Advanced Data Mining and Machine Learning Algorithms for Integrated Computer-Based Analyses of Big Environmental Databases

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Abduljabbar Asadi

Geboren am 04. 04. 1985, in Sanandaj, Iran

Tübingen

2017

Tag der mündlichen Qualifikation: 25 Sept 2107

Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Peter Dietrich
2. Berichterstatter:	Prof. Dr. Erwin Appel

Dedication

Foremost, I am highly grateful to my God for his blessing that continues to flow into my life, and because of the understanding and knowledge which gives me I am able to solve all challenges in my life.

I dedicate this work to my dear wife; Soniya Hatami who has tolerated hardship and encouraged me all the way and whose encouragement has made sure that I give it all it takes to finish which I have started. You were the most important reason for this success. I am truly thankful for having you in my life. I love you so much. I also dedicated this work to the childhood memories of my son; Ermiya, because he spent his childhood with me during doing this thesis, I promise that I have enough time for you now.

I dedicated this work to the endured hands of my father Taofigh and my mama Khanoom, who are waiting for me and always loved me unconditionally. Thank you so much, I love you unconditionally, and will never forget you. I would like to extend my thanks to my father-in-law Eskandar and mother-in-law Golbakhi; for their encouragement and involvement in my efforts, and for their love. Thank you so much.

I would like to express my deepest gratitude to my supervisor Professors Peter Dieterich and my project supervisor Dr. Hendrik Paasche for their unwavering support, collegiality, and mentorship throughout this project.

Abstract

Knowledge of the spatial distribution of geotechnical, hydrological, petroleum, and environmental parameters in the subsurface is essential in environmental earth sciences. Latest developments in engineering geophysics and remote sensing data acquisition provide a large set of techniques for non-invasive and *in situ* data recording for high-resolution ground probing. We developed different strategies based on knowledge discovery for analyzing environmental earth science databases. One important type of these databases is geophysical tomography, which offers 2D or 3D valuable and unique information about the internal composition of the ground. Based on different data mining and machine learning (i.e., feature extraction, Artificial Neural Networks, etc.) techniques we developed a data analysis strategy based on artificial neural networks (ANNs) allowing for 2D or 3D probabilistic prediction of sparsely measured earth properties constrained by geophysical tomography fully accounting for tomographic reconstruction ambiguity. Furthermore, we try to take the uncertainty or variability of the input data into account for proving the results of this method. Additionally, we have evaluated whether the training performance of the prediction model of ANNs can be used to rank geophysical tomograms. Such prediction model can contribute to solving hydrological, petroleum, or engineering exploration tasks in a data-driven manner. Another important issue is mapping the earth which is a fundamental prerequisite required to address various environmental and economic issues, such as mining target identification, soil conservation, or ecosystem management. The datasets for mapping are either technical or subjective. We show the first attempts towards integrating technical and subjective spatial datasets in an automated and rapid manner based on different data mining algorithms (i.e., graph analysis, boundary detection, clustering, etc.). Such method aims to produce a crisp or fuzzy classified map outlining dominant structures in the database optimally consistent with all available information.

Zusammenfassung

Einsicht in die räumliche Verteilung geotechnischer und hydrologischer Untergrundeigenschaften sowie von Reservoir- und Umweltparametern sind grundlegend für geowissenschaftliche Forschungen. Entwicklungen in den Bereichen geophysikalische Erkundung sowie Fernerkundung resultieren in der Verfügbarkeit verschiedenster Verfahren für die nichtinvasive, räumlich kontinuierliche Datenerfassung im Rahmen hochauflösender Messverfahren. In dieser Arbeit habe ich verschiedene Verfahren für die Analyse erdwissenschaftlicher Datenbasen entwickelt auf der Basis von Wissenserschließungsverfahren. Eine wichtige Datenbasis stellt geophysikalische Tomographie dar, die als einziges geowissenschaftliches Erkundungsverfahren 2D und 3D Abbilder des Untergrunds liefern kann. Mittels unterschiedlicher Verfahren aus den Bereichen intelligente Datenanalyse und maschinelles Lernen (z.B. Merkmalsextraktion, künstliche neuronale Netzwerke, etc.) habe ich ein Verfahren zur Datenanalyse mittels künstlicher neuronaler Netzwerke entwickelt, das die räumlich kontinuierliche 2D oder 3D Vorhersage von lediglich an wenigen Punkten gemessenen Untergrundeigenschaften im Rahmen von Wahrscheinlichkeitsaussagen ermöglicht. Das Vorhersageverfahren basiert auf geophysikalischer Tomographie und berücksichtigt die Mehrdeutigkeit der tomographischen Bildgebung. Außerdem wird auch die Messunsicherheit bei der Erfassung der Untergrundeigenschaften an wenigen Punkten in der Vorhersage berücksichtigt. Des Weiteren habe ich untersucht, ob aus den Trainingsergebnissen künstlicher neuronaler Netzwerke bei der Vorhersage auch Aussagen über die Realitätsnähe mathematisch gleichwertiger Lösungen der geophysikalischen tomographischen Bildgebung abgeleitet werden können. Vorhersageverfahren wie das von mir vorgeschlagene, können maßgeblich zur verbesserten Lösung hydrologischer und geotechnischer Fragestellungen beitragen.

Ein weiteres wichtiges Problem ist die Kartierung der Erdoberfläche, die von grundlegender Bedeutung für die Bearbeitung verschiedener ökonomischer und ökologischer Fragestellungen ist, wie z.B., die Identifizierung von Lagerstätten, den Schutz von Böden, oder Ökosystemmanagement. Kartierungsdaten resultieren entweder aus technischen (objektiven) Messungen oder visuellen (subjektiven) Untersuchungen durch erfahrene Experten. Im Rahmen dieser Arbeit zeige ich erste Entwicklungen hin zu einer automatisierten und schnellen Integration technischer und visueller (subjektiver) Daten auf der Basis unterschiedlicher intelligenter Datenanalyseverfahren (z.B., Graphenanalyse, automatische Konturerfassung, Clusteranalyse, etc.). Mit solchem Verfahren sollen hart oder weich klassifizierte Karten erstellt werden, die das Untersuchungsgebiet optimal segmentieren um höchstmögliche Konformität mit allen verfügbaren Daten zu erzielen.

Contents

Abstract i

Zusammenfassung ii

Contents vi

List of Figures xii

List of Tables xiii

List of Symbols xiv

List of Abbreviations xv

Chapter 1 Introduction 1

1.1 Knowledge Discovery in Databases (KDD) 1

1.2 Uncertainty and Error in Data 5

1.3 Subjective and Technical Data 7

1.4 Human in the Loop (HTL) 8

1.5 KDD in Earth Sicences 11

1.6 Objectives of This Thesis 12

Chapter 2 2D Probabilistic Prediction of Sparsely Measured Earth Properties Constrained by Geophysical Imaging Fully Accounting for Tomographic Reconstruction Ambiguity 15

2.1 Abstract 15

2.2 Introduction 17

2.3 Methodology 20

2.3.1 Geophysical Tomography as Feature Construction Problem 20

2.3.2 Feature Selection by Prediction 22

2.3.3 Processing Flow 24

2.4 Case Study 26

2.4.1 Tomographic Data Simulation 26

2.4.2 Feature Construction By Tomographic Reconstruction 28

2.4.3 Generation of Sparse Exploration Target Parameters 31

2.4.4 Setting up the Artificial Neural Network 32

2.5 Results and Discussion 34

2.5.1 Probabilistic Prediction of 2D Porosity Distributions 34

2.5.2 Ranking of Tomograms 38
2.6 Conclusions 41
Chapter 3 Spatially Continuous Probabilistic Prediction of Sparsely Measured Ground Properties Constrained by ill-posed Tomographic Imaging Considering Data Uncertainty and Resolution 43
3.1 Abstract 43
3.2 Introduction 45
3.3 Methodology 48
3.3.1 Artificial Neural Networks (ANNs) 48
3.4 Processing Flow 51
3.5 The Database 52
3.5.1 The Field Site 52
3.6 Data Acquisition 53
3.6.1 Tomography 53
3.6.2 Sparse Logging Data 54
3.7 Processing 55
3.7.1 Tomography 55
3.7.2 Tomographic Uncertainty 57
3.7.3 Logging Data Uncertainty 58
3.8 Setting up the ANNs 59
3.8.1 Combination of Tomograms for ANNs 59
3.8.2 Training Without Error Incorporation 60
3.8.3 Training With Error Incorporation 60
3.8.4 Training With Separate Logging Data 61
3.8.5 Training Accounting for Resolution Difference 61
3.8.6 Selecting the Number of Neurons in the Hidden Layer 62
3.9 Results and Discussion 63
3.9.1 Prediction Results of ANNs 63
3.9.2 Training Without Error Incorporation 63
3.9.3 Training With Error Incorporation 64
3.9.4 Training With Separate Logging Data 66
3.9.5 Taking Resolution Differences into Account 67
3.9.6 Comparative Discussion 68

