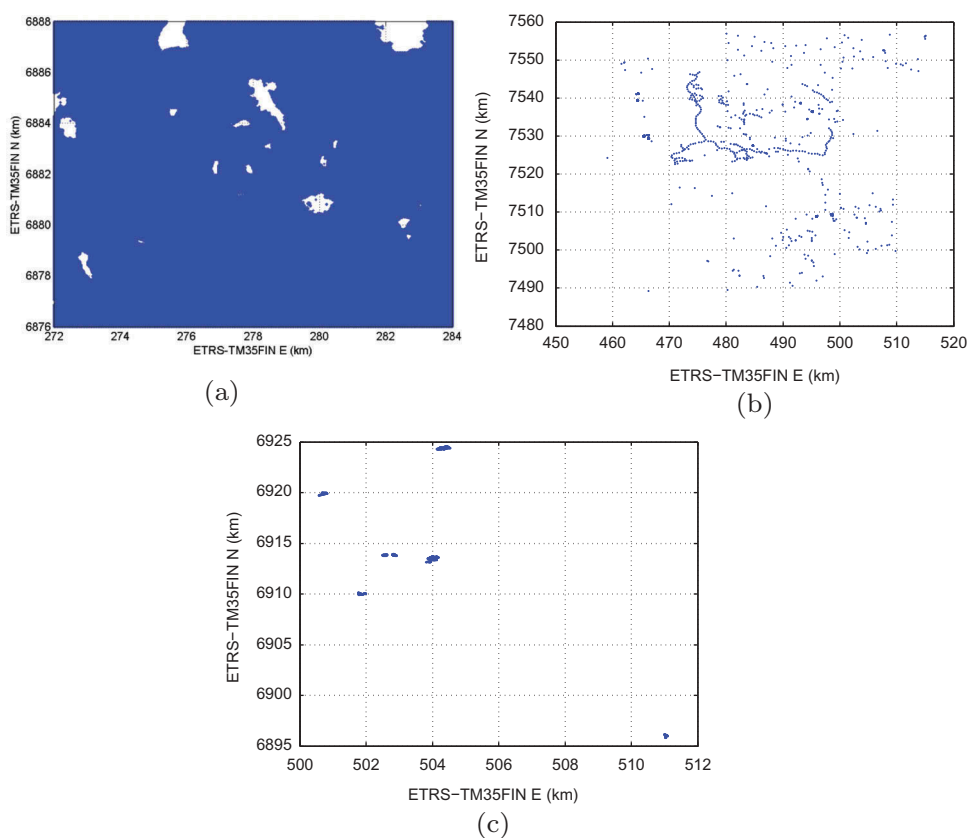


Figure 3. Coverage of data on experimental cases 1–3. (a) case 1: 361,201 data points, (b) case 2: 1691 data points, (c) case 3: 11,795 data points. Blue areas correspond to areas where data were available. The axes correspond to locations in ETRS-TM35FIN coordinates in kilometers. The dataset in case 1 is much more dense than datasets in cases 2 and 3.

imaging raster data to manually on-site collected samples of the soil (Zevenbergen and Thorne 1987, Wood 1996, Hyvönen *et al.* 2003, Tomppo *et al.* 2008, Pohjankukka *et al.* 2014a, 2014b, 2016). The formats of the datasets are TIFF-images and ASCII-files with different resolutions. A summary of the used datasets in the three cases is illustrated in Table 1. In the following paragraphs we briefly describe the used datasets. More detailed illustration of the datasets is given in the supplementary material.

3.1. Digital elevation model data (DEM)

We downloaded DEM data from the file service for open data by the National Land Survey of Finland (NLS). The DEM was made from airborne laser scanning data with grid size of 2 m. Several geomorphometric variables were derived from the NLS DEM in SAGA GIS environment. In our analysis, we used the geomorphometric features: plan curvature, profile curvature, slope, topographic wetness index, flow area, aspect, diffuse insolation and direct insolation (Beven and Kirkby 1979, Zevenbergen and Thorne

Table 1. Summary of the used datasets in all experimental cases. Response value datasets are listed in emphasized form. Also, the data format is shown either as TIFF-raster image or ASCII-vector file and grid resolution size in meters.

Dataset	Format	Grid size	Case 1	Case 2	Case 3
Digital Elevation Model	Raster	2 m	X	X	X
Multisource National Forest Inventory	Raster	20 m	X	X	X
Gamma-ray spectroscopy	Raster	50 m	X	X	X
Air-borne electromagnetic	Raster	50 m		X	
Peatland	Raster	1000 m		X	X
Weather	Raster	10000 m			X
Stoniness	Vector	–			X
Soil moisture	Vector	–			X
Water permeability exponent	Vector	–	X	X	
Harvester track damage	Vector	–			X

1987, Wood 1996, 2009, Seibert and McGlynn 2007). These derived features are more efficient for prediction than raw height data alone.

3.2. Multisource national forest inventory data

A selected set of 43 features of the state of the Finnish forests in 20 m grid size are available as the MultiSource National Forest Inventory (MS-NFI) by Natural Resources Institute Finland (LUKE). The MS-NFI dataset is derived by interpolating field measured MS-NFI samples using inverse distance-weighted k -nearest neighbor (kNN) method as the interpolation algorithm and Landsat imagery combined with DEM data as the basis of interpolation. Features include e.g. the biomass and volume of growing stock. The MS-NFI data features exhibit built-in dependencies which means the number of useful features is lower than 43. A detailed description of the MS-NFI is the work by Tomppo *et al.* (2008).

3.3. Aerial gamma-ray spectroscopy data

The aerial gamma-ray flux of potassium (K) decay with the grid size of 50 m is provided by the Geological Survey of Finland (GTK). These data are related to e.g. the moisture dynamics, frost heaving (Hyvönen *et al.* 2003) and density, porosity and grain size of the soil. High gamma radiation indicates lower soil moisture and vice versa. Several statistical and textural features were derived from the gamma-ray data. These include: 3×3 windowed mean and standard deviation, Gabor filter features (Feichtinger and Strohmmer 1997) and Local Binary Pattern features (Pietikäinen *et al.* 2011).

3.4. Peatland data

The peatland data are provided by LUKE and uses topographic information provided by NLS. The peatland data are a binary raster mask of 1000 m grid size with values 0/1 corresponding to non-peatland/peatland areas. The peatland mask is derived from four NLS topographic database elements depicting different types of peatlands. The mask bit

1 refers to a spot where the location is mostly covered by peatland vegetation and the peat thickness exceeds 0.3 m over a local area of 1000 m².

3.5. Air-borne electromagnetic (AEM) data

The AEM data were provided by the GTK. The apparent resistivity indicates the soil-type factors, e.g. grain size distribution, water content and quality in the soil and cumulative weathering.

3.6. Weather data

Weather data on temperature (C) and rainfall (mm) for years 2011–2013 were provided by the Finnish Meteorological Institute (FMI). The grid size of the dataset was 10 km. We used the mean temperature and rainfall of the last 30 days at each observation point of the response value.

3.7. Stoniness data

Stoniness was estimated by steel-rod sounding (Tamminen 1991). The rod was pushed into the soil where the penetration depth and stone hits were recorded.

3.8. Soil moisture data

Gravimetric soil water content was measured from the samples by drying the soil samples and calculating the weight difference of dry and wet soil sample (ASTM D2216-10 2010).

3.9. Water permeability exponent data

Water permeability indicates the nominal vertical speed of water through the soil sample. This feature was measured indirectly by observing the soil particle size distribution. The actual speed depends on inhomogenities (roots, rocks) and micro-cracks in the soil. The water permeability exponent is a logarithmic quantity y derived from water permeability speed v .

3.10. Harvester track damage data

Approximately 36 km of strip roads were traversed by Metsäteho Ltd. and visually assessed into damage classes by a forest operations expert. The soil damage classes used were: (1) no damage; (2) slight damage; and (3) damage. The original dataset required preprocessing by LUKE due to the inaccuracies in GPS-tracks. The strip road line segments were then converted to sample points used in the prediction process.

4. Experimental analysis with SKCV

In this section, the SKCV method is applied to three real world cases involving GIS data making them suitable to illustrate the proposed method. In the first two cases, the water permeability levels of boreal soil are predicted and in the final case the damage caused by movements of a forest harvester. The experiments provide useful results e.g. for forest industry where it is crucial to have accurate and optimistically unbiased prediction performance for soil conditions. It is estimated that forest industry in Finland alone has yearly costs of approximately 100 million euros caused by challenging trafficability conditions of the soil which increase time and fuel consumption and decrease the efficiency of timber harvesting operations (Pennanen and Mäkelä 2003, Sirén *et al.* 2013). These costs could be decreased by additional information on soil conditions, especially soil-bearing capacity by utilizing public GIS data.

The research question (1) will be addressed in cases 1, 2, 3 (sections 4.1, 4.2, 4.3) and the research question (2) will be addressed in case 1 (section 4.1). In all experimental cases, k -nearest neighbor (kNN) algorithm was used as the prediction model \mathcal{F} and the predictor features \mathbf{x}_i were z-score standardized. While there are many alternative prediction methods, the choice does not have an effect on the presence of SAC in the data. Therefore, kNN was selected due to its simplicity. As a distance function that determines the nearest neighbors, we use the Euclidean distance for the feature vectors \mathbf{x}_i . Note that this is in contrast to the spatial distance e used in SKCV. We implemented the analyses using k -values of $\{1, 3, 5, 7, 9, 11, 13, 15\}$ for kNN. The general behavior of SKCV results was similar for all tested k -values and for this reason we only report the results with $k = 9$. The performance measures used in the experiments were the standard root mean squared error (RMSE) for kNN regression (Araghinejad 2014, pp. 66–73) and classification accuracy for kNN classification. In cases 1 and 2 (regression), the predicted response value \hat{y}_i is defined as the average value of kNNs and in case 3 (classification) the mode of the kNNs.

The semivariograms and Moran's I statistics were calculated for the response variables y_i in all experimental cases to confirm the presence of SAC in the data. In a semivariogram, a variable X is spatially autocorrelated at a given distance range $[m - t, m + t] \subset \mathbb{R}^+$ with lag tolerance $t \in \mathbb{R}^+$ if its semivariogram value $\gamma_t(m) \in \mathbb{R}^+$ is lower than the sill value of the variable X (Cressie 2015b). The lag tolerance t gives us the maximum allowed deviation from $m \in \mathbb{R}^+$ when the distance between two data points is still considered to be m meters (see Figure 4). For example, if $m = 10$ meters and $t = 1$ meter, then the semivariogram value $\gamma_1(10)$ for a single data point \mathbf{d}_j is calculated from the set $\Gamma = \{\mathbf{d}_j \in \mathcal{D} \mid e(\mathbf{c}_j, \mathbf{c}_i) \in [9, 11]\}$. In other words, the data points in set Γ are considered to be 10 m away from \mathbf{d}_j . This is rarely exactly the case and hence we have to use the lag tolerance t . The lag tolerance values in the experimental cases were selected to suite the resolution of the corresponding data. In the Moran's I autocorrelation plots, we call *baseline* the 0 correlation.

4.1. Case 1: soil water permeability prediction based on soil type

In this section, we will consider the predictability of the soil water permeability levels based on the soil type. The response variable in this case is the water permeability

Calculation of semivariogram

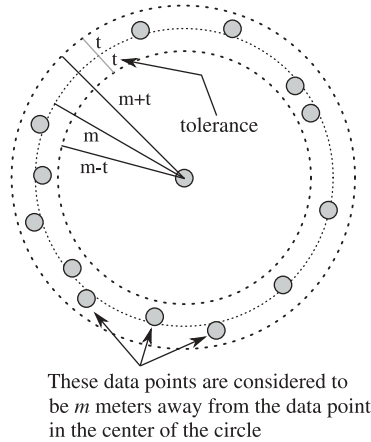


Figure 4. Calculation of the semivariogram. Data points within a distance range $[m - t, m + t]$ are considered to be m meters away from the center data point.

exponent value $y \in \mathbb{R}$, which is related to both the boreal soil type and the water permeability itself. The exact relation between these two factors is presented by Pohjankukka *et al.* (2014a). Optimal harvesting routes avoid areas with small water permeability, where soil tends to stay moist and there is an elevated risk for ground damage and logistic problems. A reliable estimate of the water permeability distribution is needed when making routing decisions during the preliminary planning phase and during the harvest operations. The aim here is to increase the efficiency and minimize the harvesting costs.

The target area is located in the municipality of Parkano, which is a part of the Pirkanmaa region of Western Finland. The size of the target area is approximately

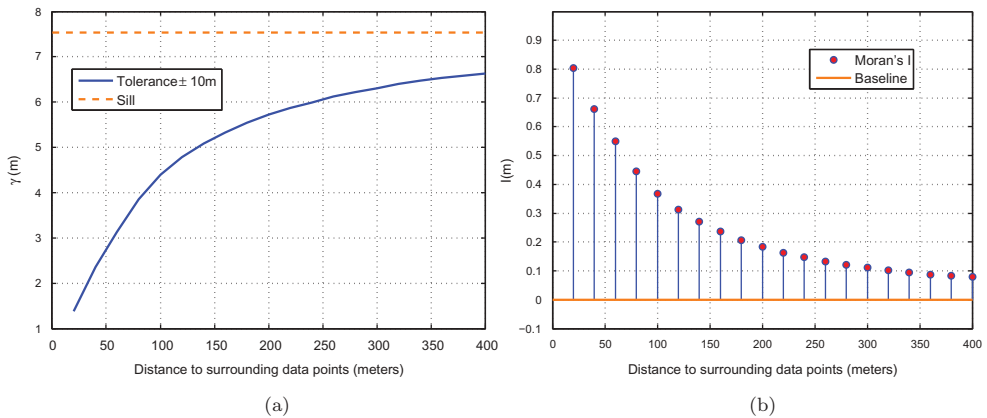


Figure 5. The semivariogram and Moran's I plot depicting the SAC of the water permeability exponent y in case 1. (a) Semivariogram showing that $\gamma(m)$ stays below the sill with $t = 10$ m. (b) Moran's I also revealing the presence of SAC in response value y .

144 km² (ETRS-TM35FIN coordinates at 278 km E, 6882 km N, zone 35). When considering all the features including the derived ones, we had a total of 49 predictor features in the dataset. In the analysis of case 1, a total of 361,201 data points were available. A summary of the datasets is illustrated in Table 1 of Section 3. In Figure 5 depicting the semivariogram and Moran's I plot for the water permeability exponent y , we can see a clear presence of SAC. The predicted water permeability exponent \hat{y}_i for the i th data point $\mathbf{d}_i = (\mathbf{x}_i, y_i, \mathbf{c}_i)$ using kNN regression is defined as

$$\hat{y}_i = \frac{1}{k} \sum_{y \in N_i} y \quad (1)$$

where N_i is the set of water permeability exponent values y of the kNNs of d_i .

The estimated prediction performance for 9-nearest neighbor (9NN) using SKCV is illustrated in Figure 6, which answers to research question (1) with various distance values r_δ . The spatial density in the results describes how many data points are in a given space, i.e. it describes the sparsity of the dataset. From Figure 6 we notice a clear rise in the prediction error (RMSE) when the distance between prediction point and training data increases. This was an expected result based on the SAC discovered in the semivariogram and Moran's I plots in Figure 5. With sparser datasets we notice the dead zone radius having a smaller effect on the results.

To measure how much the SKCV's performance decrease along the increasing dead zone radius is caused only by the decreased size of the training set, we implement additional analysis which we refer to as *SKCV random-leave-out* (SKCV-RLO). SKCV-RLO is identical to the SKCV method (see Algorithm 2 and Figure 2) with the exception that instead of removing data points from the training set that are too close to the test data, i.e. inside the dead zone perimeter, we instead remove the same number of data points randomly from the training set as we would remove in SKCV. In Figure 6, the

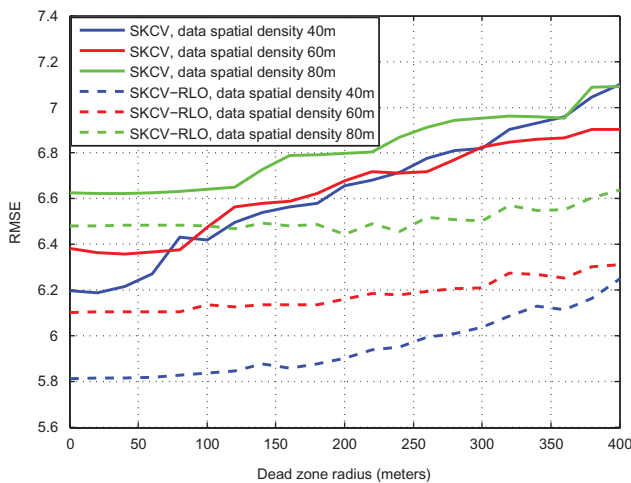


Figure 6. Prediction performance estimates for 9NN using SKCV and SKCV-RLO in case 1. The curves are plotted with three spatial densities to illustrate how the spatial density of the dataset affects the results.

estimated prediction performance for 9NN using SKCV-RLO is illustrated. On all spatial densities, we notice SKCV-RLO being less sensitive to the number of data points removed from the training set giving more optimistic results than SKCV. This reinforces our claim that the prediction algorithm prefers to use data points which are geographically close to the prediction point and shows that random removal of training data points causes negligible change in prediction accuracy when compared with SAC-based data removal.

Next, we focus our attention on research question (2), i.e. how densely we should sample data points from a new research area to achieve a given prediction level. Imagine that there are two distinct geographical areas which we refer to as areas *A* and *B*. In area *A*, there exists a dataset of measurements gathered from a certain subset of its coordinates but there are no measurements from area *B* yet. The aim is to perform a number of measurements from area *B* in order to construct a model for predicting the rest of the measurement values for every possible point in area *B*. Performing measurements used to form a training set is expensive, and hence their number should be minimized under the constraint that at least a given prediction performance level is required. This trade-off between the number of training measurements and prediction performance is not known in advance and our hypothesis is that it can be estimated with SKCV on the existing data from area *A*. Namely, if the prediction performance estimate provided by SKCV with dead zone radius r_δ on area *A* is as good or better than the required performance level, we hypothesize that we obtain as good prediction performance in area *B* if we guarantee that the closest measurement points are at most at a distance of r_δ from every point in area *B*. Given this constraint, the number of measurement points in area *B* is minimized via hexagonal sampling (see e.g. Donkoh and Opoku 2016). To support our hypothesis (i.e. using SKCV to estimate the trade-off

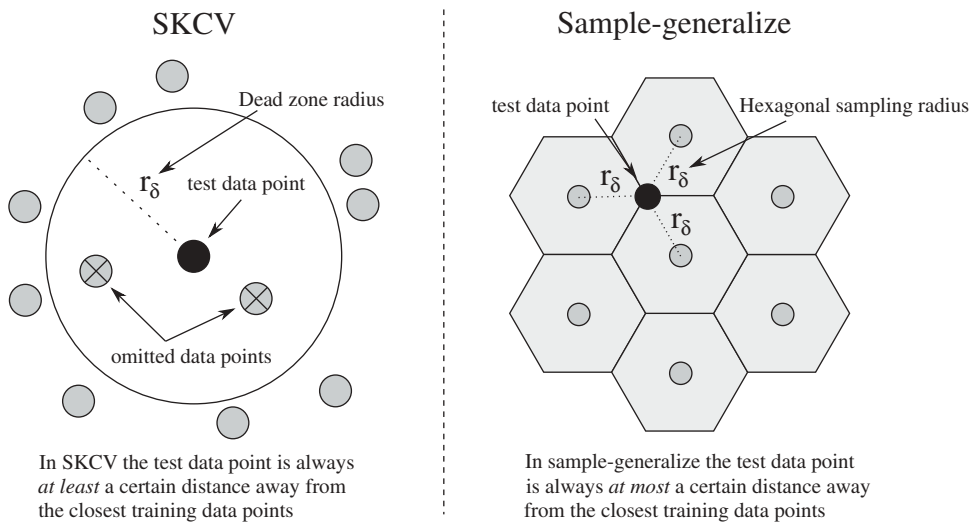


Figure 7. *Left:* In SKCV the test point is always at least r_δ meters away from training data. *Right:* In sample-generalize procedure, we sample data points (the gray points) using a hexagonal grid and predict the rest of the area around the sampled points. The black point represents a prediction point where the distance to training data is maximum, i.e. r_δ meters.

between number of measurement points and prediction performance), we use an auxiliary method called *sample-generalize*. In the sample-generalize procedure, we firstly sample training data points hexagonally (e.g. measure their response variables) with sampling radius r_δ , and secondly we use this data to train a model for predicting the responses from the rest of the area. Right side of Figure 7 illustrates the sample-generalize procedure. Note that SKCV is inherently more pessimistic than sample-generalize since the prediction point is always *at least* r_δ meters away from training data, whereas in sample-generalize the prediction point is always *at most* r_δ meters away from training data (see Figure 7).