3.9.7 Transferability and Outreach of Results 69
3.10 Conclusions 70
Chapter 4 Conceptual Developments for Clustering Mapped Data Emanating From Technical Sensors and Subjective Insights of Human Experts 72
4.1 Abstract 72
4.2 Introduction 74
4.3 Methodology 76
4.3.1 The K-means Clustering 76
4.3.2 Graph Theory and Shortest Path 78
4.3.3 Boundary Detection 78
4.3.4 Sampling Strategy 80
4.3.5 Processing Flow 81
4.4 The Datasets 85
4.4.1 The Synthetic Datasets 85
4.4.2 The Field Datasets 86
4.5 Experiments and Results 88
4.5.1 Application to the Synthetic Dataset 88
4.5.2 Application to the Field Dataset 91
4.6 Conclusions 96
Chapter 5 Summary and outlook 97
5.1 Data Mining Techniques Add Value to Geophysical Tomography and Logging Data 97
5.1.1 Taking Data Uncertainty into Account in the Machine Learning Prediction Part 99
5.1.2 Choosing Optimum Parameters for Artificial Neural Networks 100
5.1.3 Testing with Different Tomograms and Target Parameters 101
5.2 Human in the Loop for Mapping, Integration, and Segmentation of Geophysical Datasets 102
5.2.1 Further Testing the Idea of the Human in the Loop for Integration and Segmentation of Geoscientific Datasets 103
References 105
Appendix A Towards Probabilistic Prediction of Soil Moisture in the Schäfertal Catchment 113
A.1 Introduction 113
A.2. Methodology 113
A.3 Processing 114

A.4 Conclusion 116
Appendix B A New Methodology for Prediction of 2D Distributions of Sparsely Measured Logging Data under Full Consideration of Tomographic Model Generation Ambiguity 118
B.1 Abstract 118
B.2 Introduction 118
B.3 Artificial Neural Networks (ANN) 119
B.4 Methodology 120
B.5 Application to a Field Dataset 123
B.6 Conclusions 124
Appendix C 2D probabilistic prediction of sparsely measured geotechnical parameters constrained by tomographic ambiguity and measurements errors 125
C.1 Abstract 125
C.2 Introduction 125
C.3 Artificial Neural Networks (ANN) 127
C.4 The Processing Flow 128
C.5 Results 129
C.6 Conclusions 131
Appendix D Predicting Porosity According to Ensembles of Collocated Radar and Seismic Tomographic Models with Artificial Neural Networks 132
D.1 Abstract 132
Appendix E Conceptual Developments for Clustering Mapped Data Emanating from Technical Sensors and Subjective Insights of Human Experts 134
E.1 Abstract 134
Appendix F Probabilistic Integration of Tomograms and Logging Data Accounting for Tomographic Ambiguity and Logging Data Errors 136
F.1 Abstract 136
Appendix G Incorporating Hyperspectral Datasets in the Integration Strategy Introduced in Chapter 4 138
G.1 Process 138
Appendix H Histogram Normalization 142
H. 1 Process and Results 142
Curriculum Vitae 145

List of Figures

Figure 1-1: The structure of knowledge discovery in databases (KDD). It is based on the three important steps: preprocessing, data mining, and information evaluation by the end user. During these three steps the effective information will be extracted from big data bases by KDD. 3

Figure 1-2: Four combinations of accuracy and precision (Haibo, 2014).. 7

Figure 1-3: Four different interactions between KDD and humans. (a), (b), (c), and (d) show unsupervised, supervised, semi-supervised, and human in the loop approaches, respectively (Holzinger, 2016)... 10

Figure 2-1: Schematic sketch of a geophysical cross-borehole tomographic experiment with sources and receivers in the left and right boreholes.....21

Figure 2-2: Flowchart of the processing workflow for probabilistic prediction of spatially continuous models of sparsely measured target parameters and ranking non-sparse tomograms achieved from inversion or feature construction algorithms... 25

Figure 2-3: Original synthetic models representing ground truth. (a) Layered ground assumed. (b) Porosity variability in the ground. (c) Radar velocity and (d) seismic velocity are obtained from traditional deterministic petrophysical transfer functions used to convert porosity into physical parameters. (e) and (f) scatter plots of the models given in (b) and (c), and (b) and (d), respectively.....27

Figure 2-4: Traveltimes of the (a) radar and (b) seismic tomographic datasets... 29

Figure 2-5: 30 tomographic reconstructions of radar wave propagation velocity distributions achieved by fully non-linear self-organizing inversion (Paasche, 2015). The 30 tomograms are achieved by independent inversion runs and fit the underling tomographic dataset equally well.. 30

Figure 2-6: The same as in Figure 2-5 but for the seismic tomograms... 31

Figure 2-7: Sparse porosity borehole logging data acquired in the boreholes at the (a) left (x=0 m) and (b) right (x=10 m) model edges (see Figure 2-3). Original porosity represents the true information of the ground. Logging porosity represents the modelled response of a realistic borehole porosity logging probe... 32

Figure 2-8: MSE from ANN training for all combinations of spatially continuous tomograms. (a)-(e) represent the results of using ANNs with 3,5,10,20 and 50 neurons, respectively.....34

Figure 2-9: Regression results of the training procedure of the ANN when using R30 and S30 tomograms as input and the (a) original porosity (RSOP) and (b) logging porosity (RSLP) as output information. Note, only the radar and seismic velocity information of the left and right model edges has been used for training.....34

Figure 2-10: Prediction results of spatially continuous 2D porosity models based on the sparse logging data and 900 combinations of spatially continuous tomograms (Figure 2-5 and 6). (a)

Relative frequency of porosity prediction from ANN models trained with original porosity. (b) The same as (a) but overlain by minimum and maximum range of true porosity of the ground (see Figure 2-3) shown by dashed black lines. (c) and (d) are analogue to (a) and (b) but for logging porosity instead of original porosity. Predicted porosities outside the displayed range are accumulated at the bins with lowest and highest porosities. They correspond to the ANN models trained with remaining high MSE... 36

Figure 2-11: Ranking of tomograms (Figure 2-5 and 2-6) according to their relationships with reality. RSOP rank radar and seismic models according to their combination with original porosity for training the ANN models. Likewise RSLP rank radar and seismic velocity tomograms according to their combination with logging porosity for training the ANN models. This ranking has been outcome from ANN with three neurons in the hidden layer. 38

Figure 2-12: Summed squared differences of 30 tomographic radar (R) and seismic (S) velocity tomograms from the true radar and seismic velocity models (Figures 2-3c, and 2-3d). The ordering of the model number (abscissa) corresponds to those proposed by the ANNs trained with three neurons in the hidden layer (Figure 2-11). The black circles illustrate differences at the left and right edges (logging positions) of the tomograms. The gray rectangles illustrate the difference of the entire 2D area. (a) and (b) correspond to the ANN trained with original porosity. (c) and (d) correspond to the ANN trained with logging porosity.. 39

Figure 3-1: Structure of artificial neural networks (ANNs). ANNs are consisting of three interconnected layers. The input layer prepares data for feeding the ANN. The operation of hidden layer is based on sets of input information, weights of inputs w and bias parameter b. Neurons in this layer form a feedforward network with sigmoid formation. The operation of the output layer has been determined by the hidden layer and is connected to the results of the ANN training step... 48

Figure 3-2: Processing workflow to probabilistically predicting spatially continuous models for sparsely measured target parameters and geophysical tomograms achieved from fully non-linear inversion. Based on the training strategy, v determines the number of prediction models resulted from ANNs... 51

Figure 3-3: Target parameter logging data acquired by direct push technology at x=2.75 m and x=8.0 m for (a) tip resistance, (b) sleeve friction, and (c) dielectric permittivity.. 54

Figure 3-4: 30 tomographic reconstructions of radar-wave propagation velocity achieved by fully non-linear (global-search) inversion. The 30 tomograms are achieved by independent inversion runs and fit the underling tomographic dataset equally well... 55

Figure 3- 5: The same as in Figure 3- 4 but for P-wave velocity... 56

Figure 3- 6: The same as in Figure 3- 4 but for S-wave velocity... 57

Figure 3-7: Tip resistance error calculation based on a low pass zero-phase digital filter. The black line in Figure 3-7a determines the measured tip resistance; the gray line shows the filtered log. In Figure 3-7b the relative difference between measured and filtered tip resistance are shown which is considered as data noise component. Figure 3-7c determines the slope based on the low pass zero-phase digital filter in Figure 3-7a. 59

Figure 3-8: MSE from ANN training for all combinations of spatially continuous tomograms with 15, 25, 35, 50, and 100 neurons in the hidden layer. Based on this comparison 50 neurons have been selected as optimal number of neurons.. 62

Figure 3-9: Prediction results of spatially continuous 2D target parameters based on the sparse logging data and 27,000 combinations of spatially continuous tomograms (Figure 3-4, 3-5 and 3-6) shown as histographic plot. 2D probabilistic prediction plots show (a) tip resistance, (b) sleeve friction, and (c) dielectric permittivity prediction. The dotted white lines show the measured logging data of target parameters (Figure 3) that are used for training the ANN. Red colors correspond to high relative frequencies. Blue colors correspond to low relative frequencies... 64

Figure 3-10: The same as in Figure 3-9, but trained considering tomographic and logging uncertainty when training the ANN... 65

Figure 3-11: The same as in Figure 3-10, but now training was individually performed for logs at x =2.75 m and x = 8.0 m.. 66

Figure 3 -12: The same as in Figure 3- 11 but, now repeatedly considering individual samples from the logging datasets per grid cell when training the ANN, rather than averaging over all samples corresponding to a grid cell... 67

Figure 3-13: Example of comparison of tip resistance predictions at x= 6.5m drawn from Figures 3-9, 3-10, 3-11, and 3-12. Trained (a) without error incorporation, (b) with error incorporation, (c) separate logging data, and (d) accounting for resolution difference between logs and tomograms.. 69

Figure 4-1: Different types of boundaries in an image (a) step (b) line (c) ramp (d) roof... 79

Figure 4-2: Processing workflow to integration and segmentation of subjective and objective datasets. After normalizing the datasets two processing branches will be followed in the workflow. In the right part the similarity only based on the objective data will be calculated. Simultaneously in the left part the boundary of the subjective and objective maps will be extracted. Then, the new information vector based on results of these two branches will be calculated. This new information vector will be considered as new dataset and will be the input for the clustering algorithm. 82

Figure 4-3: Synthetic datasets for a 30*30 2D domain with (a) subjective map, (b) and (c) objective technical maps. The subjective map shows structures mostly the same as in the technical maps, but with step boundaries and noise free. The technical maps show structures with different boundaries (i.e., step, line, roof, and ramp), noise, and anomalies from anthropogenic effects.. 85

Figure 4-4: Two subjective maps of the Schäfertal catchment created by (a) (Borchardt 1985; Ollesch et al. 2005) and (b) Landesamt für Geologie und Bergwesen Sachsen-Anhalt. They show the structures in this catchments based on observations recorded by scientists... 86

Figure 4-5: Four attributes shown as objective or technical maps; (a) elevation, (b) slope, (c) SAGA wetness index, and (d) annual potential incoming solar radiation derived form a 2-m digital elevation model for the Schäfertal catchment (Schröter et al., 2015). 87

Figure 4-6: 900*900 Gaussian similarity matrix for the data points in the synthetic dataset based on the absolute values in technical maps. The similarity of points is a value between 0 and 1.. 88

Figure 4-7: Boundary information for the maps in the synthetic datasets. (a) Extracted boundaries based on Canny boundary detection for the subjective map, (b) and (c) extracted boundaries based on the χ2 distances for technical maps, (d) calculated total boundary map of subjective and objective boundary maps... 89

Figure 4-8: 900*900 matrices resultant from Dijkstra algorithm (a) all shortest paths and (b) all path lengths for data points in the synthetic dataset... 90

Figure 4-9: 900*900 matrix as new information vector resultant from similarity, shortest paths, and path lengths matrices. The columns show 900 samples or data points and rows show 900 attributes or variable layers. This matrix carries the information about colors and boundary information of points in the map.....90

Figure 4-10: k-means clustering results for the synthetic dataset with two strategies. (a) The k-means clustering results based on the new information vector resultant from the introduced strategy in this chapter, (b) the results of clustering without considering the boundary information of the subjective map. 6 cluster are desired in this dataset... 91

Figure 4-11: Boundary information extracted from subjective maps in the Schäfertal catchment. (a) Canny's boundary detection results for the subjective map shown in Figure 4-4a. Problems with corners and junctions exist (see inlet), (b) and (c) Gaussian filtering for extracting bulky boundary subjective maps showed in Figure 4-4a and 4b, respectively....92

Figure 4-12: Boundary information for (a) TWI and (b) insolation maps of the Schäfertal catchment based on the χ2 distance. (c) Presents the total boundary map of subjective and objective maps for the Schäfertal catchment. 93

Figure 4-13: (a) The selected s=1000 sampling points using systematic sampling selection strategy. (b) The shortest path and (c) path length resultant from Dijkstra algorithm for the exemplary point selected at x=640.2 and y=5725 km to all other points in the map... 94

Figure 4-14: The results of clustering the maps of the Schäfertal catchment with two strategies,(a) the k-means clustering results on the new information vector resultant from the introduced strategy in this chapter, (b) presents the results of clustering without considering the boundary information (Schröter et al., 2015). 30 clusters are desired in this catchment, each color determine an independent cluster. 95

Figure 4-15: The same as in Figure 4-14a, but now with sampling size reduced to (a) 500 and (b) 250 samples.. 95

Figure A-1: Land use map created by Schröter et al., (2015). Arable land is depicted by light gray, grassland by dark gray colors... 114

Figure A-2: Four topographic attributes as objective or technical maps; (a) elevation, (b) slope, (c) SAGA wetness index, and (d) annual potential incoming solar radiation derived form a 2-m digital elevation model for the Schäfertal catchment (Schröter et al., 2015). 114

Figure A-3: Soil moisture measurements in the Schäfertal. Locations for sampling the target parameter volumetric soil moisture are indicated by black dots... 115

Figure A-4: An exemplary result of regression in the training phase (a and b) and the test phase (c and d) in crop (a and c) and grass (b and d) area.. 116

Figure A-5: (a) Probabilistic map of volumetric soil moisture content using ANN models based on the grass and crop land use. The scale for each point is 50*50 m. Each glyph depicts soil moisture (color) and relative frequency (length). One exemplary glyph is zoomed in Figure A-5a to present the results of the 1000 ANN models for the related position. (b) shows the most likely soil moisture extracted from (a)... 117

Figure B-1: Structure of an Artificial Neural Network. We have three layers. The input layer prepares data for feeding the ANN. The operation of the hidden layer is determined by inputs and weights of inputs (W). The operation of the output layer is guided by the hidden layer and connected to the results of ANN training.. 120

Figure B-2: Flowchart of processing steps for the prediction of 2D tip resistance models and tomographic model ranking.. 121

Figure B-3: Geophysical velocity tomograms achieved by fully non-linear SOI. Rectangular grid cells of 1 m lateral and 0.5 m vertical side lengths have been used for model parameterization. The black lines illustrate (a) 30 radar, (b) 30 S-wave, (c) 30 P-wave velocity models... 122

Figure B-4: Results of our 2D tip resistance prediction shown as histographic plot. The dotted white lines show the measured logging data of tip resistance that are used for training the ANN. Red colors correspond to high relative frequencies. Blue colors correspond to low relative frequencies. Note the reduced sharpness of prediction at depths where the logging data are different and cannot be brought in full compliance with the velocity variations in the tomograms. 123

Figure B-5: MSE from ANN training for 30 models of radar, S-wave and P-wave velocity. Blue color for radar models, black color for S-wave and red color for P-wave seismic models. Models with low MSE can be brought more easily in compliance with the tip resistance logging data. 124

Figure C-1: Structure of a three-layer Artificial Neural Network. The input layer prepares data for feeding the ANN. The operation of the hidden layer is determined by inputs and weights of inputs (W). The operation of the output layer is guided by the hidden layer and connected to the results of ANN training.. 127

Figure C-2: Processing steps for the prediction of 2D sleeve friction models constrained by ill-posed geophysical tomography.....128

Figure C-3: 2D geophysical velocity tomograms illustrated as laterally neighboured 1D velocity panels. The black lines illustrate 30 equivalent (a) radar, (b) P-wave, (c) S-wave velocity models. Tomographic grid cells have 0.5 m and 1m vertical and lateral side lengths, respectively. 130

Figure C-4: Results of our 2D probabilistic prediction of sleeve friction. The dotted black lines show the measured logging data of sleeve friction that are used for training the ANN and calculation of the logging error. Red color corresponds to high relative frequencies. Blue color corresponds to low relative frequencies. Note that the ANNs trained with (a) with the MSE performance measure offer increased ranges of sleeve friction compared to the results (b) achieved when using the WMSE... 131

Figure G-1: Two exemplary bands of the hyperspectral data, each band carries random noise and anthropogenic effects... 138

Figure G-2: (a) 72 bands of information for one pixel in 2D area, red, gray, and blue line show measured data, fitted linear model, and difference between measured data and fitted model, respectively. (b) all measured data (red line) and difference between measured data and fitted liner model for 10000 pixel in the 2D area. The selected two position presents the decreasing of the difference between noisy points (lower bands) and normal points (higher bands), in the measured data and fitted data, which in fitted results this difference has been decreased. 138

Figure G-3: Results of the clustering method. (a) Total boundary map of selected 72 bands, (b) Clustering results for small part of Schäfertal (100*100 pixel) based on the hyperspectral datasets for desired four clusters... 139

Figure H-1: Histogram normalization for (a) elevation, (b) slope, (c) TWI, and (d) insolation maps of the Schäfertal catchment....141

Figure H-2: Boundary detection results. (a) Boundary of TWI, (b) boundary of insolation, (c) total boundary of technical maps, and (d) total boundary of subjective and technical maps... ..142

Figure H-3: (a) Distribution of 1000 samples. (b) - (d) clustering results for 30 clusters based on 1000, 500, and 250 samples, respectively. 143

List of Tables

Table 2-1: Parameters used in equations 4-1,4-2, and 4-3. All layers A-G are considered to consist of sandy or gravelly saturated sediments; Layers A and B are considered slightly consolidated, the pore fluid in layers D and G comprises a non-aqueous phase liquid component in addition to water. 28

List of Symbols

Symbol	Denotation
a	Coefficient
Σ	Relative noise component (coefficient)
C	Velocity of an electromagnetic wave in air
ε_f	Relative dielectrical permittivity of dry matrix material
ε_m	Relative dielectric permittivity of dry matrix material
ε_r	Dielectric permittivity
E	Dielelctrical permittivity/ weights of the related training tuples
f_s	Sleeve friction
q_c	Tip resistance
I / i	Coefficient
J / j	Coefficient
N	Number of tuples
${}^{P}\Delta$	Uncertainty of P-wave velocity tomogram
${}^{R}\Delta$	Uncertainty of radar-wave velocity tomogram
${}^{S}\Delta$	Uncertainty of S-wave velocity tomogram
P	First absolute derivative between related neighboured readings in the log
P_d	Value of data point p in a d-dimensions dataset
Θ	Angle
Φ	Porosity
V	Velocity
V_m	P-wave velocity of dry matrix material
V_f	P-wave velocity of pore fluid
Θ	Logging data error
Ω	Coefficient
χ^2	χ^2 distance