In order to inspect the goodness of SKCV as an estimator of the prediction performance of sample-generalize, we implement a bias-variance analysis for nine smaller subareas formed using a 3×3 grid in the Parkano research area (see Figure 8(a)). We do this by firstly forming 72 (A, B) area pairs (from 3×3 grid we get $9 \times 8 = 72$ area pairs, i.e. each smaller area has 8 pair possibilities) from the nine smaller subareas. Second, for each of the area pairs (A, B), we calculate the prediction performance estimate with SKCV on area A ($result_A$) and the prediction performance of sample-generalize on area B ($result_B$) and then we take the difference of them ($result_A - result_B$). Lastly, we calculate the mean and standard deviation of the differences on the 72 area pairs. The resulting bias-variance plot is illustrated in Figure 8(b). From the plot, we see that the SKCV estimates tend to be pessimistically biased on the range $r_\delta \in [0, 150]$ m. In range $r_\delta \in [150, 340]$ m, the SKCV estimation is almost unbiased and in range $r_\delta \in [340, 400]$ m, it is optimistically biased. The results are pretty stable on all spatial densities for SKCV; the spatial density seems to shift the results simply by a constant value.

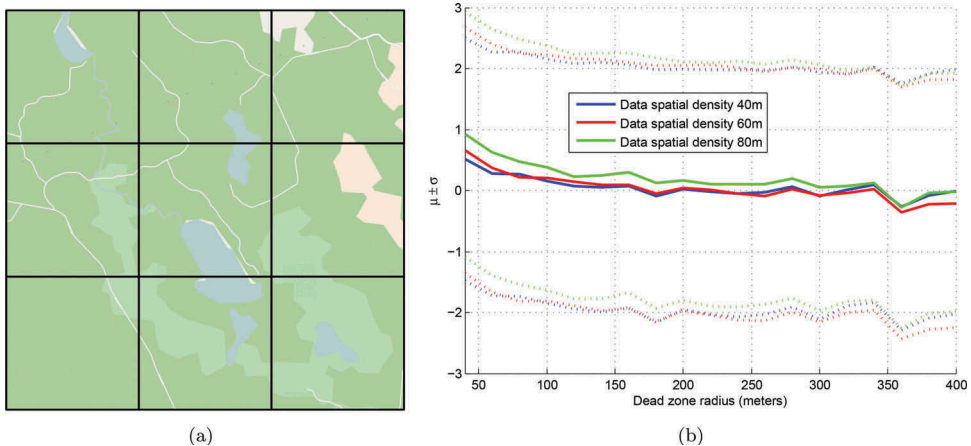


Figure 8. (a) Division of a research area into nine smaller subareas using a 3×3 grid. Each smaller area is 16 km^2 in size and consists from approximately 40,000 data points. (b) Bias-variance ($\mu \pm \sigma$) plot for the difference between the prediction performance estimate produced by SKCV and the actual prediction performance of sample-generalize of the 72 (9×8) area pairs. Solid curves represent the mean μ and dashed lines standard deviation σ . Different colors represent different spatial densities for the dataset in area A where SKCV is implemented.

4.2. Case 2: soil water permeability prediction based on field measurements

In this section, we consider the predictability of forest soil water permeability based on field measurement data. The difference between the response variables in cases 1 and 2 is that in case 1 the water permeability exponent γ is based on remote sensing data and in case 2, γ is based on field measurements. Semivariogram and Moran's I plot for the response variable is presented in Figure 9 which show clear SAC in the data. There is more variability in the SAC of case 2 than in case 1 but we must note that the dataset in case 2 was much smaller and more sparse.

The research area is located in Pomokaira, the northern part of the municipality of Sodankylä, which is a part of Finnish Lapland. The size of the target area is 18432 km². The center point of the rectangular target area is at ETRS-TM35FIN coordinates 7524 km N, 488 km E, zone 35. A total of 1691 data points were collected around the research area. The distances between the data points were much larger and they were not available from the entire research area when compared with the case 1 dataset. 102 feature variables were used for predicting the response value, i.e. the water permeability exponent γ . The used datasets in case 2 are shown in Table 1. The response variable γ is predicted in exactly the same way as in case 1 using kNN-regression in Equation (1).

Because the number of data points was significantly lower when compared with case 1, it was computationally feasible to implement SLOO and SLOO-RLO analyses on the data. The SLOO and SLOO-RLO results of case 2 are illustrated in Figure 10. The SLOO results show a clear drop in the prediction performance as the dead zone radius r_δ is increased. A high optimistic bias is observed from the SLOO-RLO results when compared with SLOO. The SLOO results indicate that the prediction performance decreases radically after the distance between test and training data is approximately 40–50 m. The effect of SAC can clearly be noted in these results.

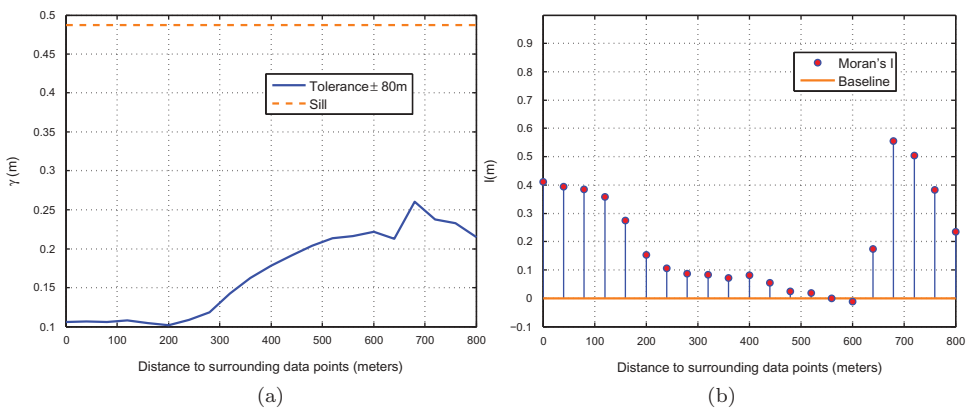


Figure 9. The semivariogram and Moran's I plot depicting the SAC of the response value of case 2. (a) Semivariogram with $t = 80$ m. (b) Moran's I.

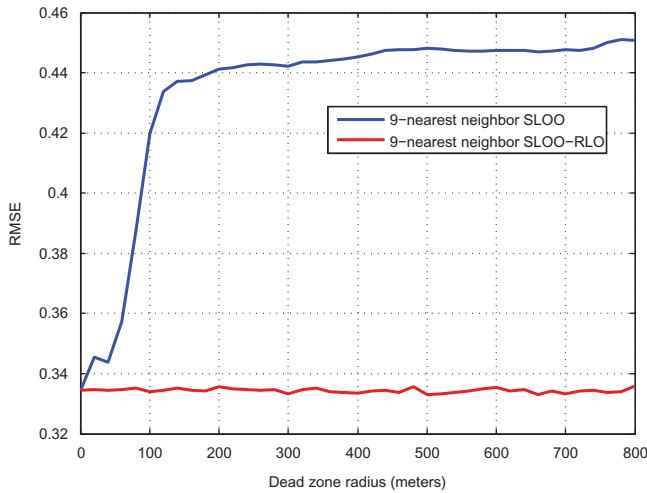


Figure 10. The SLOO and SLOO-RLO results in the Pomokaira analysis. The y -axis corresponds to the RMSE and x -axis to the length of dead zone radius r_δ .

4.3. Case 3: soil track damage classification

In this case, the goal is to assess the classification of forest harvester track damage. In other words, the task is to predict the damage that would occur to a soil point if a forest harvester drives through it. In particular, damage means the depression caused on the soil by the harvester. Track damage is affected by soil type, humidity, penetration resistance, etc. The penetration resistance of soil is an important factor in forest harvesting operations which must be accounted for in order to prevent additional costs for harvesting. Peat areas for example cause challenging soil conditions for heavy machinery and extra carefulness is needed there. It is both an expensive and laborious operation to get sunken forest harvesters out from peats. Therefore it is important to select harvesting routes which have the highest possible penetration resistance. As in cases 1

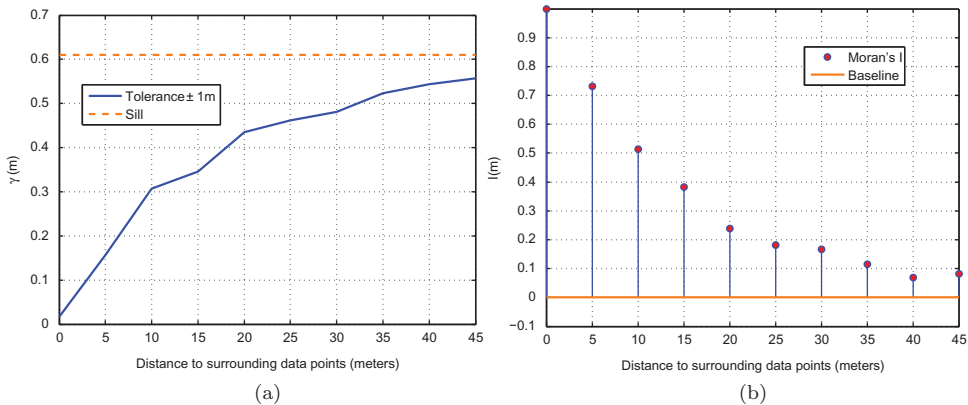


Figure 11. The semivariogram and Moran's I plot depicting the SAC of the response value of case 3. (a) Semivariogram with $t = 1$ m. (b) Moran's I.

and 2, the semivariogram and Moran's I plot for the response variable of case 3 are presented in Figure 11, which also show a clear presence of SAC. Note that the track damage is an ordinal variable consisting from three classes and hence it was also possible to construct a variogram in this case.

The research area consists from 13 different harvesting areas in Pieksämäki, a municipality located in the province of Eastern Finland $62^{\circ} 18' N$ $27^{\circ} 08' E$. A total of 83 feature variables were used for classifying the soil damage. The sizes of the datasets collected from each of these areas ranged from hundreds of samples to thousands of samples. The total number of data points was 11,795. As in cases 1 and 2, the used datasets in case 3 are shown in Table 1. In case 3, the predicted response value of \hat{y}_i (track damage class) is defined as the mode of set N_i (kNN classification), where N_i is again the set of kNNs of data point \mathbf{d}_i .

The SLOO and SLOO-RLO analyses were conducted on each of the 13 harvest areas separately because the distances between the harvest areas were in worst cases dozens of kilometers. On each of these areas, the SLOO and SLOO-RLO procedures were implemented and the results were averaged over all areas. Figure 12 presents the SLOO and SLOO-RLO results for case 3. Similarly as in cases 1 and 2, the results in case 3 confirm the effect of SAC on prediction performance estimates. One can notice an exponential form decay in the SLOO results as a function of dead zone radius r_{δ} whereas the SLOO-RLO results are almost unchanged as it was also in case 2. In the worst case, we have approximately 40% difference in the results between SLOO and SLOO-RLO.

5. Conclusion

Spatio-temporal autocorrelation is always present with GIS-based datasets and needs to be accounted for in machine-learning approaches. As discussed earlier, traditional model performance criteria such as the CV method omit the consideration of the effect of SAC

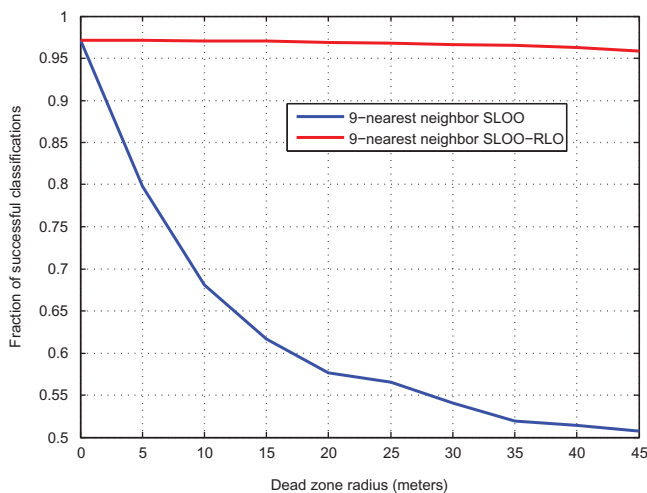


Figure 12. The SLOO and SLOO-RLO results in the Pieksämäki analysis. The y-axis corresponds to the fraction of successful classifications and x-axis to the length of dead zone radius r_{δ} .

in the performance estimations with natural datasets. To account for the SAC in GIS-based datasets, we demonstrated by the means of three experiments that the SKCV method can be used for estimating the prediction performance of spatial models without the optimistic bias due to SAC, while the ordinary CV can cause highly optimistically biased prediction performance estimates. We also showed that SKCV can be used as a data sampling density selection criterion for new research areas, which will result in reduced costs for data collection.

Acknowledgments

We want to thank the Natural Resources Institute Finland (LUKE), Geological Survey of Finland (GTK), Natural Land Survey of Finland (NLS) and Finnish Meteorological Institute (FMI) for providing the datasets. This work was supported by the funding from the Academy of Finland (Grant 295336). The preprocessing of the data was partially funded by the Finnish Funding Agency for Innovation (Tekes).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the funding from the Academy of Finland [Grant 295336]. The preprocessing of the data was partially funded by the Finnish Funding Agency for Innovation (Tekes).

ORCID

Jonne Pohjankukka  <http://orcid.org/0000-0002-5808-2577>

Tapio Pahikkala  <http://orcid.org/0000-0003-4183-2455>

References

- Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. Springer series in statistics. In: E. Parzen, K. Tanabe, and G. Kitagawa, eds. *Selected papers of Hirotugu Akaike*. New York: Springer New York, 199–213.
- Ala-Ilomäki, J., et al., 2015. *New computational methods for efficient utilisation of public data*. Technical report, Espoo, Finland: Geological Survey of Finland.
- Araghinejad, S., 2014. Regression-based models. In: A. Shahab, ed. *Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering*. Dordrecht, Netherlands: Springer, 49–83.
- ASTM D2216-10, 2010. Standard test methods for laboratory determination of water (moisture) content of soil and rock by mass. In: *Annual Book of ASTM Standards, Vol. 04.08*. West Conshohocken, PA: ASTM International.
- Azzalini, A. and Diggle, P., 1994. Prediction of soil respiration rates from temperature, moisture and soil type. *Journal of the Royal Statistical Society - Series C: Applied Statistics*, 43 (3), 505–526.
- Bahn, V., O'Connor, R.J., and Krohn, W.B., 2006. Importance of spatial autocorrelation in modeling bird distributions at a continental scale. *Ecography*, 29 (6), 835–844. doi:10.1111/j.2006.0906-7590.04621.x

- Beale, C.M., et al. 2010. Regression analysis of spatial data. *Ecology Letters*, 13 (2), 246–264. doi:10.1111/j.1461-0248.2009.01422.x
- Betts, M.G., et al. 2009. Comment on 'Methods to account for spatial autocorrelation in the analysis of species distributional data: a review'. *Ecography*, 32 (3), 374–378. doi:10.1111/j.1600-0587.2008.05562.x
- Beven, K. and Kirkby, M., 1979. A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24 (1), 43–69. doi:10.1080/02626667909491834
- Brenning, A., 2005. Spatial prediction models for landslide hazards: review, comparison and evaluation. *Natural Hazards and Earth System Sciences*, 5 (6), 853–862. doi:10.5194/nhess-5-853-2005
- Brenning, A., 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: the R package sperrorest. In: L. Bruzzone, et al., eds. *Proceedings of the Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. Munich, Germany: IEEE, 5372–5375. July.
- Cressie, N.A.C., 2015a. In: *Geostatistics*. New Jersey, USA: John Wiley & Sons, Inc, 101–104.
- Cressie, N.A.C., 2015b. In: *Geostatistics*. New Jersey, USA: John Wiley & Sons, Inc, 58–60.
- Diniz-Filho, J.A.F., Bini, L.M., and Hawkins, B.A., 2003. Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology and Biogeography*, 12 (1), 53–64. doi:10.1046/j.1466-822X.2003.00322.x
- Donkoh, E.K. and Opoku, A.A., 2016. Optimal geometric disks covering using tessellable regular polygons. *Journal of Mathematics Research*, 8 (2), 25. doi:10.5539/jmr.v8n2p25
- Dormann, F., et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30 (5), 609–628. doi:10.1111/j.2007.0906-7590.05171.x
- Feichtinger, H. and Strohmer, T., 1997. *Gabor analysis and algorithms: theory and applications*. Basel, Switzerland: Birkhäuser.
- Geary, R.C., 1954. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5 (3), 115–146. doi:10.2307/2986645
- Getis, A. and Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24 (3), 189–206. doi:10.1111/j.1538-4632.1992.tb00261.x
- Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology*, 93 (3), 679–688. doi:10.1890/11-0826.1
- Hoeting, J.A., et al. 2006. Model selection for geostatistical models. *Ecological Applications*, 16 (1), 87–98. doi:10.1890/04-0576
- Hyyönen, E., et al. 2003. Assessing site suitability for Scots pine using airborne and terrestrial gamma-ray measurements in Finnish Lapland. *Canadian Journal of Forest Research*, 33 (5), 796–806. doi:10.1139/x03-005
- Koenig, W.D., 1999. Spatial autocorrelation of ecological phenomena. *Trends in Ecology & Evolution*, 14 (1), 22–26. doi:10.1016/S0169-5347(98)01533-X
- Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology*, 74 (6), 1659–1673. doi:10.2307/1939924
- Lichstein, J.W., et al. 2002. Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, 72 (3), 445–463. doi:10.1890/0012-9615(2002)072[0445:SAAMI]2.0.CO;2
- Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1–2), 17–23. doi:10.1093/biomet/37.1-2.17
- Pennanen, O. and Mäkelä, O., 2003. Raakapuukuljetusten kelirikkohaittojen vähentäminen, Metsätehon raportti. In: *Technical report 153*. Kuopio, Finland: Metsäteho Ltd.
- Pietikäinen, M., et al., 2011. *Computer vision using local binary patterns*. London, UK: Computational Imaging and Vision Springer.
- Pohjankukka, J., et al., 2014a. Predicting water permeability of the soil based on open data. In: L. Iliadis, I. Maglogiannis, and H. Papadopoulos, eds. *Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014)*, Vol. 436 of IFIP Advances in Information and Communication Technology. Heidelberg, Germany: Springer, 436–446. January.
- Pohjankukka, J., et al., 2014b. Arctic soil hydraulic conductivity and soil type recognition based on aerial gamma-ray spectroscopy and topographical data. In: M. Borga, et al., eds.

- Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014)*. Stockholm, Sweden: IEEE, 1822–1827. August.
- Pohjankukka, J., et al. 2016. Predictability of boreal forest soil bearing capacity by machine learning. *Journal of Terramechanics*, 68, 1–8. doi:[10.1016/j.jterra.2016.09.001](https://doi.org/10.1016/j.jterra.2016.09.001)
- Rest, K.L., et al. 2014. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, 23 (7), 811–820. doi:[10.1111/geb.12161](https://doi.org/10.1111/geb.12161)
- Ruß, G. and Kruse, R., 2010. Regression models for spatial data: an example from precision agriculture. In: *Proceedings of the 10th Industrial Conference on Advances in Data Mining: applications and theoretical aspects, ICDM'10, Berlin, Germany*. Berlin, Heidelberg: Springer-Verlag, 450–463.
- Schulte, R., et al., 2005. Predicting the soil moisture conditions of Irish grasslands. *Irish Journal of Agricultural and Food Research*, 44, 95–110.
- Seibert, J. and McGlynn, B., 2007. A new triangular multiple flow direction algorithm for computing upslope areas from gridded digital elevation models. *Water Resources Research*, 43 (4), W04501. doi:[10.1029/2006WR005128](https://doi.org/10.1029/2006WR005128)
- Sirén, M., et al. 2013. Harvesting damage caused by thinning of Norway spruce in unfrozen soil. *International Journal of Forest Engineering*, 24 (1), 60–75. doi:[10.1080/19132220.2013.792155](https://doi.org/10.1080/19132220.2013.792155)
- Tamminen, P., 1991. *Kangasmaan ravinnetunnusten ilmaiseminen ja viljavuuden alueellinen vaihtelu etelä-suomessa: expression of soil nutrient status and regional variation in soil fertility of forested sites in southern Finland*. Helsinki, Finland: Folia Forestalia. Vol. 1.
- Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240. doi:[10.2307/143141](https://doi.org/10.2307/143141)
- Tomppo, E., et al., 2008. Multi-source national forest inventory - methods and applications. In: E. Tomppo, M. Haakana, M. Katila and J. Peräsaari, eds. *Managing forest ecosystems*. Dordrecht, Netherlands: Springer, Vol. 18.
- Wood, J., 1996. *The geomorphological characterisation of digital elevation models*. Thesis (PhD). University of Leicester.
- Wood, J., 2009. Geomorphometry in LandSerf. In: T. Hengl and H. Reuter, eds. *Developments in soil science*. Amsterdam, Netherlands: Elsevier, Vols. 33,12, 333–349.
- Zevenbergen, L. and Thorne, C., 1987. Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12 (1), 47–56. doi:[10.1002/\(ISSN\)1096-9837](https://doi.org/10.1002/(ISSN)1096-9837)
- Zhang, H. and Wang, Y., 2010. Kriging and cross-validation for massive spatial data. *Environmetrics*, 21 (3–4), 290–304. doi:[10.1002/env.1023](https://doi.org/10.1002/env.1023)

Publication V

Comparison of estimators and feature selection procedures in area-based forest inventory based on airborne laser scanning and digital aerial imagery

Jonne Pohjankukka, Sakari Tuominen, Juho Pitkänen, Tapio Pahikkala and Jukka Heikkonen. *Scandinavian Journal of Forest Research*, Accepted. Taylor & Francis, 2018.