List of Abbreviations

Abbreviation	Denotation
2D	2 Dimension
3D	3 Dimension
ANN	Artificial Neural Networks
BO	Boundary of objective map
BS	Boundary of subjective map
CPT	Cone Penetration Test
DLR	Deutsches Zentrum für Luft- und Raumfahrt e. V.
E	Edge
EGU	European Geosciences Union (EGU)
EMI	Electromagnetic Induction (Data)
EP	Error of P-wave
ER	Error of Radar
ES	Error of S-wave
ESF	Relative error of measured sleeve friction
F	Fluid
HTL	Human in the Loop
Hz	Hertz
IV	Information Vector
G	Graph
KDD	Knowledge Discovery in Databases
LP	Logging porosity
M	Dry matrix material
MHz	Mega-Hertz
MPa	Mega-Pascal
Ms	Millisecond
MSE	Mean squared error
Ns	Nanosecond
NASA	National Aeronautics and Space Administration
OP	Original porosity
P	P-wave velocity
P	Point
PL	Path lengths
PSO	Particle swarm optimization
RSLP	Radar Seismic Logging Porosity
RSOP	Radar Seismic original Porosity
Rms	Root mean squared
S	Size
S/s	Seismic S-wave velocity
SAGA	System for Automated Geoscientific Analyses
SP	Shortest path
SSE	Sum of squared errors
Sig	Sigmoid function

Abbreviation	Denotation
Sim	Similarity
SOI	Self-organizing inversion
SWI	SAGA wetness index
TBM	Total Boundary Map
TERENO	TERrestrial ENviromental Observatoria
TIR	Total annual incoming solar radiation
V	Vector
W	Weight
WMSE	Weighted mean squared error

“Indeed, in the creation of the heavens and the earth and the alternation of the night and the day are signs for those of understanding”

Quran (3:190)

Chapter 1
Introduction

1.1 Knowledge Discovery in Databases (KDD)

The different research disciplines (i.e., earth sciences, bioinformatics, engineering, biology, computer science, etc.), industry, or customer centered and service-oriented business are overwhelmed with the big amount of data. Data are measured or collected values about a desired quantity stored as raw material in many different types of databases that fuels different disciplines (sciences, industry, engineering, medicine, etc.) growth if only the data can be mined (Al-Hegami, 2004; Han et al., 2011). Since the 1980s database technology has been characterized by research and development activities which promote the development of application-oriented database systems, such as spatial, temporal, multimedia, stream, sensor, scientific and engineering, knowledge bases, and office information databases. The rapid progress of computer hardware and measurement tools in the past three decades has led to large supplies of powerful and affordable computers, data collection equipment, and storage devices. Based on such available and powerful technologies, it is not an exaggeration to say that the data get doubled every year due to the mechanical production of it (Stanton, 2012). Potential and abundance of big databases has been described as data-rich but information-poor situation in different disciplines and urges the need for powerful data analysis tools.

Researchers in different areas i.e., statistics, machine learning, artificial intelligence, expert systems, databases, visualization, etc., are striving to find new methods and techniques to transfer data into an effective, meaningful, and useful information that can play an important role in decision support systems. The advances in computer hardware, data collection, and database technologies make a huge number of databases and information repositories available for knowledge discovery in databases (KDD) (Olson, 2008; Han et al., 2011).

KDD is the process of extracting previously unknown, hidden, effective, and interesting patterns or information from a huge amount of data stored in databases. This type of analysis is an interactive and iterative process which involves many steps that must be done sequentially, attempting to solve the analysis and complexity in the big databases (Hegami, 2004). Figure 1-1 shows the structure of KDD. This process begins with the understanding of databases, mining the data, and ends with analysis and evaluation of the results. Actual extraction of patterns is preceded by a preliminary or pre-processing (Fayyad et al., 1996; Olson, 2008; Han et al., 2011) of data, followed by an integration or selection of appropriate data from different databases. The main tasks in this step are pre-processing (removing noise and inconsistent data), data integration (where multiple databases may be combined with each other), data selection (where relevant data for the subsequent data mining part are selected from the database by feature selection algorithms), and data transformation (where the data are transformed or

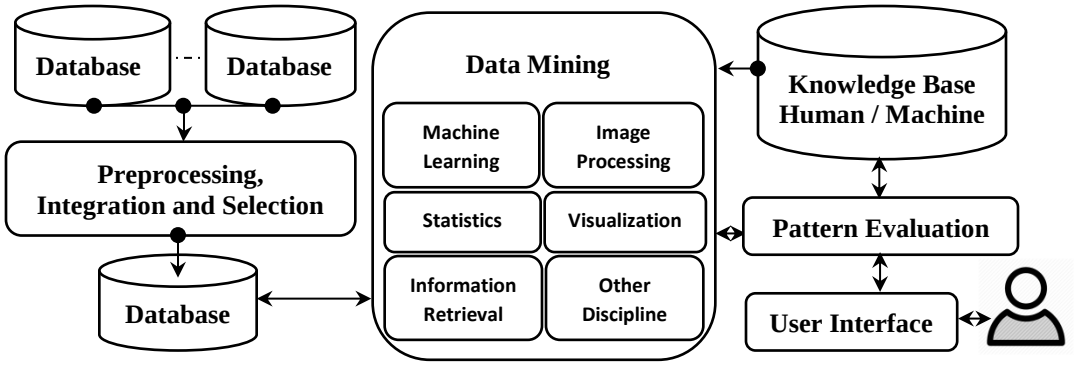

Figure 1-1: The structure of knowledge discovery in databases (KDD). It is based on the three important steps: preprocessing, data mining, and information evaluation by the end user. During these three steps the effective information will be extracted from big databases by KDD.

consolidated into forms appropriate for data mining by performing summary or aggregation operations) (Al-Hegami, 2004; Olson, 2008; Han et al., 2011).

The pre-processing step is considered to be the most time-consuming stage of KDD (Zhang et al., 2003). The results of this step will be stored in a database and are influenced by the extraction algorithms used in the data mining (second) stage. The most important step of KDD is the data mining stage which is an essential process where intelligent, statistical, and machine learning methods are applied in order to extract important patterns and information from the preprocessed database. In the last step, the visualization and knowledge representation techniques are used to present the mined knowledge to the end user. During an iterative and interactive cycle, the expert user can have interaction with the data mining algorithms and send a feedback to the data mining algorithms. Such interaction can help different KDD stages to prove their results.

In the heart of the KDD process (shown in Figure 1-1) is data mining, which refers to extracting, discovering, or “mining” effective, useful, and interesting knowledge, pattern, or information from large amounts of data stored in the databases (Han et al., 2011, Al-Hegami, 2004). Many kinds of literature refer to data mining as a synonym for KDD, but it is the core or an essential step in the process of KDD (Han et al., 2011). Data mining involves an integration of techniques from multiple disciplines such as database technology (Silberschatz et al., 1997), statistics (Hill et al., 2006), machine learning (Witten and Frank, 2005), visualization (Cleveland, 1993), information selection (Kohavi and John, 1997), pattern recognition (Gonzalez and Thomason, 1978), image and signal processing (Russ and Woods, 1995), spatial or temporal data analysis (Bailey and Gatrell, 1995), etc. Therefore, operationally data mining involves the process of discovering patterns automatically or semi-automatically from large quantities of data based on the application of the mentioned disciplines. The extracted information by the data mining algorithms (i.e., clustering, classification, regression, association rules mining, etc.) can be used for different applications ranging from science exploration, industry, engineering, medical science, environmental earth science, production control, to decision making, market analysis, fraud detection, customer retention, etc.

Noticing to the structure of KDD, there are some major issues in data mining regarding mining methodology, user interaction, performance, and diversity of the data

types, which have significant effects in the extracted knowledge or patterns by data mining algorithms. The mining methodology and user interaction issues are related to the kinds of knowledge at multiple granularities, the use of domain knowledge, and knowledge visualization. The issues of mining different kinds of knowledge in databases are due to different users which are interested in different kinds of knowledge. Therefore data mining should cover a wide aspect of data analysis and knowledge discovery algorithms, i.e., data characterization, association and correlation analysis, classification, clustering, regression, outlier detection, etc. These algorithms may use the same database in different ways and require the development of numerous pre-processing, integration, transformation and selection techniques.

Interactive mining of knowledge or human in the loop at multiple levels of KDD are another issue which should be considered during the KDD process. Because it is difficult to know exactly what can be extracted from a database, therefore the KDD models or data mining algorithms should be an interactive procedure between machine and expert users, who want to get benefit from data mining results. Such interactive mining procedure allows KDD and users to focus on the search domain for patterns and structures as well as providing and refining data mining tasks based on extracted results. In this procedure, the user can interact with the data mining algorithms to view and evaluate the data and extracted patterns at multiple granularities and from different angles. Also, for offering better results, data mining algorithms should have cooperation with background knowledge or information regarding the domain under study. This cooperation guides the data mining process and allows discovered patterns to be represented in concise terms and at different levels of abstraction. Domain knowledge related to databases from the expert user, such as integrity constraints and deduction rules, can help the data mining process to focus and speed up, or judge the interestingness and efficiency of the extracted patterns (Han et al., 2011). Another important issue in the KDD or data mining process is handling uncertain, noisy, or incomplete data. The data stored in a database may be uncertain, reflect noise, exceptional cases, or incomplete. When applying the data mining algorithm to such uncertain data, it may confuse the process, and cause the data mining algorithms to over-fit the data. Therefore, the accuracy of the data mining methods and efficiency of

discovered patterns can be poor. Understanding the different type of uncertainties, noise, or errors in the data, and applying data cleaning or uncertain data analysis methods that can handle uncertain and noisy data are required.

1.2 Uncertainty and Error in Data

In data measurements, the terms “error” and “uncertainty” are used to describe the same concept, when the measured data are unsure, noisy, or incomplete. With the emergence of new measurement technologies and application domains, such as location-based services and sensor monitoring, uncertainty is ubiquitous due to reasons such as outdated sources, environmental noise, sampling error, limited number of observations, or imprecise measurement (Taylor, 1982; Zhang et al., 2003; Han et al., 2011). Therefore uncertain and complex databases have become ubiquitous, and lead to a number of unique challenges in the KDD and data mining process. As the volume of uncertainty increases in the databases, the cost of mining and evaluating will also increase. During mining or analyzing the uncertain database, the error, or uncertainty start to have significant effects on the results of KDD and data mining, because most algorithms just assume that the input data is completely reliable and true (Aggarwal and Philip, 2009; Aggarwal, 2010). In such scenario, data records are typically represented by probability distributions reflecting the inherent uncertainty or error rather than deterministic value. These situations have created a need for uncertain data management and mining algorithms for managing, and leveraging uncertainty to improving the quality of the KDD and data mining results.

Recognizing different kinds of uncertainty or error in the data is a fundamental step to manage them during KDD process. In the literature error or uncertainty are classified in the different classes, i.e., systematic, random, gross, additive, multiplicative, absolute, relative, static, and dynamic classes (Wiley and Ltd, 1982), but the two major types of uncertainties in measured data or databases are random and systematic uncertainties. Random uncertainty is an important type of uncertainty which is due to a deficiency in defining or measuring the physical quantity or associated with unpredictable variations in the experimental conditions under which the experiment is being performed. Also, they may arise from fluctuations in either the physical quantity due to the statistical nature of

the particular phenomena or the judgment of the experimenter, such as estimation of scale reading or variation in response time (Bevington and Robinson, 2003). Random uncertainty decreases the precision (how closely two or more measurements agree with each other) of an experiment (Pengra and Dillman, 2009). Systematic uncertainty is another type of uncertainty, which is related to built-in errors in the measuring instruments either in techniques, calibration, or design of the experiment. Systematic uncertainty decreases the accuracy (how close a measured value is to the true value or accepted value) of an experiment (Bevington and Robinson, 2003; Pengra and Dillman, 2009).

In the data measurement one option to estimate the systematic uncertainty is changing every component of an experimental setup which is highly expensive. In some field of application a true or accepted value for a physical quantity may be unknown. In such cases, there is no practical efficient procedure to estimate the systematic uncertainty, and it is sometimes not possible to determine the accuracy of a measurement.Measurements may have different combinations of accuracy and precision. Four combinations of accuracy and precision, which may occur in measurements are shown in Figure 1-2. When experiment or measured data are reported, the report most represents the uncertainties (or the combination of accuracy and precision) in the measured data or calculated values for physical quantities. The uncertainty of a quantity can be represented by probabilistic (Sarma et al., 2006; Zhang et al., 2008), Interval (Abrahamsson, 2002), or fuzzy (Galindo, 2005; Zhang et al., 2008) representation. These representations state how sure we are that the ‘true value’ is within the margin. Probabilistic representation is the most common approach used to represent uncertainty with a probabilistic distribution of the measured values. Interval analysis can

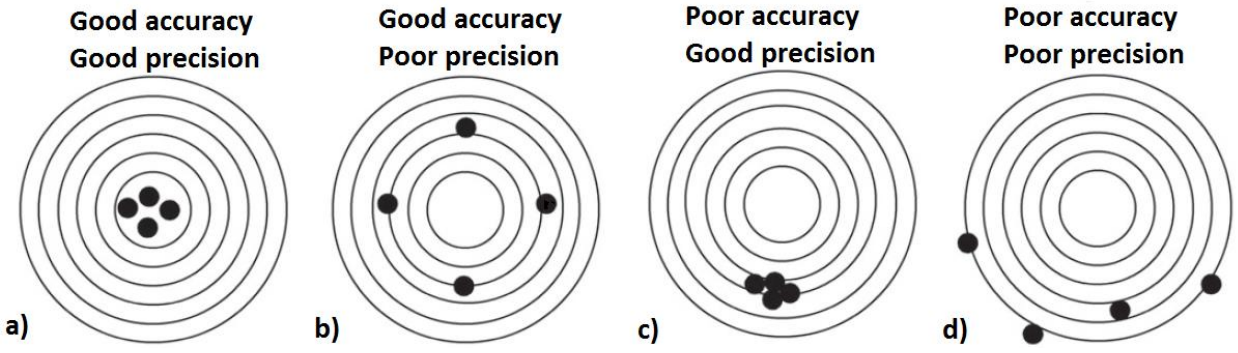

Figure 1-2: Four combinations of accuracy and precision (Haibo, 2014).

be used to estimate the possible bounds to represent uncertainty about the quantities based on the interval representation or some statistic methods (Abrahamsson, 2002). In fuzzy representation, fuzzy entities, fuzzy attributes, fuzzy aggregation, fuzzy relationship, fuzzy membership, fuzzy constraints, etc., are used to represent the uncertainty and imprecision of the measured value (Galindo, 2005; Zhang et al., 2008).

As was explained, the uncertainty of the measured data has significant effects on the results of KDD and data mining algorithms. Hence, for offering realistic data analysis results, in each stage of the KDD process (pre-processing, data mining, pattern evaluation stages) the random and systematic uncertainties of the measured data must be taken into account. This procedure is called error propagation (Taylor, 1982; Haibo, 2014). The general purpose of this step is exerting the uncertainty of the measured data to estimate the highest precision and extracting probabilistic results from KDD or data mining algorithms (Bevington and Robinson, 1992; Haibo, 2014). Random uncertainty is much easier to quantify and propagate than the systematic uncertainty in the KDD or data mining procedure. For example, it can be determined by standard statistical techniques that measure variability or standard deviation of measured values for physical quantities (Sokal and Rohlf, 1995). When data are stored in the databases the only way for determining the systematic uncertainty is using the subjective beliefs of an expert user who has skill in the measurement and is able to determine the structure and behavior of the phenomenon in the field of study (Pengra and Dillman, 2009). Therefore, it is important recognizing or noticing all irregular phenomena and structures. Some information about the surrounding physical conditions, which can become the sources of systematic uncertainty, should be recorded during the data measurement.

1.3 Subjective and Technical Data

Data are either recorded using technical measurements by sensors, referred to as technical data in the following, or human expert knowledge, often in combination with an at least partly visual inspection of the object or area of interest, referred to as subjective data. Technical data are generally objective images or values of the physical quantity providing true information within the resolution of the sensor superimposed by random and systematic uncertainty (Paasche et al., 2014). Examples are satellite imagery or

geophysical maps in the environmental earth sciences. Technical data can be quantified by using statistical methods (Meeker and Escobar, 2014), or may be based on data-driven and domain-independent methods (Kadlec et al., 2009). Subjective data are determined based on user experience and understanding of the behavior of measurements, phenomena, and patterns in the domain, and shows how the domain has been perceived by an expert human. Such information can be either fully correct or incorrect for an individual physical quantity (Paasche et al., 2014). Quantification of data accuracy is usually not possible for subjective data, but is inherently subjective and highly specific.

Use of objective or technical data in the KDD or data mining algorithms often leads to uncertain data analysis. Subjective believe of the expert user about the domain and measurements are required for considering the uncertainty, to improve, and achieve the high-quality KDD results (Haibo, 2014). Subjective believes of the user can be used in each stage of KDD process (pre-processing, data mining, pattern evaluation stages) to help the KDD or data mining algorithms to discover and extract novel, useful, realistic, and interesting knowledge (Aggarwal, 2010; Holzinger, 2016).

1.4 Human in the Loop (HTL)

Interactive or human in the loop machine learning (HTL) (Rothrock and Narayanan, 2011; Liu et al., 2014; Holzinger, 2016) is a simulation framework, which requires human knowledge and experiences about the domain in an interactive model. Traditional knowledge discovery models observe human interaction as an external input in the case study database to use in the different stages of the KDD process. In the pre-processing, data mining and pattern evaluation stages of the KDD process (Figure 1-1) there are some problems (i.e., managing uncertain databases, reducing the volume of discovered patterns, or focus on the important pattern) for which the subjective belief of expert users is necessary to prove the KDD results (Geng and Hamilton, 2006; Holzinger, 2016). The interaction between KDD and a human in a HTL model typically does not provide us with completely true knowledge about the domain or quantity in the database. But these interactions are partially subjective and simply the results of a decision made by an expert user based on his observation, deep understanding, and skills in the domain. In this interaction the expert user specifies some constraints in the form of textual, conditions,

e,g., an additional input layer. By using an interface the user can be aware of the current state of the KDD process and is enabled to manipulate a data mining algorithm through interaction (Ankerst, 2001). As showed in Figure 1-1, such interaction can be done in an iterative loop in pre-processing, data mining, or evaluating the results of data mining algorithms with sending a feedback to each stage of a KDD process. Therefore, the HTL strategy is one of the key sources of a knowledge discovery process, providing enormous potential for economical and autonomous optimization, and proving KDD models to offer more realistic and useful knowledge. In such models, gaining unprecedented amounts of world knowledge is required to solve some of the complex knowledge discovery problems (Fails and Olsen 2003; Raman).