Copyright © 2018 Taylor & Francis. Reprinted with permissions from respective publisher and authors.

Comparison of estimators and feature selection procedures in forest inventory based on airborne laser scanning and digital aerial imagery

J. Pohjankukka^a, S. Tuominen^b, J. Pitkänen^b, T. Pahikkala^a and J. Heikkonen^a

^aDepartment of Future Technologies, University of Turku, Vesilinnantie 5, FI-20500 Turku, Finland; ^bNatural Resources Institute Finland (Luke), Latokartanonkaari 9, FI-00790 Helsinki, Finland

ARTICLE HISTORY

Compiled May 8, 2018

ABSTRACT

Digital maps of forest resources are a crucial factor in successful forestry applications. Since manual measurement of this data on large areas is infeasible, maps must be constructed using a sample field data set and a prediction model constructed from remote sensing materials, of which airborne laser scanning (ALS) data and aerial images are currently widely used in management planning inventories. ALS data is suitable for the prediction of variables related to the size and volume of trees, whereas optical imagery helps in improving distinction between tree species. We studied the prediction of forest attributes using field data from National Forest Inventory complemented with ad hoc field plots in combination with ALS and aerial imagery data in Åland province, Finland. We applied feature selection with genetic algorithm and greedy forward selection and compared multiple linear and nonlinear estimators. Maximally around 40 features from a total of 154 were required to achieve the best prediction performances. Tree height was predicted with normalized root mean squared error value of 0.1 and tree volume with a value around 0.25. Predicting the volumes of spruce and broadleaved trees was the most challenging due to small proportions of these tree species in the study area.

KEYWORDS

Machine learning; feature selection; forestry; remote sensing data

Introduction

Accurate and geographically explicit information about forest characteristics is required for efficient management of forest resources. Remote sensing (RS) and earth observation techniques provide means for producing estimates of forest parameters in the form of thematic maps. In Finland, currently the most significant RS materials for forest inventory are airborne laser scanning (ALS) data and digital aerial photography. In recent years, a forest inventory method based on these data sources has been adopted for stand and sub-stand level forest mapping and the estimation of forest attributes. Normally, low-density ALS data (typically 0.5 - 2 return/m²) and digital aerial imagery with a spatial resolution of 0.25 - 0.5 m and covering visible and near-infrared wavelengths are used. ALS is currently considered the most accurate RS material for estimating stand-level forest variables (e.g. Næsset 2002, 2004;

Corresponding author J. Pohjankukka. Email: jjepoh@utu.fi

Maltamo et al. 2006). Compared to optical RS data sources, ALS data are particularly well suited to the estimation of forest attributes related to the physical dimensions of trees, such as stand height and volume. By means of ALS data, a three-dimensional (3D) surface model of the forest can be derived. Since it is possible to distinguish laser returns reflected from the ground surface from those reflected from tree canopies, both digital terrain models (DTM) and digital surface models (DSM) can be derived from ALS data (e.g. Axelsson 1999; Baltsavias 1999; Hyyppä et al. 2000; Pyysalo 2000; Gruen and Zhang 2003). On the other hand, optical imagery is typically acquired to complement the ALS data, because ALS is not considered to be well suited for estimating tree species composition or dominance, with the return densities applied for operational forest inventory (e.g. Törmä 2000; Waser et al. 2011). Of the various optical RS data sources, aerial images are usually the most readily available and best-suited for forest inventory purposes.

When using very high resolution RS data in forest inventory, the main alternatives are detection of individual trees (e.g. Koch et al. 2006) or area-based inventory method (e.g. Næsset 1997; Lefsky et al. 1999), where inventory unit is e.g. a sample plot. The primary inventory unit applied in operational ALS and aerial image based forest inventory in Finland is a square area with size of 16×16 meters. These areas form a uniform grid covering the entire inventory area. Field plots are used as a reference data for ALS and image data interpretation. Statistically, the method resembles two-phase stratified sampling although the sampling layout is not aimed at producing unbiased regional estimates, but instead, the method rather aims at locally (at plot and stand level) precise forest estimates. Thus, the method requires high correlation between the RS features and actual inventory variables (i.e. field data). Typically non-parametric estimators such as k -nearest neighbors or most similar neighbors are applied in combining RS data and field measurements.

By combining aerial photographs and ALS data it is possible to derive a very large number of RS features describing the characteristics of a field plot or a stand. The extracted RS features form a n -dimensional feature space, where n equals the number of applied RS features. It is, in general, computationally infeasible to use all possible RS features when processing large inventory areas. Also, with increasing dimensionality, the data typically become sparse in relation to the dimensions (Hinneburg et al. 2000). This causes problems, especially when using estimators based on distance or proximity in the feature space ('the curse of dimensionality'; e.g. Beyer et al. 1999). Therefore, the number of RS features needs to be reduced in a way that produces an appropriate subset of features for the estimation procedure, considering their usefulness in predicting the forest attributes as well as their mutual correlation. Various approaches have been applied for this purpose (e.g. Siedlecki and Sklansky 1989; Pudil et al. 1994; Jain and Zongker 1997; Kudo and Sklansky 1998, 2000). In RS-aided forest inventory applications e.g. correlation analysis (Tuominen and Pekkarinen 2005; Breidenbach et al. 2010), canonical analysis (Packalén et al. 2012), stepwise selection using various criteria and proceeding either forwards by adding or backwards by eliminating features, or combining these operations (Tuominen and Pekkarinen 2005; Maltamo et al. 2006; Packalén and Maltamo 2007; Haapanen and Tuominen 2008; Hudak et al. 2008; Packalén et al. 2009; Latifi et al. 2010; Breidenbach et al. 2010; Packalén et al. 2012), simulated annealing (Packalén et al. 2012) and genetic algorithms (e.g. Van Coillie et al. 2005; Haapanen and Tuominen 2008; Latifi et al. 2010) have been used.

The objective of this paper is to examine the performance of different feature selectors and estimators in predicting sample plot level forest parameters when using a combination of features extracted from ALS data and aerial imagery. We implement

feature selection with genetic algorithm and greedy forward selection and compare multiple linear and nonlinear estimators for finding near-optimal ALS and aerial imagery features.

Materials and methods

In this section, we will first describe the used data sets which are followed by a description of the estimators, feature selectors and analysis details. We denote the i th data point by a vector $\mathbf{d}_i = (\mathbf{x}_i, y_i) \in \mathbb{R}^{n+1}$, where $\mathbf{x}_i \in \mathbb{R}^n$ is the vector of $n \in \mathbb{N}$ predictor features and $y_i \in \mathbb{R}$ the corresponding response value. The measures used in this article for evaluating the goodness of the estimators are the root mean squared error (RMSE) and normalized RMSE (NRMSE). These two measures are defined as:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2}, \quad \text{NRMSE} = \text{RMSE}/\bar{y}, \quad (1)$$

where $m \in \mathbb{N}^+$ denotes the number of data points, \hat{y}_i is the model estimated value for the response value y_i and \bar{y} is the average value of the known responses, i.e. $\bar{y} = m^{-1} \sum_{i=1}^m y_i$. Other used statistics include the relative bias (BIAS%) and the well-known coefficient of determination (R^2). The relative bias is defined as:

$$\text{BIAS}\% = 100 \times \frac{1}{m} \sum_{i=1}^m \frac{y_i - \hat{y}_i}{\bar{y}}. \quad (2)$$

Study area

The study area was defined by the ALS coverage acquired by National Land Survey of Finland (NLS) in the spring of 2013, and it covered the main parts of the province of Åland but excluded the easternmost municipalities and islands in the outer archipelago (see Figure 1). The total area covered by both ALS data and aerial imagery was approximately 346,000 ha, but a large part of it was sea area. The land use in the study area, excluding sea water, is presented in Table 1. These figures are from a land use map that was created using field plot data of 10th National Forest Inventory (NFI10), satellite imagery and NLS map data (Tomppo et al. 2008, 2013). Approximately 51.2%

Table 1. Land use in the study area (excluding sea water).

	Area, ha	%
Forest land	70076	54.9
Other wooded land	20420	16.0
Other land	11259	8.8
Forestry land in total	101755	79.8
Agriculture	14270	11.2
Roads and built-up areas	7795	6.2
Inland waters	3568	2.8
Total	127589	100

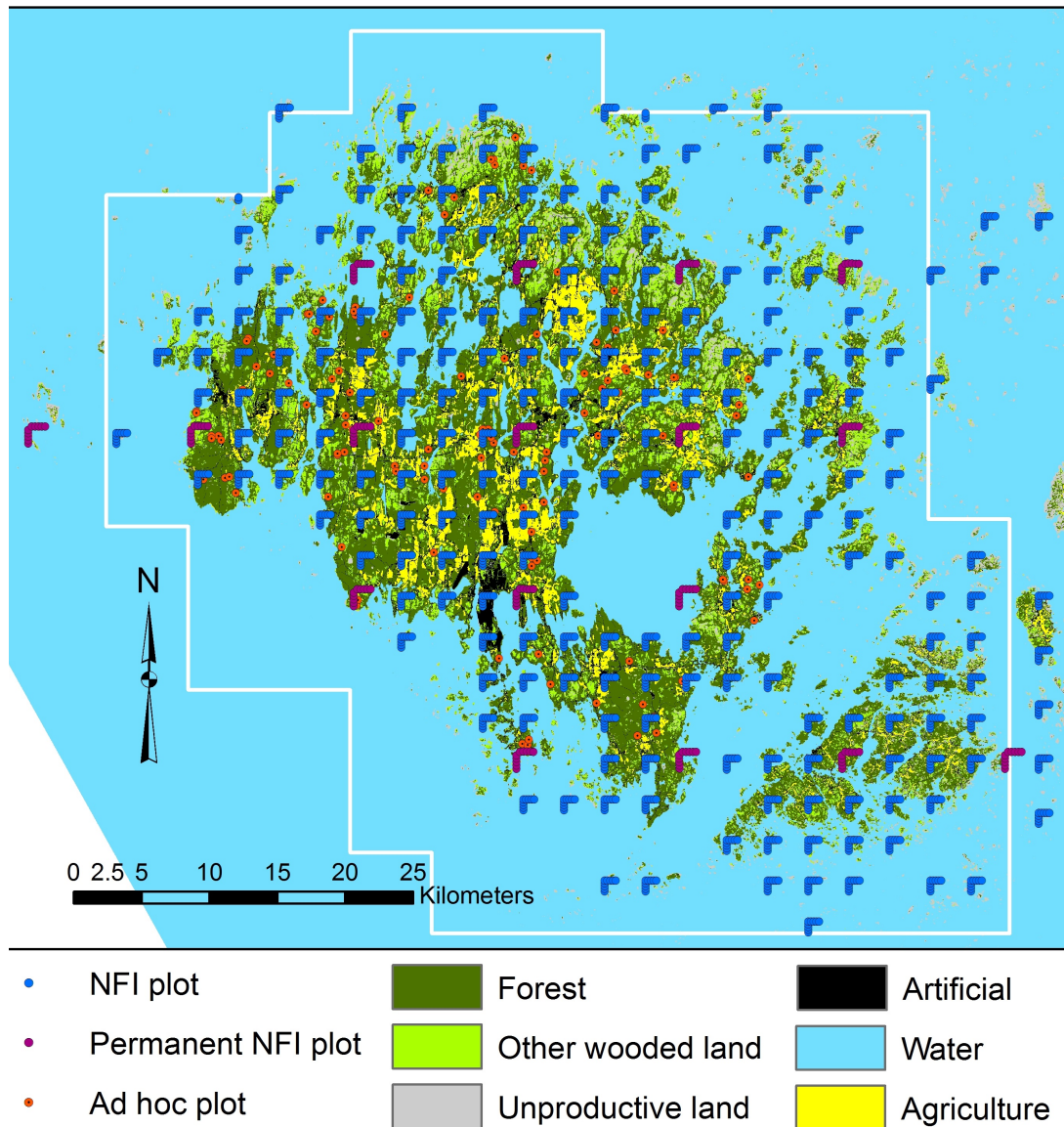


Figure 1. Map of the sampling layout in Åland. Background land cover map is based on NFI11 and topographic data. The ALS coverage area is marked with white borderline.

of the forestry land area was dominated by pine, 0.5% by spruce, 20.5% mixed coniferous (pine-spruce) and 11.2% by broadleaved species. The remaining part of forestry land area was either mixed with no dominant species (8.3%) or open regeneration areas (1.6%), young seedling stands and very sparsely stocked areas (6.7%), where no dominance was determined.

Field data

Sample plots of 11th National Forest Inventory (NFI11) were used as field reference data. The design of the NFI11 in Åland was a stratified sampling with two sampling regions: a) the ALS study area, and b) other parts of Åland, mainly remote or sparsely wooded islands (see Figure 1). In both regions systematic cluster sampling was used.

In the study area temporary sample plot clusters were established in a grid of 3 km by 3 km, whereas in the other parts the grid was 4 km by 4 km. A cluster consisted of 9 sample plots in inverted L-shaped form, having 200 meters between plots. In addition to these temporary sample plot clusters, the permanent sample plot clusters established in NFI9 (1999) were remeasured. The grids of the two cluster types were overlapping. Thus, each permanent cluster replaced one temporary cluster - in the grid of 12 by 12 km there were always 1 permanent cluster and 15 temporary clusters. The sample plots were measured in 2013.

The sample plot was a restricted Bitterlich relascope plot with a basal area factor 1 and maximum radius of 9 meters. For each sample plot, data were recorded at stand and tree levels. The stand level data included more than 100 variables describing the administrative status, site, damages, accomplished and recommended silvicultural measures. The sample plots having their center point at least 9 m from the stand boundary and located on forest land or other wooded land were positioned with Trimble Pro 6H GPS device. Using post-processed Global Navigation Satellite System (GNSS) observations the positions of plots were determined with approximately 1 m accuracy. Only the data of these plots were used in this study.

The tree level data included diameter at breast height (*dbh*), species, tree quality class, and crown class. Every 7th measured tree with *dbh* larger than 195 mm and every 14th measured tree with *dbh* less than or equal to 195 mm were selected as a sample tree with more detailed measurements. The sample tree variables included tree height, height increment of past five years (for conifers) and description of stem quality and possible damages. In addition, a bore core was taken for age and diameter increment assessment. The tree level data were used to calculate plot level stem volume, basal area, mean diameter, and mean height estimates. For volume estimation, tree heights were estimated using the mixed-effects models of (Eerikäinen 2009), including calibration at the plot and cluster levels using the sample tree heights and additionally mean tree heights assessed at the stand level. Then tree volumes were estimated using the volume functions of (Laasasenaho 1982) and expanded to per hectare values using tree specific expansion factors.

For RS-based forest inventory it is necessary to have field observations of all types of forest, otherwise the forest strata that are missing from field observations will be missing also from inventory results. In addition to the systematic NFI sample, the field data was complemented by ad hoc reference sample plots, because the systematic sample did not provide sufficient field observations in forest strata whose area was small. For the selection of ad hoc sample, an initial grid of plots was generated to inventory area with a spacing of 100×100 m between the initial plots (these points included the center point locations of the NFI plots). The initial plots in forestry land were stratified on the basis of ALS and aerial image features. The following features were used in the stratification: height where 85% of LiDAR returns have accumulated (4 strata), ratio of canopy returns to all returns (3 strata), inverse distance moment of rasterized canopy height model (4 strata) and spectral average of aerial image near-infrared (nir) band (4 strata). These features should correlate, respectively, with stand height, stand density, size/spatial organization of tree crowns and proportion of broadleaved trees (see the next two sections for more detailed descriptions of ALS and aerial image features). Ad hoc plots were not allocated into strata, whose area was less than 100 ha in the entire inventory area. Thus, although the theoretical number of strata was 192, some feature combinations are unlikely to occur, and the number of those strata representing an area more than 100 ha was 101 in the inventory area. On the basis of the distribution of the existing NFI sample plots within the strata, there

Table 2. Basic statistics of target inventory variables: tree diameter, all trees (d); tree height, all trees (h); tree basal area, all trees (b); tree volume, all trees (v_a); tree volume, pine (v_p); tree volume, spruce (v_s); tree volume, broadleaf (v_b). The units of the target variables are given in parentheses in the top row. All v -variables have the same unit.

Statistic	d (cm)	h (m)	b (m ² /ha)	v_a (m ³ /ha)	v_p	v_s	v_b
mean μ	21.1	13.6	19.5	147.6	83.1	31.3	33.2
st. deviation σ	9.52	5.62	12.87	125.85	88.32	67.89	66.55
coeff. of variation σ/μ	0.45	0.66	0.41	0.85	1.06	2.17	2.00
maximum	48.6	28.0	69.6	843.5	485.8	590.0	492.5

were 70 strata that required additional field plots. Altogether 126 ad hoc plots were selected from initial grid of plots belonging to those strata that were underrepresented in proportion to their area or missing among the systematic NFI sample. The field measurement of complementary field plots was carried out as with NFI plots. On the whole, the field data in the study region consisted of 475 sample plots on forestry land, from which 432 plots were on productive forest land, 41 on poorly productive forestry land and 2 on unproductive land.

A summary of the main statistics of the target NFI variables is presented in Table 2. The studied variables were as follows: tree diameter, all trees (d); tree height, all trees (h); tree basal area, all trees (b); tree volume, all trees (v_a); tree volume, pine (v_p); tree volume, spruce (v_s); tree volume, broadleaf (v_b). Diameter and height were calculated as basal area weighted averages. Volume of pine included also other conifer species except spruce.

ALS data

ALS data were acquired between 24th and 26th April 2013 by the company Fugro Malta Ltd. using a Piper Chieftain aircraft and a Riegl LMS-Q780 laser scanner. It is worth noticing that ALS data were acquired during leafless season for the purpose of producing a DTM of the area. The scanning altitude was 2000 m above sea level, with a maximum zenith angle of 20° and side overlap of 20%. Average density of returns was 0.54 per square meter. The returns were automatically classified into ground and vegetation returns. The automatic classification is based on the order of the LiDAR returns, which were categorized as 'only', 'first of many', 'last of many' and 'intermediate' returns. For ground classification local minima of the last returns are used as a basis of the ground level (Vilhomaa and Laaksonen 2011). The ground elevation for each LiDAR point was estimated via spatial interpolation using two nearest-neighbor ground returns and inverse distance weighting. The height above ground (H) was calculated for each LiDAR point as the difference between the z -coordinate and the estimated ground level.

ALS points within 9 m radius from the center points of the field plots were used in calculation of features from H and also from intensity (I) recorded for the points. A minimum H of two meters was used to classify LiDAR returns as canopy hits, from which most of the features were calculated. The following features were extracted from the height and intensity of the ALS points (for full list of the features see Table A and B in the appendix):

- (1) Average, standard deviation and coefficient of variation of H for canopy returns, separately from first and last returns (havg[f/l], hstd[f/l], hcv[f/l]). Only returns were included in both first and last returns

- (2) H at which $p\%$ of cumulative sum of H of canopy returns is achieved (H_p) ($hp[f/l]$, p is one of 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 85, 90, 95 and 100)
- (3) Percentage of canopy returns having $H \geq$ than corresponding H_p ($pp[f/l]$, p is one of 20, 40, 60, 80, 95)
- (4) (a) Ratio of first canopy returns to all first returns ($vegf$), and (b) Ratio of last canopy returns to all last returns ($vegl$)
- (5) Percentage of canopy returns above height limits calculated in each field plot from $H_{\min} + \frac{s}{10} * (H_{\max} - H_{\min})$ ($ds[f/l]$, s is one of 1 - 9)
- (6) Ratio of intensity percentile p to the median of intensity for canopy returns ($ip[f/l]$, p is one of 20, 40, 60 and 80)

For getting texture features, H and I values of first returns were also interpolated to raster images with 0.5 m resolution. Pixel values of the raster images were calculated using inverse squared distance weighting of the two nearest ALS points. Inverse squared distance weighting was chosen as interpolation method based on the relation between the LiDAR return density and the shape of tree canopies. Both the height and intensity images were further quantized into 16 classes for extraction of Haralicks texture features (Haralick et al. 1973). These were calculated from windows of 32 by 32 pixels around the center points of the field plots, with the co-occurrence comparison offset of five pixels. The following texture features were first calculated in four directions and then averaged out: angular second moment (ASM), contrast (Contr), correlation (Corr), variance (Var), inverse difference moment (IDM), sum average (SA), sum variance (SV), sum entropy (SE), entropy (Entr), difference variance (DV) and difference entropy (DE).