Figure 1-3 shows different scenarios for interaction between a KDD process and an expert user. Figure 1-3a shows unsupervised learning (Albalate and Minker, 2013; Holzinger, 2016) where the learning algorithm is applied to the raw data and learning procedure (i.e., clustering or association rules mining) is fully automatic. This strategy does not require a human to manually label the data, but in the evaluation stage of the KDD process (Figure 1-1), the expert user can analyze the discovered knowledge or patterns objectively and/or subjectively to form a filter that minimizes the number of discovered rules and pattern which are easier to understand (Figure 1-3a). Figure 1-3b shows supervised machine learning (Kotsiantis et al., 2007) in which a human provides labels for the training data and/or selects features to feed the learning algorithm. In supervised learning (i.e., classification,) the subjective belief of the user can be used as a filter to concentrate and select a set of instances that should be given more attention and determining features, which are more important to the learning algorithms. Figure 1-3c shows semi-supervised learning (Chapelle et al., 2009; Albalate and Minker, 2013) which can be a mixture of supervised and unsupervised learning that uses mixing of labeled and unlabeled data to find labels according to a similarity measure to one of the given groups in the learning algorithm. Figure 1-3d illustrates the human in the loop strategy where the human expert is seen as an agent directly involved in the actual learning or data mining stage of the KDD process. In this strategy, the subjective belief of

the user can guide the mining process to form a constraint in order to discover rules or patterns which are more efficient and realistic.

With HTL based knowledge discovery models, on the one hand, users greatly benefit from the knowledge extracted by these KDD models. On the other hand, these KDD models can greatly benefit from the domain knowledge, which users provide and communicate through their interactions with the model. During this interaction, first, the domain knowledge of the human (e.g., focus on certain patterns, information about attribute values, or volume and type of uncertainty or error in the data) can be transferred to the KDD process and vice versa. Second, if the subjective belief of a human can specify how to search, or focus patterns in the domain it can make KDD or data mining algorithms

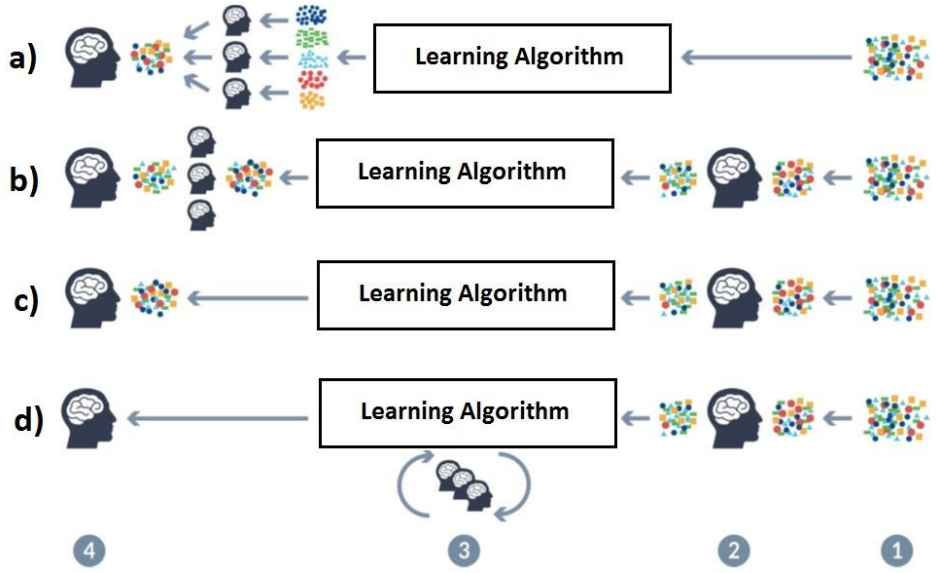

Figure 1-3: Four different interactions between KDD and humans. (a), (b), (c), and (d) show unsupervised, supervised, semi-supervised, and human in the loop approaches, respectively (Holzinger, 2016).

more effective. Because a data mining algorithm typically searches in large search spaces, which the involvement of the user can narrow down significantly and prove the data mining algorithm by more accurate search (Ankerst, 2001). The key challenge in the HTL based KDD process is that the subjective belief of users typically does not fit the standard machine learning or data mining algorithms outcome. However, for proving and offering realistic results of complex knowledge discovery models in today's technological landscape humans as active participants must be included in their pre-processing, data mining, and evaluation stages.

1.5 KDD in Earth Sicences

Earth sciences typically contain many interrelated components and involve several disciplines, i.e., biology, geophysics, hydrology, chemistry, etc. Earth sciences are in part concerned with the observation of environment variables for the purpose of describing the process, pattern and structure understanding. Observing, analyzing or modeling the terrestrial environments provide the knowledge-based model to handle a wide variety of environmental, societal, or industrial issues such as ecosystem management or resource exploration (Recknagel, 2001). Recently, due to rapid developments of measurement tools, the earth science discipline experienced a rapid transformation from a data-poor to a data-rich situation (Kumar, 2010). In particular, geophysical, biological, hydrological and environmental observations of spatial data, time related data, acquired by remote sensors or on-site recording systems, as well as outputs of the large-scale computational platforms used for earth monitoring or exploring provide terabytes of temporal, spatial and spatio-temporal data (Kumar, 2010). These worth and big datasets offer a great potential for discovering, understanding, and predicting the behavior of the earth's system to advance the different disciplines, i.e., biology, geophysics, hydrology, chemistry, etc.

A variety of technical, economic, ecological, social and environmental factors increase the complexity of the earth sciences (Spate et al., 2006). For real world environmental problems, the measured data, which explain the physical environmental quantities, come from different sources with different format, resolution, and uncertainties. In some cases, the environmental databases carry a big volume of technical (i.e., satellite imagery and geophysical maps) and subjective information (i.e., soil maps and geological maps) about the structures in the terrestrial environment. Offering knowledge discovery models for extraction and analysis of interesting patterns in such databases is a big challenge in the earth sciences (Spate et al., 2006; Kumar, 2010; Paasche et al., 2014). Recently, different KDD models have been introduced in this area to discover patterns, knowledge, and structures from the earth science databases (i.e., Ramachandran et al., 2000; Li and Narayanan, 2004; Hoffman et al., 2011; Siegel et al, 2016). Many researchers, universities, and organization (i.e., NASA, DLR, EGU, and etc.) are focusing on KDD applied to earth sciences databases. More studies are necessary to tackle different challenges of environmental databases (i.e., big data analysis, heterogeneous

spatio-temporal datasets, noisy or uncertain data management and analysis, integrating subjective and technical data, probabilistic analysis, dimensionality reduction, etc.) to prove the results of KDD models and make them trustable models for analyzing the earth science databases (Kumar, 2010; Hoffman et al., 2011; Paasche et al., 2014). The ultimate goal of this thesis involves the advancement of new strategies to setting up different KDD processes for application in earth science databases focusing on probabilistic analysis under uncertainty consideration and integration of subjective and technical data.

1.6 Objectives of This Thesis

This thesis is organized into three different parts, putting forward different objectives to offering knowledge discovery in environmental databases. One important type of environmental databases is geophysical tomography, which offers valuable and unique information about the internal composition of the ground. Geophysical tomographic datasets uniquely offer the ability to image physical parameter variations, e.g., radar or seismic wave propagation velocities, in a spatially continuous manner. The most important challenge when using geophysical tomography in hydrological, environmental or engineering exploration is, how to link the tomographically reconstructed physical parameter variations to the aquifer, reservoir or geotechnical target parameters of interest, which are usually different from those imaged by geophysical tomography (Paasche et al., 2006; Rumpf and Tronicke, 2014). The second chapter of this thesis shows a data analysis model allowing 2D or 3D probabilistic prediction of sparsely measured earth properties constrained by geophysical imaging fully accounting for tomographic reconstruction ambiguity. The main focus of this chapter is in the pre-processing and data mining stage of the KDD process (Figure 1-1). This model tries to take the uncertainty or variability of the input data into account for offering a probabilistic prediction of the target parameters. Additionally, this chapter evaluates, whether the training performance of the prediction model can be used to rank geophysical tomograms. If this ranking is successful then, the feature selection, in an iterative relation between pre-processing and data mining stage, can be applied to the input data to decrease the volume of input data and increasing the performance of the KDD models.

Another main issue in the environmental databases is uncertainty or measurement error in the measured data (e.g., tomographic ambiguity, borehole logging data errors). For a realistic analysis (i.e., predictions of geotechnical target parameters) the uncertainty, or measurement errors, and the differences in spatial resolution must be taken into account (Rumpf and Tronicke, 2014; Asadi et al., 2016). The third chapter represents a spatially continuous probabilistic prediction model of sparsely measured ground properties constrained by ill-posed tomographic imaging considering data uncertainty and resolution. Such prediction model can contribute to solving hydrological, petroleum, or engineering exploration tasks. The main focus of this chapter is improving the results of the data mining stage (Figure 1-1) of the KDD process with application to an environmental earth database by feeding the uncertainty and measurement errors in the learning phase of the prediction model. Four different training strategies taking into account the uncertainty of logging data and geophysical tomographic ambiguity to avoid data overfitting of the prediction model are considered. This chapter shows a successful transformation of the uncertainty of logging data and geophysical tomographic reconstruction ambiguity as well as differences in spatial resolution of logging and tomographic models into the probabilistic 2D or 3D prediction of our target parameters in a data-driven manner. Such strategy allows application of the presented methodology to any combination of geophysical tomograms and hydrologic, petroleum or engineering target parameters solely measured in boreholes. Furthermore, the concept is also applicable to predict spatially continuous maps on the basis of geoscientific maps, e.g., probabilistically interpolating sparse soil moisture measurements on the basis of multiple geophysical, geochemical or remotely sensed maps.

Mapping the earth is a fundamental prerequisite required to address various environmental and economic issues, such as mining target identification, soil conservation, or ecosystem management (Odeh et al., 1990; Paasche and Eberle, 2009; Behrens et al., 2010). Typical databases comprise disparate spatial datasets mapping the spatial variability of physical, chemical, biological or other properties of the earth considered to realize educated decisions on land and resource utilization. The datasets for mapping are either technical or subjective. In the fourth chapter of this thesis I present first attempts towards integrating technical and subjective spatial datasets in an

automated and rapid manner aiming to produce a crisp or fuzzy classified map outlining dominant structures in the database optimally consistent with all available information. This chapter focus on developing a KDD model based on the human in the loop (Figure 1-1) to integrating subjective and technical databases in the earth sciences. This new model may potentially allow for multi-map integration and map analysis according to different features, such as color or absolute value, edge, and texture information in the mapped technical information and additional consideration of knowledge provided by human experts. The analyses in this chapter are based on a real dataset acquired in the Schäfertal, Germany, which is part of the TERENO Harz/Central German Lowland Observatory.

Finally, the fifth chapter of this thesis presents conclusions and outlook where the major findings of this thesis are explained. In this chapter indications arising from this thesis for future research directions and recommendations for proving the discussed models for environmental earth data analysis are discussed. Furthermore, new application areas for which such models can be applied for discovering patterns, structures and knowledge are addressed.

Chapter 2

2D Probabilistic Prediction of Sparsely Measured Earth Properties Constrained by Geophysical Imaging Fully Accounting for Tomographic Reconstruction Ambiguity

Abduljabbar Asadi, Peter Dietrich, and Hendrik Paasche
Manuscript published in Environmental Earth Sciences, 2016

2.1 Abstract

Many hydrological, environmental, or engineering exploration tasks require predicting spatially continuous scenarios of sparsely measured borehole logging data. We present a methodology to probabilistically predict such scenarios constrained by ill-posed geophysical tomography. Our approach allows for transducing tomographic reconstruction ambiguity into the probabilistic prediction of spatially continuous target parameter scenarios. It is even applicable to datasets where petrophysical relations in the survey area are non-unique, i.e., different facies related petrophysical relations may be present. We employ static two-layer Artificial Neural Networks (ANNs) for prediction and additionally evaluate, whether the training performance of the ANNs can be used to rank geophysical tomograms, which are mathematically equal reconstructions of physical parameter distributions in the ground. We illustrate our methodology using a realistic synthetic database for maximal control about the prediction performance and ranking potential of the approach. For doing so, we try to link geophysical radar and seismic tomography as input parameters to porosity of the ground as target parameter of ANN. However, the approach is flexible and can cope with any combination of geophysical tomograms and hydrologic, environmental or engineering target parameters. Ranking of

equivalent geophysical tomograms based on additional borehole logging data is found to be generally possible, but risks remain that the ranking based on the ANN training performance does not fully coincide with the closeness of geophysical tomograms to ground truth. Since geophysical field datasets do usually not offer control options similar to those used in our synthetic database, we do not recommend the utilization of recurrent ANNs to learn weights for the individual geophysical tomograms used in the prediction procedure.

Keywords: Geophysics, Petrophysics, Probabilistic Prediction, Tomography. ANN, global search inversion

2.2 Introduction

Geophysical tomography offers valuable and unique information about the internal composition of the ground. Such information is essential for supporting many hydrological, environmental, and engineering exploration tasks. Geophysical tomographic datasets uniquely offer the ability to image physical parameter variations, e.g., radar or seismic wave propagation velocities, in a spatially continuous manner. The most important challenge when using geophysical tomography in hydrological, environmental or engineering exploration is to link the tomographically reconstructed physical parameter variations to the aquifer, reservoir or geotechnical target parameters of interest, which are usually different from those imaged by geophysical tomography. Numerous examples, where geophysical tomography is used for the characterization of aquifers and hydrocarbon reservoir (Hubbard et al., 2001; Binley et al., 2001; Tronicke and Holliger, 2005; Paasche et al., 2006; Boisclair et al., 2011; Ruggeri et al., 2013), e.g., with porosity or hydraulic conductivity being typical exploration target parameters, exist. Geophysical tomography is also used for geotechnical ground characterization (Yamamoto, 2001; Angioni et al., 2003; Rumpf and Tronicke, 2014), e.g., with compression and shear strengths as exploration target parameters. In these studies the geophysical tomography offers spatially continuous 2D or 3D information about the subsurface, but the exploration target parameters can only be measured laterally sparse along one dimension in sparse boreholes or by direct-push probing.

Numerous approaches are available to link geophysical tomograms to hydrological or engineering target parameters. Traditional techniques rely on empirical, semi-empirical, or theoretically founded deterministic transfer functions (e.g., Archie, 1942; Gassmann, 1951; Wyllie et al., 1956; Topp et al., 1980; Angioni et al., 2003). Generally, these approaches convert the spatial distribution of one physical parameter imaged by geophysical tomography into a spatial distribution of the desired target parameter. Unfortunately the relationship between physical and exploration target parameters is usually non-linear, non-unique, spatially and temporally variable and hence often not exactly known (Schön, 1998). Recently, methodological frameworks have been proposed which allow for improved incorporation of uncertainty and non-unique inter-parameter relations building on statistical analysis methods, e.g., artificial neural networks (Cawley

et al., 2007), co-kriging (Cassiani et al., 1998; Gloaguen et al., 2001), Bayesian inference (Ezzedine et al., 1999; Hubbard et al., 2001; Chen et al., 2001; Bosch et al., 2010; Boisclair et al., 2011; Ruggeri et al., 2013), fuzzy systems (Paasche et al., 2006), or conditional stochastic simulations (Tronicke and Holliger, 2005; Dafflon et al., 2009). These approaches require measured information about the target parameter to be present, e.g., by borehole measurements or material extraction followed by laboratory analysis.

Artificial Neural Networks (ANN) have been applied to a variety of problems in the geophysical domain, particularly addressing seismic data processing issues (e.g., energy onset picking, deconvolution, trace editing, event classification, waveform recognition, etc.), well log data analysis (e.g., detection of subsurface layer boundaries), geophysical data inversion, lithology classification and porosity prediction based on attributes derived from reflection seismic data and borehole logging data, as well as stratigraphic feature interpretation based on seismic attribute analyses (e.g., Poulton, 2002; Van der Baan and Jutten, 2000; Leite and de Souza Filho, 2009; Khoshdel and Riahi, 2011; Leite and Vidal, 2011; Raeesi et al., 2012). To our knowledge, ANNs have not been used before to link geophysical tomograms imaging different physical parameters with sparse logging data for 2D or 3D probabilistic prediction of sparsely measured earth properties.

For most geophysical tomographic imaging problems, the tomographic reconstruction suffers ambiguity due to limited number of observations and observational accuracy (e.g., Friedel 2003). Traditionally, deterministic approaches based on local-search gradient-based optimization techniques (e.g., Aster et al., 2005) are employed to reconstruct a single geophysical tomogram explaining the data and an additional constraint employed to regularize the ill-posed optimization problem. Local-search optimization techniques do not allow for realistic and quantitative appraisal of ambiguity inherent to geophysical tomographic datasets since they are inherently deterministic. Traditional approaches based on deterministic transfer functions and geostatistical concepts for predicting the target parameters based on geophysical tomograms did not incorporate the tomographic reconstruction ambiguity since they build on geophysical tomograms achieved by local-search optimization techniques. However, quantitative and realistic prediction of uncertainty in the exploration target parameter estimation would

offer valuable information for decision taking in hydrological, environmental, and engineering exploration tasks with regard to risk quantification and minimization.

Recently fully non-linear optimization techniques, e.g., Particle Swarm Optimization (Kennedy and Eberhart, 1995), or Genetic algorithms (Mitchell, 1998), have been employed to reconstruct ensembles of geophysical tomograms fitting the underlying dataset equally well. While some approaches artificially limit the complexity of the subsurface variability to be tomographically reconstructed by using simple geological concepts, such as a layered subsurface (e.g., Velis, 2001; Roy et al., 2005; Tronicke et al., 2012), others are suitable to achieve tomograms of arbitrary and data driven complexity (Bodin and Sambridge, 2009; Bodin et al., 2012). The resultant tomogram ensembles can be used to assess the tomographic ambiguity in a realistic manner. While all these tomograms are mathematically equivalent answers to the geophysical tomographic reconstruction problem, they may resemble the internal composition of the ground to variable degrees. Currently it is not known whether these models can be ranked according to their approximation of reality using additional information from 1D exploration target parameter measured in a borehole. If this could be done successfully, the geophysical tomograms could be ranked and correspondingly weighted for the subsequent linkage to sparsely measured additional target parameters and the prediction of 2D or 3D distributions of these target parameters.