Aerial imagery data

The aerial imagery from the study area was acquired in June 2013. Vexcel Ultracam Eagle M1 f80 camera sensor was used in image acquisition, and the imaging altitude was 7670 m (corresponding approximately ground resolution with this sensor). Stereo overlap was 60% within flight line and 30% between flight lines. Thus, contrary to the ALS data, the aerial imagery was from the season when the vegetation was in full leaf. Both RGB and colour-infrared (CIR) images were acquired. All bands of CIR imagery green (g), red (r) and nir, and blue (b) band of RGB imagery were used in this study. The aerial images were acquired using digital camera sensor. The images were orthorectified into resolution of 0.5 m. Image features for field plots were calculated again within local windows of 32 by 32 pixels. The following features were extracted from the aerial image bands (for full list of the features see Table A and B in the appendix):

- (1) Average (mean), standard deviation (std) and coefficient of variation (cv) from each of the four bands
- (2) The next transformations from band averages within the pixel windows:
 - (a) NDVI as $\frac{nir - r}{nir + r}$ (ndvi)
 - (b) A modified NDVI as $\frac{nir - g}{nir + g}$ (gndvi)
 - (c) $\frac{nir}{r}$ (nir.r)
 - (d) $\frac{nir}{g}$ (nir.g)
- (3) The same Haralick features as from the ALS based canopy height and intensity

images with similar processing and parameters

K-nearest neighbors

K-nearest neighbors (KNN, see e.g. Araghinejad 2014, p. 66–73) is a simple, yet effective nonlinear estimator used for classification and regression. Given a data point \mathbf{d}_i , a prediction \hat{y}_i for its response value in KNN-regression is given by:

$$\hat{y}_i = \frac{\sum_{j \in N_i} w_j y_j}{\sum_{j \in N_i} w_j}, \quad (3)$$

where w_j s are the weight values for the neighbors y_j and N_i is the set of indexes of the k -nearest neighbors ($|N_i| = k \in \mathbb{N}$). The k neighbors of \mathbf{d}_i are determined by the feature vectors \mathbf{x}_i using some metric function, typically the Euclidean distance function e . In case of KNN-classification the predicted response value \hat{y}_i is defined, e.g., as the mode of the k -nearest neighbors. In many applications of KNN-regression, the weights w_j in Equation 3 are usually set as $w_j = 1$. In a distance weighted KNN the weights are set as $w_j = \frac{1}{e(\mathbf{x}_j, \mathbf{x}_i)}$, i.e. neighbors \mathbf{d}_j which are closer to \mathbf{d}_i in the feature space are given higher weight. In forest applications the distance weighted KNN is also commonly used. KNN has an upside that several response variables can be predicted simultaneously, and then the predicted quantity variables of subgroups, e.g. volumes of tree species, sum up to the predicted total value. Also, covariance of variables in training observations is retained in predictions, albeit fully only if k is one.

Regularized least squares

Regularized least squares (known as RLS or ridge regression, Hoerl and Kennard 2000) is a regularized version of the linear regression method. Linear regression models are one of the most widely used methods in statistical inference. In RLS our aim is to find a weight vector $\mathbf{w} \in \mathbb{R}^n$, such that the cost function:

$$E(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m (y_i - \mathbf{x}_i^T \mathbf{w})^2 + \frac{\lambda}{m} \mathbf{w}^T \mathbf{w} \quad (4)$$

is minimized. The term $\lambda > 0$ is the regularization parameter, which will increase the value of the cost function for high magnitude vectors \mathbf{w} . This will balance the trade-off between minimizing the training error and model complexity, which will result in better generalization. The optimal λ parameter is selected from a predefined set using e.g. cross validation (CV). In our analyses we selected optimal λ from the set $\Lambda = \{2^{-16}, 2^{-15}, \dots, 2^{15}, 2^{16}\}$ by using leave-one-out CV (LOOCV). The selection of the set Λ is a practical rule-of-thumb (see e.g. Hsu et al. 2010) since it covers a good representative grid of the regularization parameter values from very small to very large. Exponential relation between the regularization parameters also assures the values are not too close to each other.

In some of the target inventory variables the data distributions were highly skewed. For example, with the tree volume targets the data mainly consisted of a large number of low volume values and relatively small amount of high volume values. This naturally causes the RLS model to favor data points with low volume values over high volume values. Imbalanced data sets are rather common in many applications of data analysis and often used countermeasure is to use under- or oversampling techniques (see e.g.

Chawla 2005). Since large volume values are of more practical interest in our case we introduce in this article an additional estimator which we call balanced RLS (BRLS). BRLS is almost identical to RLS, except that first it implements oversampling with replacement to the data, so that the skewed target distribution is balanced (i.e. a uniform distribution). After the balancing we continue with building a RLS model using this extended data set. In the tree volume example, this is equivalent to giving more weight for learning to predict data points with large volume value than data points with low volume value. When selecting the BRLS model via CV, we made sure that no repetitions of the data points were included into both training and test data sets in a single CV round. In other words, the intersection of training and test data sets was always an empty set. Otherwise, the results would trivially be very good for BRLS.

Multilayer perceptron

Multilayer perceptron (MLP, see e.g. Bishop 1995) is a feedforward neural network, where the goal is to train the network by minimizing the cost function:

$$E(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^m (y_i - a(\mathbf{x}_i, \mathbf{w}))^2, \quad (5)$$

where the function $a(\mathbf{x}_i, \mathbf{w})$ is the output of the network given an input \mathbf{x}_i and network weights \mathbf{w} . The network weights are defined as: $\mathbf{w} = \{w_{jk}^{(l)} \mid 2 \leq l \leq L, 1 \leq j \leq d^{l-1}, 1 \leq k \leq d^l\}$, where $L \in \mathbb{N}$ is the total number of layers in the network (including input and output layer) and $d^l \in \mathbb{N}$ is the number of nodes on layer l . For example $w_{13}^{(2)}$ is the network weight between nodes 1 and 3 of layers 1 and 2. Layer 1 in the network corresponds to the input layer.

A popular ad hoc version of the MLP network is the MLP early stopping committee (MLP-ESC) which consists from a committee of MLP networks. The prediction for a response value y_i is defined as the average output of the committee networks. In MLP-ESC, we use early stopping with a validation data set to regularize the MLP training in order to avoid over-fitting. In other words, we stop training the MLP network when the validation error begins to increase. Due to its simplicity MLP-ESC is easy and fast to implement and in many cases gives good results.

In our analyses, the MLP experiment was implemented using Netlab (Nabney 2004) library and the MLP-ESC model consisted from two layers, two hidden units and 10 committee members.

Greedy forward selection

Greedy forward selection (GFS, see e.g. Pahikkala et al. 2010) is a depth-first type feature selection method. The idea in GFS is to sequentially select the best feature to be used along with the already selected features. Specifically, the steps of the GFS are the following:

- (1) Initially a set of selected features F is an empty set $F = \emptyset$ and a set G consists from all the predictor features.
- (2) Form a candidate feature set $C_i = F \cup \{f_i\}$, where $f_i \in G$ and evaluate model performance (via e.g. CV) using features in set C_i . Repeat this process for all possible candidate sets C_i .

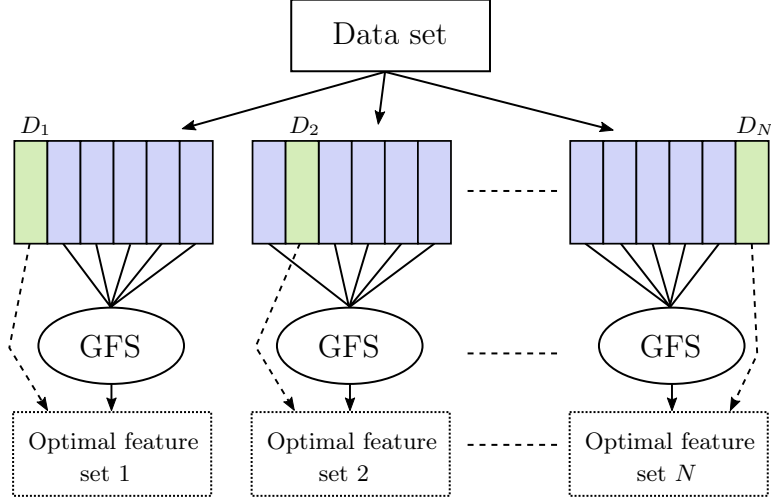


Figure 2. Illustration of the NGFS procedure. The data set is first split into N subsets D_i (green rectangles), after which GFS is implemented for each set D_{-i} . The prediction performance of the resulting N feature sets are then tested (via CV) using the corresponding test sets D_i .

- (3) Save the candidate set C_j with the best model prediction performance and set $F = C_j$.
- (4) Set $G = G \setminus \{f_j\}$, i.e. we remove the best found feature from G . Repeat steps 2-4 until $G = \emptyset$.

The resulting set F contains all the predictor features of G with ordering based on sequential maximum improvement of prediction performance. For example the first element in F is the feature with the best performance if a single feature is used in the model. The second element in F is the feature with the best performance together with the first feature and so on.

In addition to GFS we also used a nested version of the GFS, which we call nested GFS (NGFS). In NGFS the data set is first split into N random subsets $\{D_1, D_2, \dots, D_N\}$, after which we implement GFS feature selection N times, with each set D_i at its turn being a test set for evaluating the prediction performance of the features selected using the set $D_{-i} = \cup_{j \neq i} D_j$. In other words, with NGFS we do GFS feature selection N times with slightly different data sets and at each time we test the prediction capability of the features selected by the GFS. The NGFS procedure is illustrated in Figure 2. In this article we used $N = 10$ folds for the NGFS. With NGFS we can analyze the stability of the feature selection process, i.e. it will indicate how sensitive the feature selection is to changes in the data set.

Genetic algorithm

A genetic algorithm (GA) was also used for feature selection so that the goal of the algorithm was to minimize the sum of the RMSE and one and a half times the bias of the target variable in CV. The general GA procedure begins by generating an initial population of strings (chromosomes or genomes), that consist of random combinations of predictor variables (genes). Each chromosome is considered a binary string having values 1 or 0 indicating that certain variable is either ‘selected’ in the subset or ‘not selected’. The strings evolve over a user-defined number of iterations (generations). This evolution includes the following operations: selecting strings for mating by applying

a user-defined objective criterion (the more copies in the mating pool, the better), allowing the strings in the mating pool to swap parts (cross over), causing random noise (mutations) in the offspring (children), and passing the resulting strings to the next generation. The process is repeated until a predefined criterion is fulfilled or a predetermined number of iterations have been completed (see e.g. Broadhurst et al. 1997; Tuominen and Haapanen 2013; Moser et al. 2017).

The population size for GA was 100 and the number of generations 120. Because GA is a stochastic process, we carried out nine feature selection runs for each target variable both with KNN-regression (later KNN-GA) and RLS (RLS-GA). In addition to this univariate modeling, all the target variables were modeled together using multivariate KNN (MVKNN-GA). For this, we selected weights subjectively for target variables that were normalized before the use in the evaluation function of GA. The weights were 1 for diameter, basal area and volume of pine, 1.2 for volume of spruce, 1.5 for height and volume of broad-leaved trees and 4 for total volume. Due to increased complexity from multivariate optimization, we carried out 18 feature selection runs with MVKNN-GA. The number of nearest neighbors, k , was set to 5 in both KNN-GA and MVKNN-GA and inverse squared distance weighting of the KNN was used. In feature selection with RLS-GA, 10-fold cross validation (10-fold-CV) was applied for speeding computations. Otherwise, LOOCV was used both in feature selection and in calculation of result statistics.

Analysis implementation details

Multiple feature selection scenarios using different estimators were implemented in our analyses. For some estimators we used both GFS and NGFS feature selectors. LOOCV based model selection was used in general for maximum utilization of the data in model training. When it was computationally infeasible to implement LOOCV based model selection, we used 10-fold-CV instead. This does not however diminish these analyses since the overall difference in practice between the results of 10-fold-CV and LOOCV is small. In Table 3, we have listed all the conducted analyses in detail with their abbreviations. In KNN analyses using GFS we set the neighbor weights as $w_j = 1$ and with GA we set $w_j = \frac{1}{e(\mathbf{x}_j, \mathbf{x}_i)}$ correspondingly.

In summary of the analyses, the goal was to find the best combination of ALS and digital aerial imagery data features for predicting the target inventory variables. The predictor features consisted from a total of 154 variables (see sections about ALS and aerial imagery data or Tables A and B in the appendix) and the predicted target inventory variables are listed in Table 2. The results with estimators RLS or MLP-ESC

Table 3. List of all implemented analyses and corresponding abbreviations.

estimator	Feature selector	Cross validation	Data balancing	Analysis abbreviation
RLS	GFS	LOOCV	no	RLS-GFS
RLS	NGFS	LOOCV	no	RLS-NGFS
RLS	GFS	LOOCV	yes	BRLS-GFS
RLS	NGFS	LOOCV	yes	BRLS-NGFS
KNN	GFS	LOOCV	no	KNN-GFS
KNN	NGFS	LOOCV	no	KNN-NGFS
MLP-ESC	GFS	10-fold-CV	no	MLP-ESC-GFS
MVKNN	GA	LOOCV	no	MVKNN-GA
KNN	GA	LOOCV	no	KNN-GA
RLS	GA	10-fold-CV	no	RLS-GA

were calculated so that negative predictions were replaced with zero, because our response variables cannot have negative values. A total of 475 data points were available in these studies, making it worth mentioning that generalization of these results is a challenging task. It is for example widely known that as the data dimensionality increases, the number of data points needed to achieve successful generalization increases exponentially. One must of course note that small data sets are a common problem in many applications.

Results

In Figure 3 we have illustrated the GFS and NGFS results for target inventory variables height h (all trees), volume v_p (pine) and volume v_s (spruce). The NRMSE is shown as a function of number of predictor features. The results of h and v_s target variables represent the best and worst cases in terms of predictability. The rest of the target variables had results somewhere between those of h and v_s . Similar graphs for GA were not possible to produce since GA is not a sequential feature selector. We notice that around a maximum of 40 features are enough to achieve the optimum prediction performance in all cases. There was similar behavior in other target variables. After around 40 features the results stabilize or begin to get worse, possibly due to uninformative features. For tree height, MLP-ESC-GFS gets the lowest NRMSE value of 0.1 with just 9 features. For spruce tree volume, KNN-GFS is the best method and

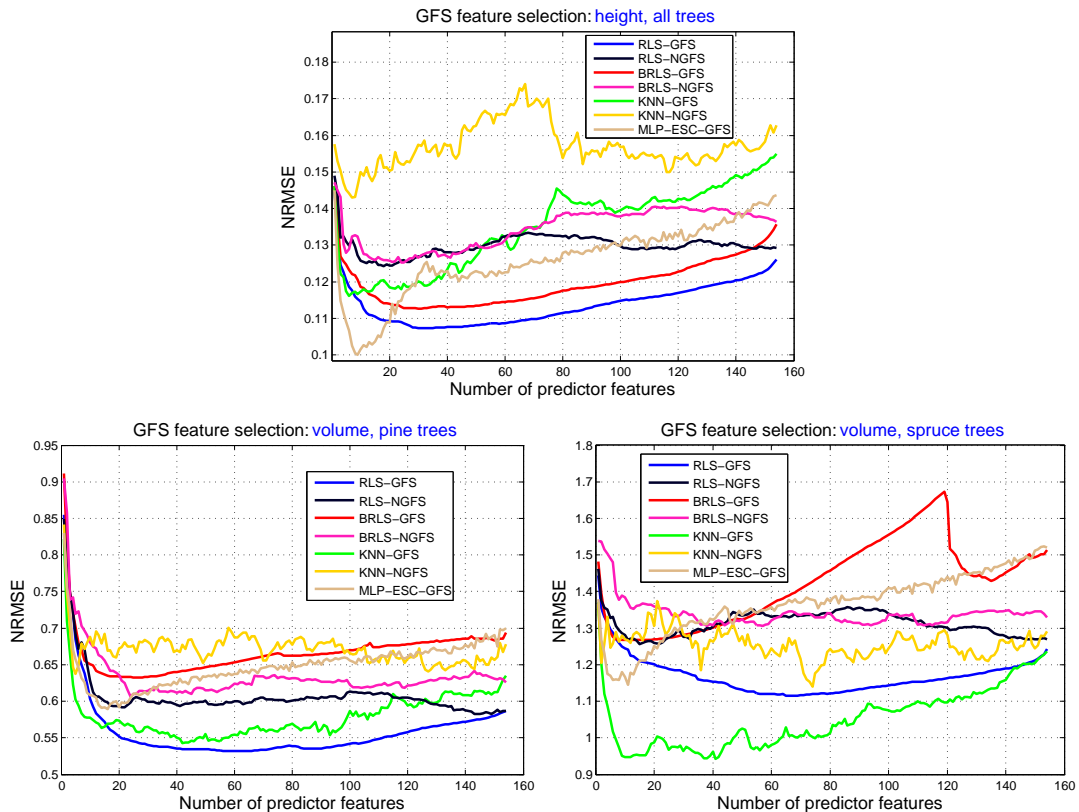


Figure 3. GFS and NGFS feature selection results for h , v_p and v_s . NRMSE is shown as a function of number of predictor features.

has a NRMSE value of 0.943 with 41 features. Linear methods tend to be more stable but do not have enough expressive power to capture the (most likely a nonlinear) pattern in the data.

Detailed results for the analyses are presented in Table 4. We can see that the NRMSE results of each variable are pretty stable with small variance across all the analyses. The results are mixed with linear methods being the best choice with some inventory variables (diameter d , basal area b , volumes v_a and v_p), whereas for others (height h , volumes v_s and v_b) nonlinear methods obtain the best results. Overall, the prediction performance results in terms of R^2 are lowest for pine and spruce volumes and diameter. Linear methods seem to require more features for successful prediction than nonlinear methods. After zeroing the negative predictions, they also produce biased predictions for volumes of spruce and broadleaf, which have large proportions of zero observations. Data balancing with the BRLS shows worse results than their unbalanced counterparts. This could result from the fact, that even though low frequency data points are given more weight the balancing causes now more error for the large frequency data points. Also nesting by NGFS shows worse results than without nesting, giving 0.01–0.08 and 0.04–0.10 lower R^2 values for RLS-NGFS and KNN-NGFS.

As there were no big differences in results of RLS-GA and KNN-GA with the feature sets selected in nine runs, results of each target variable in Table 4 are after that run that gave median NRMSE for the variable. For MVKNN-GA, one feature selection run out of 18 was chosen by calculating ranks of NRMSEs for each target variable and then finding the run with the minimum highest rank among the variables. In the chosen run, 26 features were selected and highest rank was 11th for volume v_b .

When comparing MVKNN-GA to the univariate methods, excluding balanced and nested ones, R^2 values of MVKNN-GA are 0.03–0.11 lower than that of the best univariate method on the target variables. The difference is smallest for volume v_a and largest for diameter d and volume v_s . Prediction error levels of a variable by MVKNN-GA are to some degree affected by the subjective weights given for the target variables in feature selection and further by the chosen selection run. The variability of results in the 18 selection runs was largest for volumes v_s and v_b .

It was of interest to also study the stability of the feature selection processes if subjected to changes in the data set. As we discussed earlier, the NGFS gives us a way to do this since the feature selection is implemented multiple times (depending on the fold size N) with slightly different data sets each round. With the three NGFS analyses (RLS-NGFS, BRLS-NGFS, KNN-NGFS) and fold size $N = 10$ we get a total of 30 different feature selection results. These results are illustrated in the Figures A, B and C in the appendix for target variables h , v_p and v_s respectively. The graphs show which features among the 154 were in general selected most often and ranked the highest. The higher the feature (represented by a colored circle in the plot) is in the graph, the more often it was selected to the optimum feature set. The color of the circle in the graphs represents the feature’s average rank. This will help to rank features which are on the same level on the y-axis. For example, two features f_1 and f_2 might have been selected the same number of times to the optimal feature set, but f_1 could always have been the first feature to be selected to optimum feature set, while f_2 could always have been the last one to be selected. The greener the feature’s circle in the graph is, the higher it is ranked among the features. It can be noticed from the figures that the feature selection for the target inventory variable h is more clear than for v_s . For variable h there seems to be a more stronger connection between it and the features than for v_s . This can be seen from the fact that for h in Figure A the graph has a steeper rise than in Figure C, meaning the feature selection is more

Table 4. Results of comparison of estimators. Negative bias indicates underestimation. The #features row refers to the number of features in the feature set with optimal model prediction performance. The definitions for the other statistics can be found from Materials and methods section.

Statistic	Analysis	d	h	b	v_a	v_p	v_s	v_b
NRMSE	RLS-GFS	0.210	0.107	0.220	0.250	0.532	1.114	0.857
	RLS-NGFS	0.236	0.124	0.241	0.272	0.582	1.255	0.959
	BRLS-GFS	0.239	0.113	0.241	0.358	0.632	1.267	0.938
	BRLS-NGFS	0.238	0.125	0.245	0.349	0.606	1.306	1.001
	KNN-GFS	0.232	0.116	0.241	0.266	0.542	0.943	0.716
	KNN-NGFS	0.257	0.143	0.270	0.315	0.638	1.140	0.879
	MLP-ESC-GFS	0.213	0.100	0.228	0.271	0.590	1.145	0.971
	MVKNN-GA	0.261	0.142	0.269	0.295	0.604	1.187	0.852
	KNN-GA	0.241	0.122	0.248	0.263	0.553	0.979	0.701
	RLS-GA	0.219	0.112	0.225	0.256	0.560	1.217	0.913
BIAS%	RLS-GFS	0.3	0.1	0.0	0.3	4.1	19.9	13.5
	RLS-NGFS	0.2	0.1	0.2	0.0	2.1	15.7	10.2
	BRLS-GFS	-2.7	0.0	1.3	-12.8	-13.0	19.9	1.1
	BRLS-NGFS	-0.4	0.0	-0.1	-11.1	4.8	19.5	14.3
	KNN-GFS	-0.8	-0.9	-1.0	-1.0	2.2	-0.7	-7.9
	KNN-NGFS	-0.5	-0.7	-1.3	-2.5	1.2	2.3	-7.9
	MLP-ESC-GFS	0.3	0.1	-0.1	-0.8	-5.0	-7.3	-6.3
	MVKNN-GA	0.0	-0.6	0.7	-0.9	0.1	2.1	-6.0
	KNN-GA	0.1	0.0	0.0	0.0	-0.1	0.1	0.0
	RLS-GA	0.2	0.0	0.1	0.1	3.6	17.0	11.8
R^2	RLS-GFS	0.78	0.93	0.89	0.91	0.75	0.74	0.82
	RLS-NGFS	0.73	0.91	0.87	0.90	0.70	0.66	0.77
	BRLS-GFS	0.72	0.93	0.87	0.82	0.65	0.66	0.78
	BRLS-NGFS	0.72	0.91	0.86	0.83	0.67	0.64	0.75
	KNN-GFS	0.74	0.92	0.87	0.90	0.74	0.81	0.87
	KNN-NGFS	0.68	0.88	0.83	0.86	0.64	0.72	0.81
	MLP-ESC-GFS	0.78	0.94	0.88	0.90	0.69	0.72	0.76
	MVKNN-GA	0.67	0.88	0.83	0.88	0.68	0.70	0.82
	KNN-GA	0.72	0.91	0.86	0.91	0.73	0.80	0.88
	RLS-GA	0.77	0.93	0.88	0.91	0.72	0.69	0.79
#features	RLS-GFS	45	30	35	41	59	66	73
	RLS-NGFS	24	18	18	14	143	15	146
	BRLS-GFS	24	30	46	19	25	17	50
	BRLS-NGFS	13	24	33	2	24	57	41
	KNN-GFS	18	6	44	51	42	41	43
	KNN-NGFS	67	7	13	2	5	74	42
	MLP-ESC-GFS	31	9	11	2	16	11	6
	MVKNN-GA	26	26	26	26	26	26	26
	KNN-GA	22	16	18	22	25	16	12
	RLS-GA	22	21	23	17	30	22	21

definite with h than v_s . According to the figures, features 139 (h_DE, selected 30 out of 30 times), 67 (i60f, selected 20 out of 30 times) and 118 (nir_ASM, selected 19 out of 30 times) were the top three features for h . The corresponding features for v_p were 52 (d7l, selected 27 out of 30 times), 140 (i_ASM, selected 25 out of 30 times) and 119 (nir_Contr, selected 24 out of 30 times), and for v_s they were 69 (i20l, selected 28 out of 30 times), 48 (d3l, selected 27 out of 30 times) and 71 (i60l, selected 27 out of 30 times). For descriptions of these features see the Tables A and B in the appendix.

Discussion

In area-based forest inventory, ALS data gives a distribution of height and intensity values, which can be described with numerous features, e.g., percentiles or proportional canopy point densities. Further, spectral information from aerial images help in tree species-specific modeling (e.g. Fassnacht et al. 2016) and texture features both from the rasterized ALS data and aerial images may contain structural information about the target element. The number of possible predictor features is thus large, as shown by the 154 features derived in this study. This feature set clearly has multicollinearity problems. Many of the features can also be irrelevant or noisy predictors for some response variable, so we used two feature selection methods to get near-optimal feature sets for our seven response variables. There is no guarantee of finding the optimal predictor variable subset (Garey and Johnson 1979), since the algorithms do not go through all possible combinations but solutions close to optimal can usually be found in a feasible computation time. GFS was used because as a sequential method it selects features in order, enabling examination of feature rankings. One stochastic method, GA, was selected for comparison, as it has been used earlier with one of our estimators, KNN (Tuominen et al. 2014). The other non-parametric estimator tested was MLP-ESC. Third estimator was RLS, also known as ridge regression, for which a fast training algorithm (GFS) for sparse linear predictors has been developed (Pahikkala et al. 2010). Ridge regression is commonly used to address the problem of multicollinearity.

Both of the feature selectors were used with two estimators, KNN and RLS. The differences in prediction accuracy between GFS and GA were minor, as NRMSE and thus also R^2 accuracies were almost equal on all response variables with the same estimator. The largest differences were between RLS-GFS and RLS-GA in the volumes by tree species, for which RLS-GFS produced 0.03-0.05 higher R^2 values than RLS-GA, but then, in these volumes KNN performed better or about equally. GFS selected in general more features than GA but this is affected by the parameters used in GA as well. Also the estimator affects the number of selected features as linear GFS-RLS selected a great number of features whereas non-parametric MLP-ESC-GFS selected clearly less. While large feature sets can be processed on current computers, more features can still cause more time and memory consumption problems in the calculation of the actual inventory area results.

One property of estimators is whether they extrapolate. Contrary to KNN, RLS and to some degree MLP-ESC extrapolate, which has both pros and cons. In our data set, the volumes of spruce and broadleaf had a lot of zero or close to zero observations, leading to negative predictions that are not feasible. These were replaced with zero, which caused bias and thus lower NRMSE values with RLS. In this case, the amount of bias could probably have been reduced by some transformations of the dependent or independent variables, but due to the already large number of features this was not

tried. However, transforming variables could also had reduced the number of features selected by RLS-GFS for the best combination.

A nested version of GFS was run to study the sensitivity of the feature selection to changes in the data set. In a nested cross validation, model building, including feature selection, and prediction are repeated so that the same observation is not used in both. In a 10-fold nested CV, prediction accuracies of KNN decreased more than those of RLS. This shows the other side of the capability to extrapolate: KNN does not extrapolate and thus it can be more sensitive to the changes in the data set. Another point is that the training data CV results without nesting can be too optimistic for an unknown observation, because the model building is affected by each training observation. In a study of KNN variants, each training observation was used for nesting and then LOOCV without nesting increased RMSE in general less than 5% but in some cases over 10% (Packalén et al. 2012). With our 10-fold nested CV, the corresponding increase was 8-14% for RLS-GFS and 10-19% for KNN-GFS.

In the NGFS analysis of the stability of the feature selection, mean height and volume of spruce were the extremities. Height has more straightforward relation with a number of features, mainly those based on the height distribution of reflected ALS returns and textural features of the rasterized canopy height model. Still, the top three features `h_DE`, `i60f` and `nir_ASM` are somewhat surprising, as there is a texture feature of the aerial imagery and an intensity feature. The height feature is also from texture and not from more direct height distribution statistics, of which `h100f` is the 4th feature. These may be linked to the used relascope field plot, where the distance for inclusion of a tree depends on its diameter: only trees having diameter 18 cm or larger are measured at all distances up to the maximum radius of 9 m. This can cause that on some field plots no or a few small trees are measured but e.g. `h100f` can have a value of over 10 m. Part of the crowns from trees outside the plot can extend into the plot in a fixed radius plot as well though this problem is pronounced in a relascope plot. We could calculate predictor features according to the inclusion radius of the largest tree measured but then the sample plot size in the calculation of the features would decrease when the largest tree has a small diameter.

The volume of spruce has a more complex relation with the predictor features in this data, and the distinction between frequently selected and other features in the NGFS analysis is not obvious: there were no features with zero occurrence in the feature set producing optimal prediction performance. This is probably linked to the problems in the identification of tree species, because the lowest prediction performance results were for pine and spruce volumes. There was no spruce on many field plots and otherwise spruce was the dominant species only on a few plots. As a shade tolerant tree species, spruce can also exist in the lower canopy storeys, which complicates the separation of spruce from the other species further. Again here in the NGFS analysis, the top features were somewhat surprising: three out of four top features were from the ALS intensity distributions. It can be that the intensity features help here in finding plots without spruce and in separation of the volumes of spruce and pine.

To conclude, we implemented analyses for seven key inventory variables and noted predictions to be the most reliable for tree height h (all trees) and the least reliable for tree volumes v_p (pine) and v_s (spruce) and diameter d . On average, the most difficult target inventory variables to predict were tree volumes by tree species. In addition to the prediction performance, this was noted from the feature selection results. In the case of tree height for example, the NGFS analysis (see Figure A in the appendix) showed that particular and relatively small amount of features were

selected often, suggesting the presence of stronger connection between the features and the target variable. With tree volume on the other hand, the NGFS analysis showed the estimators utilizing greater number of features, indicating a weaker connection between the features and target variable. The low prediction performance for volumes of broadleaf and spruce can be partly explained by the small proportion of these species in Aland region. The results demonstrated that for an individual variable around 20-40 features were sufficient for near-optimal prediction, while the ranking of linear (RLS) and nonlinear (KNN and MLP-ESC) estimators varied between the target variables.

References

- Araghinejad S. 2014. Regression-based models. Springer Netherlands. p. 49–83.
- Axelsson P. 1999. Processing of laser scanner data - algorithms and applications. *Journal of Photogrammetry and Remote Sensing*. 54(23):138–147.
- Baltsavias EP. 1999. A comparison between photogrammetry and laser scanning. *Journal of Photogrammetry and Remote Sensing*. 54(2-3):83–94.
- Beyer K, Goldstein J, Ramakrishnan R, Shaft U. 1999. When is “nearest neighbor” meaningful? Springer Berlin Heidelberg. p. 217–235.
- Bishop C. 1995. Neural networks for pattern recognition. Oxford University Press.
- Breidenbach J, Næsset E, Lien V, Gobakken T, Solberg S. 2010. Prediction of species specific forest inventory attributes using a nonparametric semi-individual tree crown approach based on fused airborne laser scanning and multispectral data. *Remote Sensing of Environment*. 114(4):911–924.
- Broadhurst D, Goodacre R, Jones A, Rowland JJ, Kell DB. 1997. Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta*. 348(1):71–86.
- Chawla NV. 2005. Data mining for imbalanced datasets: An overview. In: Maimon O, Rokach L, editors. *The Data Mining and Knowledge Discovery Handbook*. p. 875–886.
- Eerikäinen K. 2009. A multivariate linear mixed-effects model for the generalization of sample tree heights and crown ratios in the finnish national forest inventory. *Forest Science*. 55(6):480–493.
- Fassnacht FE, Latifi H, Stereńczak K, Modzelewska A, Lefsky M, Waser LT, Straub C, Ghosh A. 2016. Review of studies on tree species classification from remotely sensed data. *Remote Sensing of Environment*. 186:64–87.
- Garey MR, Johnson DS. 1979. Computers and intractability; a guide to the theory of np-completeness. New York, NY, USA: W. H. Freeman & Co.
- Gruen A, Zhang L. 2003. Automatic dtm generation from tls data. In: Gruen, Kahman (eds.), *optical 3-d measurement techniques vi*. vol. 1. p. 93–105.
- Haapanen R, Tuominen S. 2008. Data combination and feature selection for multi-source forest inventory. *Photogrammetric Engineering and Remote Sensing*. 74(7):869–880.
- Haralick RM, Shanmugam K, Dinstein I. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*. SMC-3(6):610–621.
- Hinneburg A, Aggarwal CC, Keim DA. 2000. What is the nearest neighbor in high dimensional spaces? In: *Proceedings of the 26th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc. p. 506–515. VLDB '00.
- Hoerl AE, Kennard RW. 2000. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. 42(1):80–86.
- Hsu Cw, Chang Cc, Lin Cj. 2010. A practical guide to support vector classification.
- Hudak A, Crockston N, Evans J, Hall D, Falkowski M. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*. 113(1):289–290.

- Hyypä J, Pyysalo U, Hyypä H, Samberg A. 2000. Elevation accuracy of laser scanning-derived digital terrain and target models in forest environment. In: Proceedings of EARSeL-SIG-Workshop LIDAR. p. 14–17.
- Jain A, Zongker D. 1997. Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 19(2):153–158.
- Koch B, Heyder U, Weinacker H. 2006. Detection of individual tree crowns in airborne lidar data. *Photogrammetric Engineering & Remote Sensing*. 72(4):357–363.
- Kudo M, Sklansky J. 1998. Classifier-independent feature selection for two-stage feature selection. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 548–554.
- Kudo M, Sklansky J. 2000. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*. 33:25–41.
- Laasasenaho J. 1982. Taper curve and volume functions for pine, spruce and birch [dissertation]. *Communicationes Instituti Forestalis Fenniae* 108.
- Latifi H, Nothdurft A, Koch B. 2010. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. *Forestry: An International Journal of Forest Research*. 83(4):395–407.
- Lefsky MA, Harding D, Cohen WB, Parker G, Shugart HH. 1999. Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern maryland, usa. *Remote Sensing of Environment*. 67(1):83–98.
- Maltamo M, Malinen J, Packalén P, Suvanto A, Kangas J. 2006. Nonparametric estimation of stem volume using airborne laser scanning, aerial photography, and stand-register data. *Canadian Journal of Forest Research*. 36(2):426–436.
- Moser P, Vibrans AC, McRoberts RE, Næsset E, Gobakken T, Chirici G, Mura M, Marchetti M. 2017. Methods for variable selection in LiDAR-assisted forest inventories. *Forestry: An International Journal of Forest Research*. 90(1):112–124.
- Nabney I. 2004. *Netlab: Algorithms for pattern recognition*. Springer. 1.
- Næsset E. 1997. Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment*. 61(2):246–253.
- Næsset E. 2002. Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment*. 80(1):88–99.
- Næsset E. 2004. Accuracy of forest inventory using airborne laser scanning: evaluating the first nordic full-scale operational project. *Scandinavian Journal of Forest Research*. 19(6):554–557.
- Packalén P, Maltamo M. 2007. The k-msn method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. *Remote Sensing of Environment*. 109(3):328–341.
- Packalén P, Suvanto A, Maltamo M. 2009. A two stage method to estimate species specific growing stock by combining als data and aerial photographs of known orientation parameters. *Photogrammetric Engineering and Remote Sensing*. 75(12):1451–1460.
- Packalén P, Temesgen H, Maltamo M. 2012. Variable selection strategies for nearest neighbor imputation methods used in remote sensing based forest inventory. *Canadian Journal of Remote Sensing*. 38(5):557–569.
- Pahikkala T, Airola A, Salakoski T. 2010. Speeding up greedy forward selection for regularized least-squares. In: 2010 Ninth International Conference on Machine Learning and Applications. p. 325–330.
- Pudil P, Novovičová J, Kittler J. 1994. Floating search methods in feature selection. *Pattern Recognition Letters*. 15(11):1119–1125.
- Pyysalo U. 2000. A method to create a three dimensional forest model from laser scanner data. *Photogrammetric Journal of Finland*. 17(1):34–42.
- Siedlecki W, Sklansky J. 1989. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*. 10(5):335–347.
- Tomppo E, Haakana M, Katila M, Peräsaari J. 2008. Multi-source national forest inventory - methods and applications. *Managing Forest Ecosystems*:374.
- Tomppo E, Katila M, Mäkisara K, Peräsaari J. 2013. The multi-source national forest inventory

- of Finland - methods and results 2009. Working Papers of the Finnish Forest Research Institute 273:216.
- Törmä M. 2000. Estimation of tree species proportions of forest stands using laser scanning. *International Archives of Photogrammetry and Remote Sensing*. 33:1524–1531.
- Tuominen S, Haapanen R. 2013. Estimation of forest biomass by means of genetic algorithm-based optimization of airborne laser scanning and digital aerial photograph features. *Silva Fennica*. 47(1):20.
- Tuominen S, Pekkarinen A. 2005. Performance of different spectral and textural aerial photograph features in multi-source forest inventory. *Remote Sensing of Environment*. 94(2):256–268.
- Tuominen S, Pitkänen J, Balazs A, Korhonen K, Hyvönen P, Muinonen E. 2014. Nfi plots as complementary reference data in forest inventory based on airborne laser scanning and aerial photography in Finland. *Silva Fennica*. 48(2):1–20.
- Van Coillie F, Verbeke L, De Wulf R. 2005. GA-driven feature selection in object-based classification for forest mapping with IKONOS imagery in Flanders, Belgium. In: Olsson H, editor. *Proceedings of ForestSat 2005 in Borås May 31-June 3 : Skogsstyrelsen report 8b*. National Board of Forestry. p. 11–15.
- Vilhomaa J, Laaksonen H. 2011. Valtakunnallinen laserkeilaus - testityöstä tuotantoon. *The Photogrammetric Journal of Finland*. 22(3):82–91.
- Waser LT, Ginzler C, Kuechler M, Baltsavias E, Hurni L. 2011. Semi-automatic classification of tree species in different forest ecosystems by spectral and geometric variables derived from airborne digital sensor (ads40) and {RC30} data. *Remote Sensing of Environment*. 115(1):76–85.

Appendix

Table A. List of the ALS and aerial imagery predictor features used in the analyses. Each feature has an identifier number (gray cell) which is followed by the feature name. The descriptions for the corresponding features can be found in Table B.

1	havgf	32	h100l	63	p80l	94	b_DV	125	nir_SE
2	havgl	33	hcvf	64	p95l	95	b_DE	126	nir_Entr
3	hstdf	34	hcvl	65	i20f	96	g_ASM	127	nir_DV
4	hstdl	35	vegf	66	i40f	97	g_Contr	128	nir_DE
5	h0f	36	vegl	67	i60f	98	g_Corr	129	h_ASM
6	h5f	37	d1f	68	i80f	99	g_Var	130	h_Contr
7	h10f	38	d2f	69	i20l	100	g_IDM	131	h_Corr
8	h20f	39	d3f	70	i40l	101	g_SA	132	h_Var
9	h30f	40	d4f	71	i60l	102	g_SV	133	h_IDM
10	h40f	41	d5f	72	i80l	103	g_SE	134	h_SA
11	h50f	42	d6f	73	b_mean	104	g_Entr	135	h_SV
12	h60f	43	d7f	74	b_std	105	g_DV	136	h_SE
13	h70f	44	d8f	75	b_cv	106	g_DE	137	h_Entr
14	h80f	45	d9f	76	g_mean	107	r_ASM	138	h_DV
15	h85f	46	d1l	77	g_std	108	r_Contr	139	h_DE
16	h90f	47	d2l	78	g_cv	109	r_Corr	140	i_ASM
17	h95f	48	d3l	79	r_mean	110	r_Var	141	i_Contr
18	h100f	49	d4l	80	r_std	111	r_IDM	142	i_Corr
19	h0l	50	d5l	81	r_cv	112	r_SA	143	i_Var
20	h5l	51	d6l	82	nir_mean	113	r_SV	144	i_IDM
21	h10l	52	d7l	83	nir_std	114	r_SE	145	i_SA
22	h20l	53	d8l	84	nir_cv	115	r_Entr	146	i_SV
23	h30l	54	d9l	85	b_ASM	116	r_DV	147	i_SE
24	h40l	55	p20f	86	b_Contr	117	r_DE	148	i_Entr
25	h50l	56	p40f	87	b_Corr	118	nir_ASM	149	i_DV
26	h60l	57	p60f	88	b_Var	119	nir_Contr	150	i_DE
27	h70l	58	p80f	89	b_IDM	120	nir_Corr	151	ndvi
28	h80l	59	p95f	90	b_SA	121	nir_Var	152	gndvi
29	h85l	60	p20l	91	b_SV	122	nir_IDM	153	nir.r
30	h90l	61	p40l	92	b_SE	123	nir_SA	154	nir.g
31	h95l	62	p60l	93	b_Entr	124	nir_SV	-	-

Table B. Descriptions of the ALS and aerial imagery features listed in Table A. Left column contains the feature names and the right column descriptions. The square brackets in the left column mean that multiple options can be placed there, e.g. veg[f/l] means either vegf or vegl and h[p][f/l] can be e.g. h50f or h85l. For more information see the article sections about ALS and aerial imagery data. H means height above ground.

hav[f/l], hstd[f/l], hcv[f/l]	Average, standard deviation and coefficient of determination of first/last returned canopy returns
h[p][f/l]	H at which $p\%$ of cumulative sum of first/last canopy returns is achieved (H_p)
veg[f/l]	Ratio of first/last canopy returns to all first/last returns
d[s][f/l]	Percentage of first/last canopy returns above height limits calculated in each field plot from $H_{\min} + \frac{s}{10} \times (H_{\max} - H_{\min})$, $s \in \{1, 2, \dots, 9\}$
p[p][f/l]	Percentage of first/last canopy returns having $H \geq$ than corresponding H_p
i[p][f/l]	Ratio of intensity percentile p to the median of intensity for first/last canopy returns
[r/g/b/nir]_[mean/std/cv]	Average, standard deviation and coefficient of determination of red/green/near-infrared bands of CIR imagery and blue band of RGB imagery
[r/g/b/nir]_[ASM/Contr/Corr/Var/IDM/SA/SV/SE/Entr/DV/DE]	Texture features: angular second moment/contrast/correlation/variance/inverse difference moment/sum average/sum variance/sum entropy/entropy/difference entropy of red/green/near-infrared bands of CIR imagery and blue band of RGB imagery
[h/i]_[ASM/Contr/Corr/Var/IDM/SA/SV/SE/Entr/DV/DE]	Texture features: angular second moment/contrast/correlation/variance/inverse difference moment/sum average/sum variance/sum entropy/entropy/difference entropy of ALS based canopy height and intensity
ndvi	Transformation from band averages within the pixel windows: nir-r/nir+r
gndvi	Transformation from band averages within the pixel windows: nir-g/nir+g
nir.r	Transformation from band averages within the pixel windows: nir/r
nir.g	Transformation from band averages within the pixel windows: nir/g

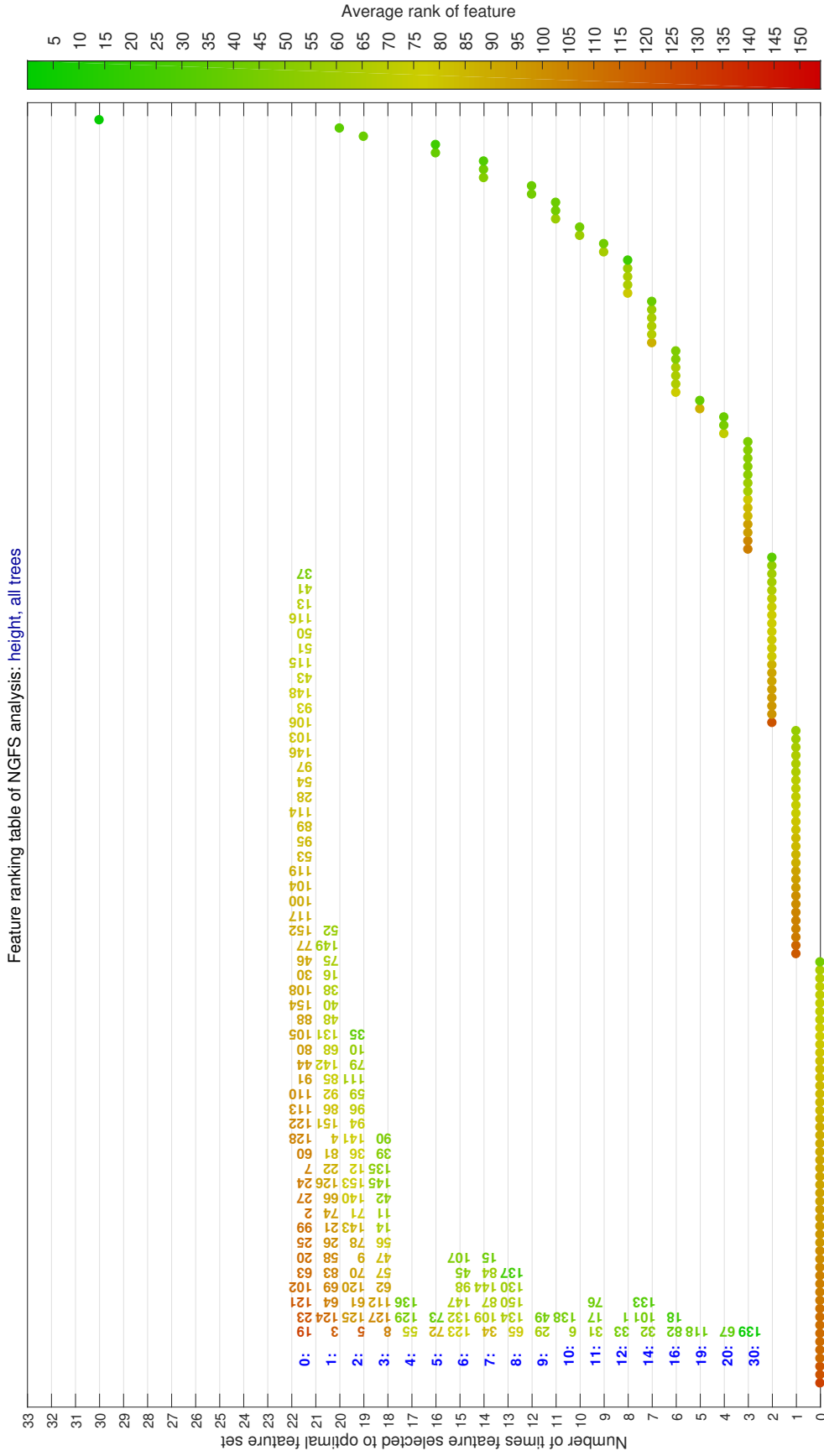


Figure A. Feature ranking table of NGFS analysis for height (*h*, all trees) target variable. Each circle represents a single feature of the 154 ALS and aerial imagery predictor features. The position of each circle on the y-axis tells how many times that feature was selected to the optimal feature set (out of the 30 GFS analyses) with the best prediction performance. The color of a circle represents the average relative ranking (relative to other features) of that feature. The feature identifier numbers (colored numbers) are also listed in the plot, which can be referred to the numbers in Table A. The blue number indicates the position of the corresponding features in the y-axis of the plot.

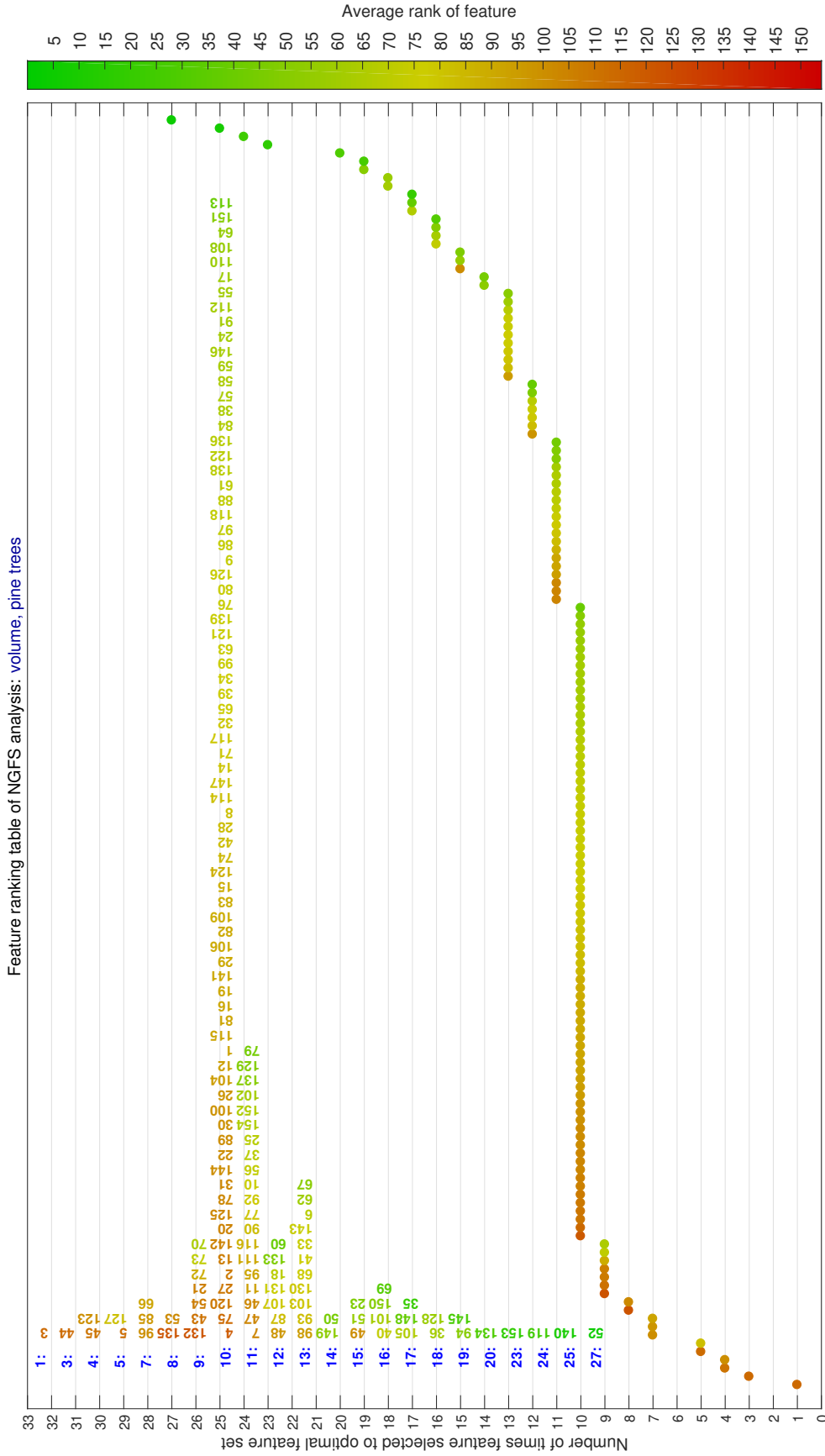
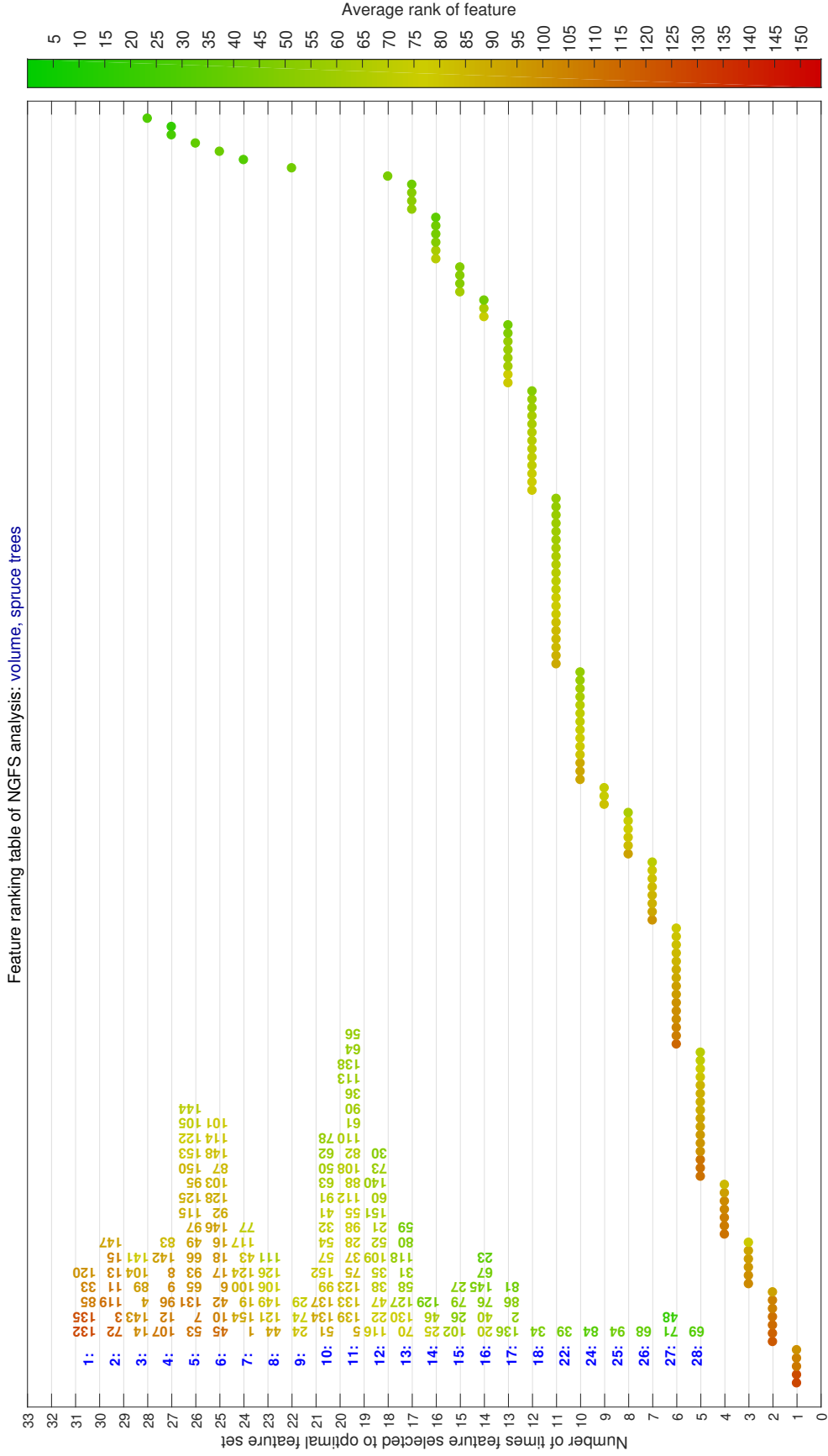


Figure B. Feature ranking table of NGFS analysis for volume (v_p , pine trees) target variable. Each circle represents a single feature of the 154 ALS and aerial imagery predictor features. The position of each circle on the y-axis tells how many times that feature was selected to the optimal feature set (out of the 30 GFS analyses) with the best prediction performance. The color of a circle represents the average relative ranking (relative to other features) of that feature. The feature identifier numbers (colored numbers) are also listed in the plot, which can be referred to the numbers in Table A. The blue number indicates the position of the corresponding features in the y-axis of the plot.



Publication VI

Reliable AUC estimation of spatial classifiers, with application to mineral prospectivity mapping

Antti Airola, Jonne Pohjankukka, Johanna Torppa, Maarit Middleton, Vesa Nykänen, Jukka Heikkonen and Tapio Pahikkala. Data Mining and Knowledge Discovery, Accepted. Springer, 2018.

Copyright © 2018 Springer. Reprinted with permissions from respective publisher and authors.

Reliable AUC estimation of spatial classifiers, with application to mineral prospectivity mapping

Antti Airola · Jonne Pohjankukka ·
Johanna Torppa · Maarit Middleton ·
Vesa Nykänen · Jukka Heikkonen ·
Tapio Pahikkala

Abstract Machine learning based classification methods are widely used in geoscience applications, including mineral prospectivity mapping. Typical characteristics of the data, such as small number of positive instances, imbalanced class distributions and lack of verified negative instances, necessitate the use of ROC analysis and cross-validation for classifier evaluation. However, recent literature has identified two sources of bias, that can affect reliability of area under ROC curve (AUC) estimation via cross-validation on spatial data. The pooling procedure, performed by methods such as leave-one-out can introduce a substantial negative bias to results. At the same time, spatial dependencies leading to spatial autocorrelation can result in overoptimistic results, if not corrected for. In this work, we introduce the spatial leave-pair-out cross-validation method, that corrects for both of these biases simultaneously. The methodology is used to benchmark a number of classification methods on mineral prospectivity mapping data from Central Lapland greenstone belt. The evaluation highlights the dangers of obtaining misleading results on spatial data and demonstrates how these problems can be avoided. Further, the results show the advantages of simple linear models on this classification task.

Keywords area under ROC curve · classifier evaluation · cross-validation · mineral prospectivity mapping · spatial data mining

This work was supported by the Academy of Finland (grants 289903, 311273).

Antti Airola · Jonne Pohjankukka · Jukka Heikkonen · Tapio Pahikkala
Department of Future Technologies, University of Turku, FI-20014, Turku, Finland
E-mail: forname.surname@utu.fi

Johanna Torppa · Maarit Middleton · Vesa Nykänen
Geological Survey of Finland
E-mail: forname.surname@gtk.fi

1 Introduction

Mineral prospectivity mapping or mineral potential mapping (MPM) techniques are used to delineate areas favorable for mineral exploration (see e.g. Bonham-Garter (1994); Carranza (2008); Nykänen (2008)). By integrating information derived from spatial geological, geophysical and geochemical datasets the MPM methodology is used to quantify the likelihood of presence of a specific type of mineral occurrence within a study area. In supervised MPM learning techniques, the locations of known mineral occurrences are used to relate the occurrences to the mapped quantities that are indicative of the corresponding mineral deposit type. Known mineral occurrences can be also used for validating the models (Bonham-Carter, 1994).

In this work, we consider the issue of supervised binary classification in spatial prediction problems. Here the goal is to train a classifier that can predict some property of a geographical area, such as the presence or absence of a mineral deposit. Training and evaluation of such classifiers is challenging because the available data is typically highly imbalanced and the amount of positive instances denoting known mineral occurrences is small. Further, instead of known negative instances, data sets usually contain only positive and unlabeled instances (see e.g. Nykänen (2008); Rigol-Sanchez et al (2003)); a setting known as positive-unlabeled (PU) learning (Elkan and Noto, 2008). Works such as Bradley (1997); Fawcett (2006); Huang and Ling (2005) have suggested the use of area under ROC curve (AUC) for classifier evaluation on imbalanced data, as the criterion is insensitive to relative class distributions on the test set. Further, AUC has also been established as a recommended metric for PU-learning problems (Elkan and Noto, 2008; Jain et al, 2017). Thus, AUC is a natural performance measure for MPM classifier evaluation, and studies such as Brown et al (2003); Nykänen (2008); Nykänen et al (2015); Rodriguez-Galiano et al (2015) have used AUC for evaluating MPM models. Further, since adequately large separate test data may not be afforded for MPM, cross-validation (CV) is necessary for validating the models (see e.g. Abedi et al (2012); Rigol-Sanchez et al (2003); Carranza (2008); Rodriguez-Galiano et al (2015)).

Based on recent literature we suggest that there are two major sources of bias, that can affect results when using cross-validation for estimating the AUC of spatial classification problems. First, standard CV methods such as leave-one-out (LOOCV) and K-fold are often affected by a negative bias resulting from pooling together predictions from different folds for AUC computation, as shown by Airola et al (2009, 2011); Forman and Scholz (2010); Parker et al (2007); Smith et al (2014). Airola et al (2009, 2011) propose a leave-pair-out CV (LPOCV) method for correcting such bias in AUC estimation. LPOCV is further validated by Smith et al (2014) on clinical data. Second, spatial autocorrelation causes standard CV methods to produce optimistically biased prediction performance estimates for spatial data. This is caused by fact that leave-one-out and K-fold relying on the assumption that the data is drawn independent and identically distributed (i.i.d.), an assumption violated

by spatial data where close instances tend to be more similar than ones distant from each other. Recently, Pohjankukka et al (2014, 2017); Le Rest et al (2014) have proposed spatial CV (SCV) methods for correcting this bias.

In this work, combining the leave-pair-out and spatial CV methods, we introduce the leave-pair-out spatial CV (LPO-SCV) method for evaluating MPM classifiers. As a case study, we use the approach to benchmark a number of machine learning methods on an orogenic gold MPM classification task. In our experiments, we first show that one can obtain completely misleading results, if the spatial and pooling biases are not corrected for. At worst, one can obtain with standard CV methods close to perfect AUC values for classifiers, that are in reality not much better than random on making predictions for new data. We demonstrate, how the LPO-SCV corrects the pooling and spatial biases, allowing one to reliably estimate the AUC of spatial classifiers. Finally, in the LPO-SCV based classifier comparison, we show simple linear models to be surprisingly competitive on the MPM data.

2 Cross-validation for AUC estimation with spatial data

First, we present our mathematical notation. Let us assume a set of m instances, divided into the so-called positive and negative classes. Further, let $\mathcal{I} = \{1, 2, \dots, m\}$ denote the index set of these instances, with $\mathcal{I}_+ \subset \mathcal{I}$ and $\mathcal{I}_- \subset \mathcal{I}$ denoting the indices of the positive and the negative instances, respectively. Further, let $f : \mathcal{I} \rightarrow \mathbb{R}$ denote a classifier, that maps each instance to a real-value, representing how likely it is to belong to the positive class. We can use f to classify data, by assigning each $f(i) > t$ to the positive class, and the rest to the negative class for some threshold t . Finally, when defining the cross-validation methods we refer by $f_{\mathcal{H}}$, where $\mathcal{H} \subseteq \mathcal{I}$, to a classifier trained with a machine learning method on the subset of the instances indexed by \mathcal{H} .

2.1 AUC

Area under the ROC curve (AUC) is a common criterion for evaluating the quality of a classifier. It estimates the probability, that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Hanley and McNeil, 1982). AUC is invariant to prior class distributions, and does not require one to define class specific error costs or a threshold t . These advantages make it a especially popular metric for classifier evaluation and comparison, especially in applications dealing with imbalanced data (Bradley, 1997; Fawcett, 2006; Huang and Ling, 2005).

AUC can be computed based on the Wilcoxon-Mann-Whitney statistic (Bamber, 1975) as

$$\frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} H(f(i) - f(j)), \quad (1)$$

where

$$H(a) = \begin{cases} 1, & \text{if } a > 0 \\ 0.5, & \text{if } a = 0 \\ 0, & \text{if } a < 0 \end{cases} \quad (2)$$

is the Heaviside step function, and $|\mathcal{I}_+|$ and $|\mathcal{I}_-|$ denote the number of positive and negative instances, respectively. That is, the fraction of positive-negative pairs, where the positive has higher prediction than the negative is computed. The AUC value is between 0 and 1, with 0.5 corresponding to a random classifier, or one that always predicts the same value.

2.2 Pooling bias and LPOCV

When dealing with data sets where at least one of the classes has only a small number of instances belonging to it, cross-validation is typically used for computing the AUC. However, recently it has been shown that standard cross-validation methods such as leave-one-out (LOO) and K-fold cross-validation can have a large negative bias, when used for computing AUC (Airola et al, 2009, 2011; Forman and Scholz, 2010; Smith et al, 2014; Parker et al, 2007). This effect is related to a procedure known as pooling, where predictions from different rounds of cross-validation are compared when computing AUC. Thus we refer to this effect as *pooling bias*. Recently, Airola et al (2009, 2011); Smith et al (2014) have shown that the pooling bias can be eliminated by using leave-pair-out cross-validation (LPO). In LPO, each positive-negative pair is left out in turn of the training set, and the classifier trained on the remaining instances. The LPO AUC estimate is then computed as the fraction of pairs, where the positive instance has a higher prediction than the negative one.

Formally, this can be defined as

$$\frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} H(f_{\mathcal{I} \setminus \{i,j\}}(i) - f_{\mathcal{I} \setminus \{i,j\}}(j)), \quad (3)$$

where $f_{\mathcal{I} \setminus \{i,j\}}$ is the classifier trained without the i :th and j :th instances.

For an example of pooling bias, let us consider a trivial classifier $f(i) = \frac{|\mathcal{I}_+|}{m}$, that just predicts the fraction of positive instances in the training set. In leave-one-out, the classifier would always obtain AUC of 0, since it would predict $\frac{|\mathcal{I}_+|}{m-1}$ when a negative instance is left out, and $\frac{|\mathcal{I}_+|-1}{m-1}$ when a positive instance is left out. While this is an extreme example, the strong effect pooling bias can have has been established experimentally in several studies (Airola et al, 2009, 2011; Forman and Scholz, 2010; Smith et al, 2014; Parker et al, 2007), and is further validated by our results. LPO avoids this problem, as it never compares predictions made by different classifiers. Airola et al (2011) show that LPO is almost unbiased, meaning that it provides an unbiased estimate of AUC for a classifier trained on $m - 2$ instances.

2.3 Spatial bias and SCV

Most of the methodologies in statistical inference rely on the assumption that data samples are realizations from (i.i.d.) random variables. In cases where we are concerned with spatio-temporal data sets this assumption can have major drawbacks. Take for example geographical instances sampled from the soil. We are given three instances i , j , and k with i and j located geographically much closer to each other than k to both the previous two. Anyone could argue in this scenario that i and j are probably the most similar to each other among the three instances due to the small geographical distance between them. In 1970 Waldo R. Tobler stated in his work (Tobler, 1970) the Tobler’s first law of geography: *“Everything is related to everything else, but near things are more related than distant things”*.

The relationship of being near versus being similar in spatial data analysis is called *spatial autocorrelation* (SAC). SAC in spatial data sets is usually measured quantitatively using e.g. variograms or Moran’s index (Cressie, 2015; Longley et al, 2005). SAC tends to be naturally high for instances close to each other and small for instances more distant from each other. It is therefore clear that when we have a set of geographical data samples, they are most certainly not i.i.d., which needs to be addressed in the model evaluation and selection.

To estimate model’s prediction performance where the effect of SAC has been reduced Pohjankukka et al (2014); Le Rest et al (2014); Pohjankukka et al (2017) proposed *spatial cross validation* (SCV) to be used for this purpose. The idea in SCV is to estimate a model’s prediction performance for a test point r_δ units away from the closest known instances. This is conducted by altering the data in the CV procedure, so that a test point will always be at least r_δ units away from the training data. Following Pohjankukka et al (2017), we call this left out area the *deadzone*. SCV produces a prediction performance estimate of our model as a function of r_δ , i.e. the distance of closest known data to the predicted instance. Thus SCV simulates the situation, where the trained model is used to make predictions for data that is further than r_δ units of distance from the instances in the training data.

2.4 Spatial Leave-pair-out Cross-Validation

In order to eliminate both the biases caused by pooling and spatial autocorrelation simultaneously, we now introduce the LPO-SCV method, that combines the LPO and SCV methods. The method is illustrated in Figure 1. In LPO-SCV, on each round of CV a positive-negative pair, and all the instances within r_δ radius of these two points, are left out of the training set. The model is trained on the rest of the training set, and predictions are made for the left out positive and negative instance. The AUC estimate is the fraction on pairs, for which the positive instance has a higher predicted value, than the negative one. The procedure is repeated for all possible positive-negative pairs.

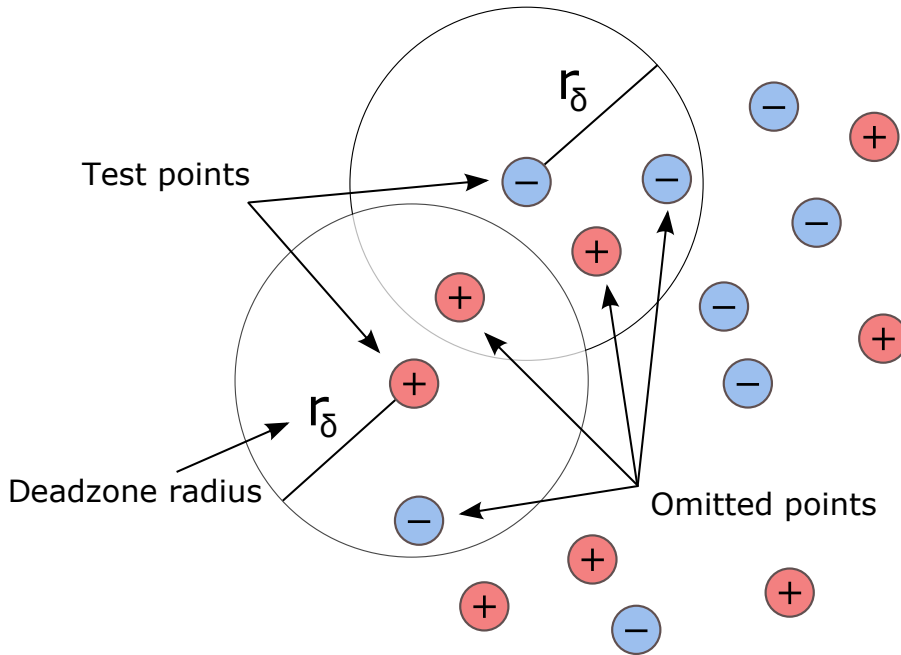


Fig. 1 Leave-pair-out Spatial Cross-Validation. On each round, a positive and negative instance are left out, as well as all the instances within the deadzone circles surrounding them. Thus the CV procedure simulates the setting, where the left out test pair is at least r_δ distance away from nearest training instance.

Formally, the estimate can be defined as follows. Let $d(i, j)$ denote the geographical distance (e.g. Euclidean) between the i :th and j :th training instances. Further, let $\mathcal{U}(i, j) = \{k \in \mathcal{I} | d(i, k) > r_\delta \wedge d(j, k) > r_\delta\}$ denote all training instances that have a larger distance than r_δ from both i :th and j :th training instance. Then, the LPO-SCV is computed as

$$\frac{1}{|\mathcal{I}_+||\mathcal{I}_-|} \sum_{i \in \mathcal{I}_+} \sum_{j \in \mathcal{I}_-} H(f_{\mathcal{U}(i,j)}(i) - f_{\mathcal{U}(i,j)}(j)), \quad (4)$$

where $f_{\mathcal{U}(i,j)}(i)$ is the classifier trained on all data outside circles of radius r_δ around instances i and j .

Similarly to the ordinary LPO-CV, the approach corrects for pooling bias by ensuring that only predictions made on the same round of cross-validation are ever compared. At the same time, the method corrects for spatial bias by excluding instances too near test data from the training data. The deadzone ensures that the AUC result holds for data further than r_δ units from the training instances, not just in the immediate neighborhood of the training data.

A downside of the approach is computational complexity, as full LPO-SCV requires training the classifier $|\mathcal{I}_+||\mathcal{I}_-|$ times. When this is not computationally feasible, one may approximate full LPO-SCV by randomly sampling a

subset of all the possible pairs. Further, for ridge regression classifiers, fast LPO-SCV cross-validation can be implemented using the fast holdout algorithms (Pahikkala et al, 2012) implemented in the RLScore open source library (Pahikkala and Airola, 2016).

3 Data

We chose to experiment the LPO-SCV method for prospectivity modelling of orogenic gold occurrences in the Central Lapland greenstone belt (CLGB). As positive instances, we used the locations of known orogenic gold occurrences, and as negative instances, a random selection of locations in the study area. As evidence features, we used rasters derived from airborne and ground based geophysics, till geochemistry and geological interpretations. Two datasets were generated: one with pixel size of 200 m x 200 m and another one with 50 m x 50 m. The coarser grid is a compromise between the resolutions of the original data, while the more accurate grid reveals the details in the geophysical data sets, but is over accurate for geochemical and gravity data. Overall dimensions of the study area are approximately 170 km in the East-West and 110 km in the North-South direction, yielding 508944 and 8146792 points for the 200 m and 50 m rasters, respectively. The datasets are illustrated in Figure 2.

3.1 Training data

Positive instances were extracted from the Geological survey of Finland's (GTK) database of mineral deposits and occurrences in Finland, and contain all the 27 gold deposits and other occurrences in CLGB that have been categorized as orogenic. Definition of the exact location of the occurrences is somewhat vague since they are not point-like. Usually orogenic gold deposits are no more than 100 meters in width, but can extend hundreds of meters along structures. Defining whether an occurrence is a single one or consists of multiple separate occurrences is subject to interpretation. Here, the deposits with undefined extents are represented as single pixels in the coarser grid, and extended using a linear smoothing filter to cover a square area of 32 pixels in the 50 m x 50 m grid.

Negative instances are generated by randomly sampling pixels in the study area. Random sampling for the negatives is justified, since the vast majority of the study area can be considered unprospective. Both versions of the data set contain a total of 1000 instances. In the first one there are 27 positive, and 973 negative instances. In the second data set, we randomly sample 16 pixels from the 32 pixels representing each deposit, leading to $27 \times 16 = 432$ positive, and 568 negative instances.

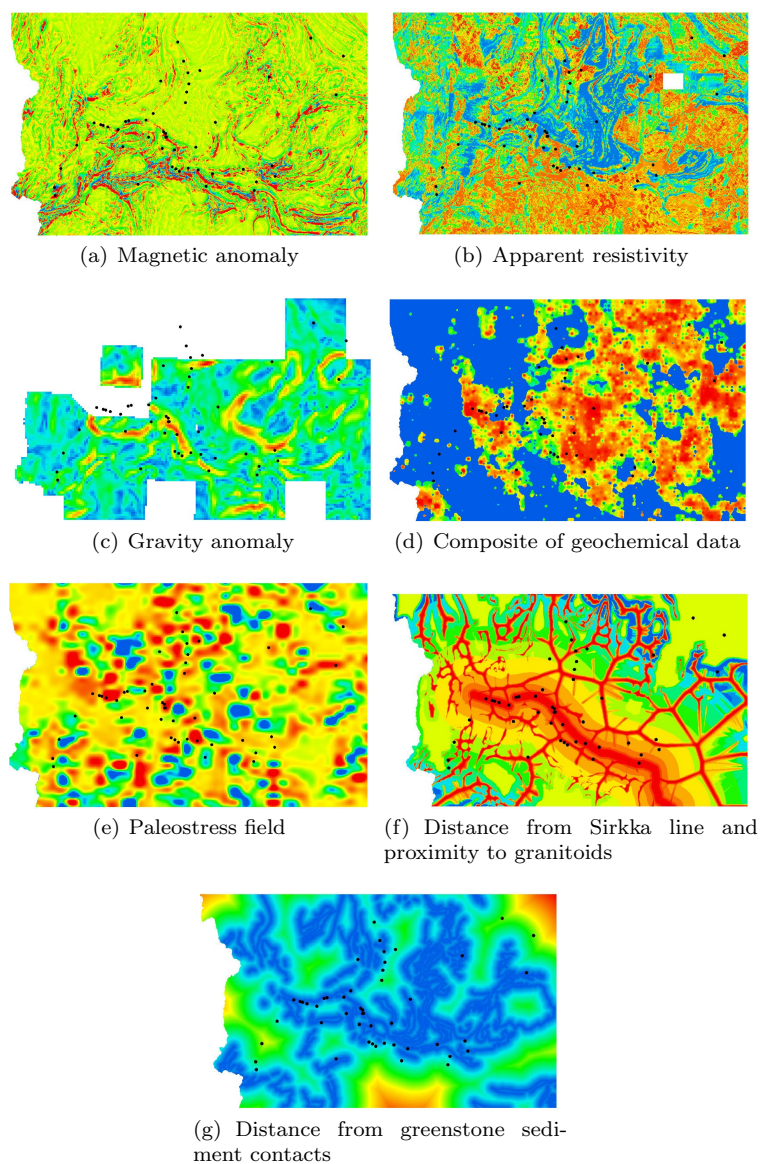


Fig. 2 Spatial representation of the evidence features in the study area. Coloring goes from blue (low values) to red (high values). Black dots represent the locations of the known orogenic gold occurrences.

3.2 Evidence features

The evidence feature set was the same as the one generated by Nykänen (2008) and consists of typical mineral exploration related geoscientific spatial data

that are derived from airborne geophysics (magnetic and electro-magnetic), ground geophysics (gravity), regional till geochemistry and a 1:200,000 scale digital geological map. Two evidence feature sets were generated with cell sizes of 0.04 km^2 and 0.0025 km^2 , while the resolution of the original measurements varies from 1 point/ 0.01 km^2 to 1 point/ 4 km^2 . Data preprocessing is briefly described below, while the geological basis and more detailed description of the preprocessing steps can be found in Nykänen (2008) and other references provided. Grid cell dimensions used by Nykänen (2008) were $250 \text{ m} \times 250 \text{ m}$, and resampling of cell values for the $200 \text{ m} \times 200 \text{ m}$ grid was done using the nearest neighbour method. Resampling for the $50 \text{ m} \times 50 \text{ m}$ grid was done using a linear smoothing filter. All the features are standardized to zero mean and unit variance.

Airborne magnetic and electromagnetic data were derived from the nationwide airborne geophysical measurements collected by GTK in 1973-2007 (Airo, 2005). Measurements were carried out with 200 m line spacing at a nominal $30\text{-}40 \text{ m}$ altitude using a fixed-wing aircraft, with vertical coplanar coils (coaxial until 1979) for the electromagnetics (Hautaniemi et al, 2005). Magnetic data were interpolated to grids with $50 \text{ m} \times 50 \text{ m}$ cell size, and deviation from the Definitive Geomagnetic Reference Field was computed following Korhonen (2005). Further, deviation of each pixel value from the median of pixel values within a radius of 4 km was calculated by Nykänen (2008). Electromagnetic response was interpreted as apparent resistivity and interpolated to grids with $50 \text{ m} \times 50 \text{ m}$ cell size following Suppala et al (2005), and further resampled to $250 \text{ m} \times 250 \text{ m}$ by Nykänen (2008).

The regional scale gravity map was derived from the ground-based gravity measurements collected by GTK and the Finnish Geospatial Research Institute (former Finnish Geodetic Institute) in 1990's (Kääriäinen and Mäkinen, 1997) with 1 point/ 1 km^2 . Gravity is the only evidence feature that does not cover the entire study area. Nykänen (2008) computed the horizontal gradient of Bouguer anomaly derived from the gravity measurements.

Geochemical data were derived from GTK's national geochemical survey of glacial tills, conducted in 1970's and 80's (Salminen and Tarvainen, 1995). 3-5 samples, taken at a density of 1 sample/ km^2 , were combined for analysis. The concentrations, thus, represent the average till concentration in an area of approximately 4 km^2 . Data for Au, As, Cu, Fe, Ni and Te were interpolated by Nykänen (2008) using the inverse distance weight method with the weight decreasing as the square of the distance. Since the grid cell size was much smaller than the sampling density, anomalous average concentrations appear spot-like near the locations associated to the combined sample. Nykänen (2008) further combined the different element concentration grids by setting conditions that Cu must always be elevated for a prospective area, at least one of As, Fe, Ni or Te must be elevated and presence of Au increases prospectivity.

From the digital 1:200,000 scale bedrock map of northern Finland (Lehtonen et al, 1998), three evidence features were derived. The first feature is the paleostress model computed following Holyland and Ojala (1997) by geomechanical interpretation at 1:100,000 scale using faults and lithological contacts

from the digital 1:200,000 scale bedrock maps and 1:100,000 scale geophysical maps. The second feature is the combination of proximity to granitoids in the Kittilä, Savukoski and Sodankylä Groups and distance to the Sirkka shear zone. The mean distance to granitoids within a 2500 m neighborhood is subtracted from the original proximity grid resulting in a grid which defines the midpoint between the granitoids within the greenstone belt, and this grid is combined with the proximity grid to the Sirkka Shear Zone. Values are discretized to 10 classes. The third feature derived from the bedrock map is the distance to contact zones between the greenstone belt lithological units and the overlying sedimentary units.

The geospatial data covers a 20 000 km² area centered on the Central Lapland Greenstone Belt (CLGB), located in the Northern Fennoscandian Shield. This area is a typical Paleoproterozoic greenstone belt composed of mafic to ultramafic volcanic successions and largely overlying sedimentary units surrounded and intruded by younger granitoids and mafic intrusions (Lehtonen et al, 1998). There has been noticeable amount of mineral exploration activity within the area during the recent years resulting more than 30 drill-defined gold occurrences and one currently operating gold mine. Majority of the gold occurrences within the CLGB are classified as orogenic gold deposits, as defined by Groves et al (1998). Indirect age constraints suggest two separate gold mineralization events within the Fennoscandian Shield at 1.9-1.86 and 1.85-1.79 Ga (Weihed et al, 2005). The assumption is that gold mineralization occurred during late orogenic events, enabling use of the current geometries on the bedrock map as a source of inputs for the spatial modeling because they approximate the geometries at the time of gold mineralization (Nykänen, 2008).

4 Experiments

In the experiments, we demonstrate the effects of both pooling and spatial bias, and how LPO-SCV allows correcting for both of them. Then we proceed to benchmark a number of different classifiers on the prospectivity mapping data sets. We consider three linear methods, support vector machine (SVM), logistic regression and ridge regression, as well as two non-linear ones, k-NN and random forest (Hastie et al, 2001; Breiman, 2001).

For ridge regression, we used the training and fast cross-validation algorithms implemented in RLScore library (Pahikkala and Airola, 2016). For the other methods, we used the scikit-learn library (Pedregosa et al, 2011), where the SVM implementation is based on the LIBLINEAR package (Fan et al, 2008). All 1000 instances of both datasets are used in the experiments. For the 200 m x 200 m data we run full LPO-SCV, whereas on the 50 m x 50 m resolution data, a random subsample of 50000 positive-negative pairs is used in LPO-SCV to speed up validation.

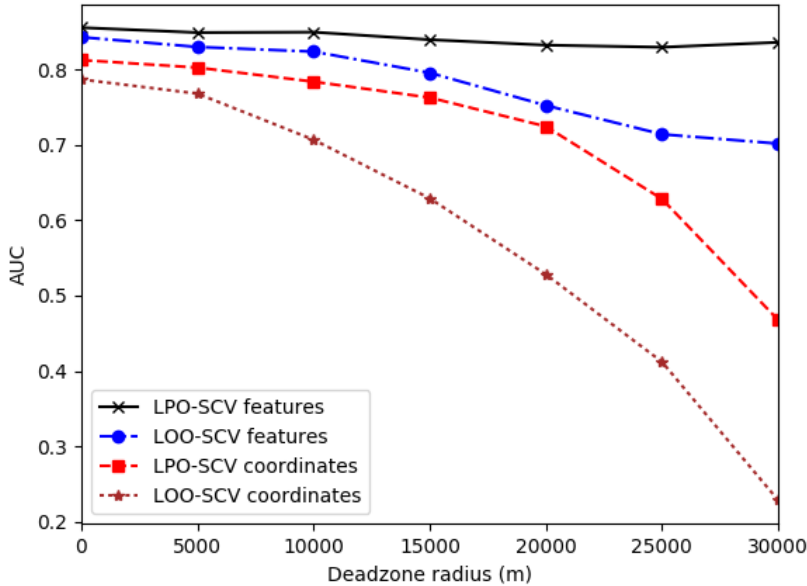


Fig. 3 Comparison of leave-pair-out (LPO-SCV) and leave-one-out (LOO-SCV) spatial cross-validation results on the version of the data with 200 m x 200m resolution and single pixel representing each deposit. k-NN ($k=250$) is trained with both using the regular feature set, and only x and y coordinates as features.

4.1 Pooling and spatial bias

In the first set of experiments, we compare a number of cross-validation approaches with k-NN classifier, in order to demonstrate both pooling and spatial biases. We used a large number of neighbors ($k=250$), as we noticed the method gave very poor results for small values of k . The experiments are performed on the data set with 200 m x 200 m resolution and single pixel per deposit.

The first classifier is trained normally on the evidence features. The second one is trained only on the x and y coordinates of the instances. The second classifier is used to demonstrate the spatial bias, as clearly it cannot learn to generalize to new areas. Based on the coordinates one can merely predict "gold deposits are found near other gold deposits".

We compare both LOO-SCV and LPO-SCV on deadzone radii ranging from 0 to 30000 meters. When $r_\delta = 0$, the methods are equivalent to ordinary LOO and LPO with no correction for spatial bias. In Figure 3 we can see a clear demonstration of both the pooling and spatial biases.

The LOO results are for both methods much worse than the LPO methods due to the pessimistic bias of LOO. The pooling bias increases as the deadzone grows larger; with 30 km deadzone the LPO-SCV result with model trained

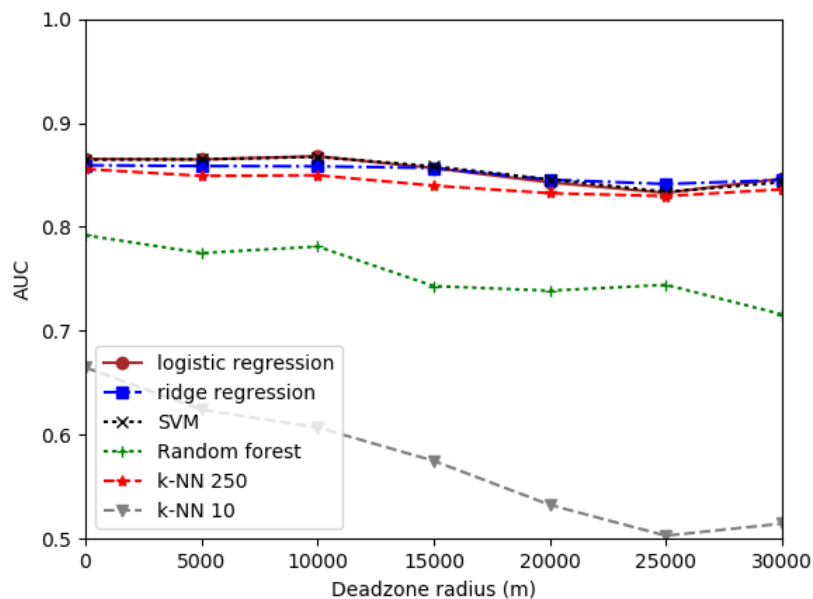


Fig. 4 Comparison of different classifiers on data with 200 m x 200 m resolution and one pixel per deposit.

on features is 0.84 AUC, whereas with LOO-SCV the result is only 0.70 AUC. Most noticeably, for the model trained on only the coordinates, the results even drop substantially below the 0.5 random level of AUC. These results are in line with the pessimistic bias of LOO for AUC estimation shown in earlier works of Airola et al (2009, 2011); Parker et al (2007); Smith et al (2014).

Spatial bias: For ordinary LPO and LOO with no deadzone ($r_\delta = 0$), x and y coordinates are enough to predict well (AUC 0.81). The predictions, however, drop to random level by $r_\delta = 30$ km, showing that based on only the coordinates the model cannot predict at all at 30 km distance and further from the training instances. In contrast, the model trained on the evidence features can generalize outside the training area.

LPO-SCV eliminates both sources of bias. On one hand, it eliminates the substantial pessimistic pooling bias that can be seen in the LOO-SCV results. At the same time, it shows that whereas the model trained on the features can generalize outside the immediate surrounding area of training data, the coordinate based models cannot.

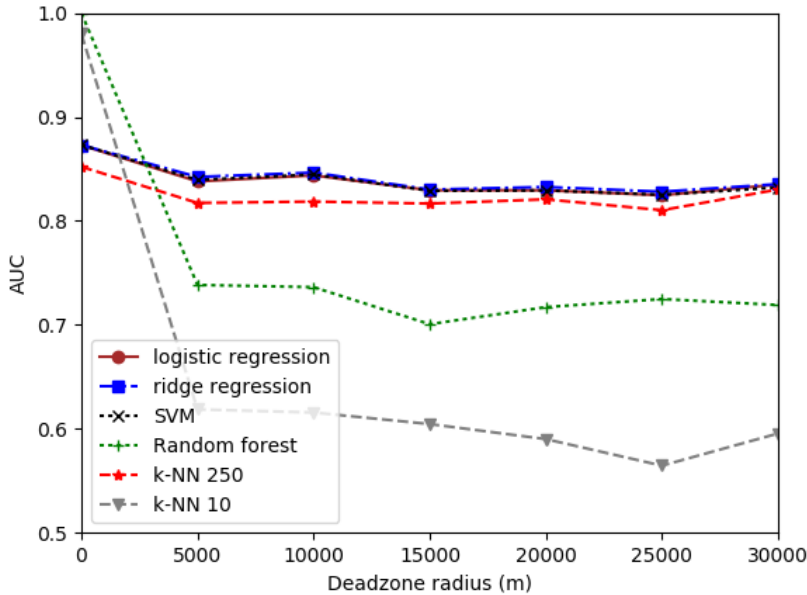


Fig. 5 Comparison of different classifiers on data with 50 m x 50 m resolution and sixteen pixels per deposit. Results for k-NN 10 and random forest are highly overoptimistic when deadzone is not used ($r_\delta = 0$).

4.2 Classifier comparison

We tested five different classification methods on the data set, using LPO-SCV. For SVM, logistic regression and ridge regression we present results for regularization parameter 1, as the results for a large range of parameter values were very similar. For random forest, the results are presented for 100 trees, as little improvement was observed after increasing the number of trees beyond this point. For k-NN we present the results both for $k=10$ and $k=250$, as the method behaved very differently depending on whether the number of neighbors was small or large. For SVM and logistic regression we used balancing to weight both classes equally, for the ridge regression and k-NN implementations such option was not available, for random forests balancing proved harmful and was not used.

The results are presented in Figure 4 for the data with 200 m x 200 m resolution and single pixel per deposit, and in Figure 5 for the data with 50 m x 50 m resolution and sixteen pixels per deposit.

The major difference between these two experiments is, how k-NN with $k=10$, and random forest behave on $r_\delta = 0$, where no deadzone correction is done (compare Figure 4 and 5). On the data set with a single pixel per deposit, the AUC of k-NN is 0.66, and that of random forest 0.79. On the data

with 16 pixels, k-NN AUC is 0.98, and random forest AUC 1.00. Thus on one of the data sets the two methods appear to work poorly, while on another it would seem that they can classify the data perfectly. The second result is a clear example of spatial bias. On each round of CV, the methods overfit to the 15 deposit pixels left in the training set, and can thus predict the left-out pixel. When deadzone radius is increased, the effect disappears and the poor ability of the classifiers to predict beyond their immediate neighborhood is revealed. This effect is not nearly as strong for the linear methods, as they are not expressive enough to overfit to the data as much as the non-linear k-NN and random forest models.

Otherwise the behavior of the methods is similar on the two versions of the data set. It can be seen that the linear methods (SVM, logistic and ridge regression) outperform the non-linear ones. Their AUC starts around 0.87, and decreases to 0.85 AUC as deadzone radius grows. There are no substantial differences between the performances of these three methods. k-NN 250 results are also very close to those of the linear classifiers with AUCs ranging from 0.86 to 0.84 on the single pixel data. The Random forest works poorly, with AUC always below 0.8.

Surprised by the poor performance of Random forest, we also performed limited experiments to see whether by further parameter tuning, or by using other types of tree-based ensemble methods such as the Extremely randomized trees (Geurts et al, 2006), results would improve. We did not find this to be the case. We also tested nonlinear kernel ridge regression (Evgeniou et al, 2000) using the RBF kernels of various widths. This did not lead to improvements over the linear ridge regression, but resulted in substantial increase in running time.

5 Discussion

The results demonstrate the clear need for spatial cross-validation of spatial prediction models, such as MPM classifiers. Due to small number of positive instances available in many applications, CV is crucial for validating the models. We show that if the spatial dependencies are not taken into account, one can obtain high AUCs even with classifiers that completely fail in generalizing outside the training area.

The data resolution and using multiple pixels versus using a single pixel to represent the deposits did not affect much the results for the best performing methods, when deadzone correction was properly done. However, when using several pixels to represent a deposit together with non-linear classifiers, we obtained very biased results if deadzone correction is not used.

The method comparison showed that simple linear models worked well on the MPM prediction problem. Whether the model was fitted by minimizing the logistic, least-squares (ridge), or hinge (SVM) loss did not affect the results much. The result is likely due to the small sample size, as there are only 27 positive instances of gold mineralization available in the data set. More

complex models are likely to overfit to the noise in the data, rather than discover patterns that would improve the predictions beyond what the linear model already captures. This could also be seen in the k-NN results, where averaging over a very large number of neighbors (k=250) provided the best results, whereas more complex local models based on a smaller number of neighbors (k=10) did not yield high AUC when properly validated.

In earlier work, Nykänen (2008) has shown 0.99 AUC results for both logistic regression and radial basis functional link nets on orogenic gold MPM data from the same study area. Our results are lower, though not directly comparable due to differences in data processing and experimental setup in model validation. Still, the different outcomes demonstrate the high degree to which the results depend on the chosen model validation strategy. These choices can often have much larger effect on results than the chosen classifiers. Thus we encourage researchers dealing with spatial data to provide also comprehensive spatial CV evaluations of their models, in order to establish how well they can predict at different distances from training data. This approach provides additional insights about the characteristics of the data, that the classical model validation methods are not able to provide.

6 Conclusion

In this work, we considered the problem of evaluating the AUC of classifiers on spatial data. Standard CV methods that have been developed for i.i.d. data suffer from two sources of bias, the pooling and spatial biases. In our experiments on MPM data we demonstrated the dangers of ignoring these biases, as one can obtain incorrect AUC values ranging from much worse than random to perfect with existing CV methods. We introduced the novel LPO-SCV method, that allows correcting for both the pooling and spatial biases inherent in classical CV methods. We demonstrate experimentally how the method allows reducing these biases, and benchmarked a number of MPM classifiers showing the advantages of simple linear models. While we have considered only one MPM classification problem, the introduced evaluation approach is general and could be applied in a wide range of different types of spatial classification or ranking problems.

References

- Abedi M, Norouzi GH, Bahroudi A (2012) Support vector machine for multi-classification of mineral prospectivity areas. *Computers & Geosciences* 46(Supplement C):272 – 283
- Airo ML (2005) Aerogeophysics in Finland 1972-2004. Special paper, Geological survey of Finland 39
- Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T (2009) A comparison of AUC estimators in small-sample studies. In: Džeroski S, Geurts P,

- Rousu J (eds) Proceedings of the third International Workshop on Machine Learning in Systems Biology, PMLR, Ljubljana, Slovenia, Proceedings of Machine Learning Research, vol 8, pp 3–13
- Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T (2011) An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis* 55(4):1828–1844, DOI 10.1016/j.csda.2010.11.018
- Bamber D (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology* 12(4):387–415
- Bonham-Garter G (1994) Geographic information systems for geoscientists – modelling with GIS, *Computer methods in geosciences*, vol 13. Pergamon Press, Oxford, UK
- Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30(7):1145–1159
- Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32
- Brown WM, Gedeon TD, Groves DI (2003) Use of noise to augment training data: A neural network method of mineral-potential mapping in regions of limited known deposit examples. *Natural Resources Research* 12(2):141–152
- Carranza E (2008) Geochemical Anomaly and Mineral Prospectivity Mapping in GIS, *Handbook of Exploration and Environmental Geochemistry*, vol 11. Elsevier Science
- Cressie N (2015) *Geostatistics*. John Wiley & Sons, Inc.
- Elkan C, Noto K (2008) Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '08, pp 213–220
- Evgeniou T, Pontil M, Poggio T (2000) Regularization networks and support vector machines. *Advances in Computational Mathematics* 13:1–50
- Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874
- Fawcett T (2006) An introduction to ROC analysis. *Pattern recognition letters* 27(8):861–874
- Forman G, Scholz M (2010) Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explorations* 12(1):49–57
- Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. *Machine Learning* 63(1):3–42
- Groves D, Goldfarb R, Gebre-Mariam M, Hagemann S, Robert F (1998) Orogenic gold deposits: A proposed classification in the context of their crustal distribution and relationship to other gold deposit types. *Ore Geology Reviews* 13(1):7 – 27
- Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143(1):29–36

- Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning*. Springer Series in Statistics, Springer New York Inc., New York, NY, USA
- Hautaniemi H, Kurimo M, Multala J, Leväniemi H, Vironmäki J (2005) The "three in one" aerogeophysical concept of GTK in 2004. Special paper, Geological survey of Finland 39:21–74
- Holyland PW, Ojala J (1997) Computer aided structural targeting: two and three dimensional stress mapping. *Australian Journal of Earth Sciences* 44:421–432
- Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 17(3):299–310
- Jain S, White M, Radivojac P (2017) Recovering true classifier performance in positive-unlabeled learning. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp 2066–2072
- Kääriäinen J, Mäkinen J (1997) Airborne magnetic method: special features and review on applications. Special paper, Geological survey of Finland 39:77–102
- Korhonen JV (2005) Airborne magnetic method: special features and review on applications. Special paper, Geological survey of Finland 39:77–102
- Le Rest K, Pinaud D, Monestiez P, Chadoeuf J, Bretagnolle V (2014) Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography* 23(7):811–820, DOI 10.1111/geb.12161
- Lehtonen M, Airo M, Eilu P, Hanski E, Kortelainen V, Lanne E (1998) The stratigraphy, petrology and geochemistry of the Kittilä greenstone area, northern Finland. In: *Report of Investigation*, vol 140, Geological Survey of Finland, pp 140–144
- Longley P, Goodchild M, Maguire D, Rhind D (2005) *Geographic Information Systems and Science*
- Nykänen V (2008) Radial basis functional link nets used as a prospectivity mapping tool for orogenic gold deposits within the central Lapland greenstone belt, northern Fennoscandian shield. *Natural Resources Research* 17(1):29–48, DOI 10.1007/s11053-008-9062-0
- Nykänen V, Lahti I, Niiranen T, Korhonen K (2015) Receiver operating characteristics (roc) as validation tool for prospectivity models - a magmatic Ni-Cu case study from the central lapland greenstone belt, northern finland. *Ore Geology Reviews* 71(Supplement C):853 – 860
- Pahikkala T, Airo A (2016) RLScore: Regularized least-squares learners. *Journal of Machine Learning Research* 17(221):1–5
- Pahikkala T, Suominen H, Boberg J (2012) Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Machine Learning* 87(3):381–407, DOI 10.1007/s10994-012-5287-6
- Parker BJ, Gunter S, Bedo J (2007) Stratification bias in low signal microarray studies. *BMC Bioinformatics* 8:326

- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
- Pohjankukka J, Nevalainen P, Pahikkala T, Hyvönen E, Hänninen P, Sutinen R, Ala-Ilomäki J, Heikkonen J (2014) Predicting water permeability of the soil based on open data. In: Iliadis L, Maglogiannis I, Papadopoulos H (eds) *Proceedings of the 10th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2014)*, Springer, IFIP Advances in Information and Communication Technology, vol 436, pp 436–446
- Pohjankukka J, Pahikkala T, Nevalainen P, Heikkonen J (2017) Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science* 31(10):2001–2019, DOI 10.1080/13658816.2017.1346255
- Rigol-Sanchez JP, Chica-Olmo M, Abarca-Hernandez F (2003) Artificial neural networks as a tool for mineral potential mapping with gis. *International Journal of Remote Sensing* 24(5):1151–1156
- Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M (2015) Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews* 71(Supplement C):804 – 818
- Salminen R, Tarvainen T (1995) Geochemical mapping and databases in Finland. *Journal of geochemical exploration* 55:321–327
- Smith GC, Seaman SR, Wood AM, Royston P, White IR (2014) Correcting for optimistic prediction in small data sets. *American journal of epidemiology* 180(3):318–324
- Suppala I, Oksama M, Hongisto H (2005) GTK airborne EM system: characteristics and interpretation guidelines. Special paper, Geological survey of Finland 39:103–118
- Tobler WR (1970) A computer movie simulating urban growth in the detroit region. *Economic Geography* 46(sup1):234–240
- Weihed P, Arndt N, Billström K, Duchesne JC, Eilu P, Martinsson O, Papunen H, Lahtinen R (2005) Precambrian geodynamics and ore formation: The fennoscandian shield. *Ore Geology Reviews* 27(1):273 – 322, special Issue on Geodynamics and Ore Deposit Evolution in Europe

Turku Centre for Computer Science

TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspñäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4 -Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen**, On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Sääntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Methods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performance Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyötiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases
202. **Fredrik Abbors**, Model-Based Testing of Software Systems: Functionality and Performance
203. **Inna Pereverzeva**, Formal Development of Resilient Distributed Systems
204. **Mikhail Barash**, Defining Contexts in Context-Free Grammars
205. **Sepinoud Azimi**, Computational Models for and from Biology: Simple Gene Assembly and Reaction Systems
206. **Petter Sandvik**, Formal Modelling for Digital Media Distribution

207. **Jongyun Moon**, Hydrogen Sensor Application of Anodic Titanium Oxide Nanostructures
208. **Simon Holmbacka**, Energy Aware Software for Many-Core Systems
209. **Charalampos Zinoviadis**, Hierarchy and Expansiveness in Two-Dimensional Subshifts of Finite Type
210. **Mika Murtojärvi**, Efficient Algorithms for Coastal Geographic Problems
211. **Sami Mäkelä**, Cohesion Metrics for Improving Software Quality
212. **Eyal Eshet**, Examining Human-Centered Design Practice in the Mobile Apps Era
213. **Jetro Vesti**, Rich Words and Balanced Words
214. **Jarkko Peltomäki**, Privileged Words and Sturmian Words
215. **Fahimeh Farahnakian**, Energy and Performance Management of Virtual Machines: Provisioning, Placement and Consolidation
216. **Diana-Elena Gratie**, Refinement of Biomodels Using Petri Nets
217. **Harri Merisaari**, Algorithmic Analysis Techniques for Molecular Imaging
218. **Stefan Grönroos**, Efficient and Low-Cost Software Defined Radio on Commodity Hardware
219. **Noora Nieminen**, Garbling Schemes and Applications
220. **Ville Taajamaa**, O-CDIO: Engineering Education Framework with Embedded Design Thinking Methods
221. **Johannes Holvitie**, Technical Debt in Software Development – Examining Premises and Overcoming Implementation for Efficient Management
222. **Tewodros Deneke**, Proactive Management of Video Transcoding Services
223. **Kashif Javed**, Model-Driven Development and Verification of Fault Tolerant Systems
224. **Pekka Naula**, Sparse Predictive Modeling – A Cost-Effective Perspective
225. **Antti Hakkala**, On Security and Privacy for Networked Information Society – Observations and Solutions for Security Engineering and Trust Building in Advanced Societal Processes
226. **Anne-Maarit Majanoja**, Selective Outsourcing in Global IT Services – Operational Level Challenges and Opportunities
227. **Samuel Rönqvist**, Knowledge-Lean Text Mining
228. **Mohammad-Hashem Hahgbayan**, Energy-Efficient and Reliable Computing in Dark Silicon Era
229. **Charmi Panchal**, Qualitative Methods for Modeling Biochemical Systems and Datasets: The Logicome and the Reaction Systems Approaches
230. **Erkki Kaila**, Utilizing Educational Technology in Computer Science and Programming Courses: Theory and Practice
231. **Fredrik Robertsén**, The Lattice Boltzmann Method, a Petaflop and Beyond
232. **Jonne Pohjankukka**, Machine Learning Approaches for Natural Resource Data

TURKU CENTRE *for* COMPUTER SCIENCE

<http://www.tucs.fi>

tucs@abo.fi



University of Turku

Faculty of Science and Engineering

- Department of Future Technologies
- Department of Mathematics and Statistics

Turku School of Economics

- Institute of Information Systems Science



Åbo Akademi University

Faculty of Science and Engineering

- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics

- Information Systems

ISBN 978-952-12-3710-2

ISSN 1239-1883

Jonne Pohjankukka

Jonne Pohjankukka

Jonne Pohjankukka

Machine Learning Approaches for Natural Resource Data

Machine Learning Approaches for Natural Resource Data

Machine Learning Approaches for Natural Resource Data