Very recently, such ensembles of equivalent tomograms have been used for constraining the probabilistic inference of spatially continuous 2D predictions of exploration target parameter distributions. Rumpf and Tronicke (2014) employ Alternating Conditional Expectation (Breiman and Friedman, 1985) to link ensembles of 125 radar, seismic P-wave and S-wave tomograms with sparsely measured exploration target parameters, i.e., sleeve friction and effective grain size. In their tomographic reconstruction they rely on the concept of a layered ground and illustrate their prediction uncertainty by mean and median values in combination with percentile ranges. Paasche (submitted a) and Paasche (submitted b) employ fuzzy sets to translate tomographic ambiguity into the probabilistic inference of 2D target parameter models.

In this paper we focus on the probabilistic prediction of 2D distributed porosity of the ground as target parameters, based on ensembles of equivalent radar and seismic wave propagation velocity tomograms. Additionally, we evaluate whether it is possible to rank the mathematically equivalent tomograms based on the measured target parameter information. For doing so a prediction based feature selection (Liu and Motoda, 1998; Liu and Motoda, 2001) method using an ANN (e.g., Seteiono and Liu, 1997; Verikas and Bacauskiene, 2002; Ganivada et al., 2013; Frénay et al., 2013; Yan and Yang, 2015), is applied for probabilistic prediction of porosity target parameters measured in two sparse boreholes and selecting the high quality related geophysical tomograms. We illustrate our analyses using a simulated experiment allowing for full evaluation of the results of the prediction and ranking procedures. Finally we discuss the results and highlight the limitations inherent to the approach particularly when it comes to ranking the tomograms.

2.3 Methodology

2.3.1 Geophysical Tomography as Feature Construction Problem

Feature construction (Liu and Motoda, 1998) transforms a given set of input features to generate a new set that is more utilizable in the subsequent data mining tasks. For example the constructed feature may allow for linear or at least a more unique relation to a desired target parameter than the original data, which enables improved prediction performance. In geophysical tomography a dataset is inverted to achieve one or more geophysical tomograms imaging the spatial variability of a physical parameter (Aster et al., 2005), e.g., seismic velocity. Tomographic reconstruction techniques act here as feature construction step to achieve tomograms that are more suitable for geoscientific interpretation and linkage to exploration target parameters than the measured data itself. For example in Figure 2-1 we illustrate a geophysical cross-borehole tomographic experiment. Energy sources generating a signal, e.g., a seismic wave, are placed at different depths in one borehole. In a second borehole spaced several meters or tens of meters, sensors are mounted recording the excited energy after it has travelled from a source to the receivers. In a seismic or georadar tomographic experiment, we are interested in the traveltime required for the excited wave to reach the receivers. This measured feature is a function of source-receiver distance and material-specific

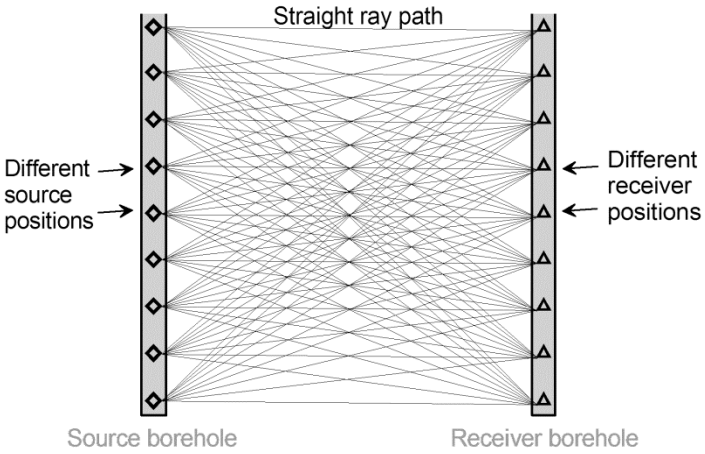

Figure 2-1: Schematic sketch of a geophysical cross-borehole tomographic experiment with sources and receivers in the left and right boreholes.

properties, e.g., velocities, between the boreholes. These velocities are yet unknown, spatially variable, and determine the fastest pathway of energy between sources and receivers. For linking the material-specific velocities with material-specific exploration target parameters, we have to construct new features out of the observed traveltimes that are free of the effects of the experimental setup, e.g., the source-receiver distances. In doing so, we strive to reconstruct the velocity distribution between the boreholes from the set of traveltime observations with different source and receiver positions.

We use a global-search inversion technique (Paasche, 2015) to construct an ensemble of new features, i.e., velocity tomograms, on the basis of the observed traveltimes. The tomographic reconstruction problem is formulated as an optimization problem and solved using particle swarm optimization (PSO; Kennedy and Eberhardt, 1995). We follow Schwarzbach et al., (2005) and set up a bi-objective optimization problem concurrently addressing a data misfit objective (rms error) and a regularization objective enforcing spatially smooth model parameter variations. Opposite to Schwarzbach et al. (2005) who use Tikhonov regularization (e.g., Aster et al., 2005), we employ a spatially acting smoothness constraint originally developed for constrained fuzzy cluster analyses (Pham, 2001). Other than Tikhonov regularization, this constrain does not tend to damp the total model parameter amplitudes and keep them artificially close to the global mean of all model parameters. Instead, it favours solutions that exhibit a piecewise smooth model parameter variability but also allows for significant contrasts wherever found to be necessary to explain the data. When solving the bi-objective optimization problem we balance the interests of the data misfit objective and the smoothness constraint using a game theoretic approach (Balling, 2003) implemented in

the PSO algorithm. The inversion results provide realistic information about the velocity variations between the two boreholes as well as quantitative information about the ambiguity of the tomographic model reconstruction. This method is data driven and does not require prior information about the ground. We obtain several different 2D velocity tomograms from the traveltime dataset, which are considered equivalently acceptable solutions of our tomographic reconstruction problem. Every found tomogram explains the underlying data to the same degree, i.e., the data misfit measure is of equal size for all found tomograms. While these tomograms are more suitable for target parameter prediction than the originally measured traveltimes, we have increased the dimensionality of the feature space during feature construction. Instead of one dataset we have now multiple tomograms, which cannot be ranked according to a feature construction quality measure, such as the data misfit.

2.3.2 Feature Selection by Prediction

The goal of feature selection (e.g., Liu and Motoda, 1998; Ganivada et al., 2013; Frénay et al., 2013; Yan and Yang, 2015) is to identify an optimal subset of features that is particularly suitable for the subsequent prediction task. Albeit feature construction resulted in an ensemble of tomograms that are mathematically equally plausible solutions of the tomographic reconstruction problem, some tomograms may be more realistic images of the internal composition of the ground than others. Since measured information about the exploration target parameters, e.g., acquired in boreholes, provides sparse information about the same ground (reality), we want to investigate, whether it can be used for ranking the tomograms. Selecting the realistic tomograms could improve the accuracy of the subsequent target parameter prediction and can be addressed by solving a typical feature selection problem.

The most straight forward feature selection techniques are filter methods (Liu and Motoda, 1998), which select a subset of features based on general characteristics of features in datasets like correlation. They are run as preprocessing step and operate independently of the subsequent prediction. When employing filter methods the comparison between tomograms and target parameters can for example be done by linear or exponential correlation analysis. This is similar to the traditional petrophysical

transfer function concept for predicting the target parameters by a priori assuming a certain relation between tomographically imaged parameter and desired target parameter, e.g, linear or exponential, as valid across the survey area. When ranking the tomograms according to the correlation with the target parameter the results are not independent from the chosen correlation function. Hence, a priori information about the correct correlation, or transfer function would be required, which is, among others, dependent on measurement resolution, material composition, depositional history, and usually not known beforehand.

This leads us to the utilization of nonlinear feature selection methods avoiding dependence on the proper selection of the correlation measure. Wrapper methods (e.g., Liu and Motoda 1998; Seteiono and Liu, 1997; Liu and Motoda, 2001; Ganivada et al., 2013; Frénay et al., 2013; Yan and Yang, 2015), evaluate the performance and accuracy of feature subsets by integrating the feature selection problem into a learning method that can be used for prediction. Particularly, in situations with unknown correlations between features and target parameters wrapper methods offer more accurate feature selection or ranking than the filter methods (Liu and Motoda, 1998).

In this paper we employ static two layer feed forward ANNs (Hornik, 1991; Ban and Chang, 2013; Widrow et al., 2013) as learning method for prediction based feature selection (e.g., Jain et al., 1996; Verikas and Bacauskiene, 2002; Leray and Gallinari, 2002). Using ANNs we pairwise link physically different tomograms emanating from radar (electromagnetic) and seismic (mechanical) datasets to the exploration target parameter at locations where radar, seismic, and porosity are available for training the ANNs. This is at the left and right edges of the 2D tomographic plane. The prediction models obtained by the ANNs are used to generate spatially continuous distributions of the target parameter in regions not used for training, i.e., for all tomographic grid cells for which no logging data are available. The mean squared error (MSE) of the learned prediction models, which is a performance parameter of ANNs, appears attractive as a measure for optimal tomogram or feature subset selection, i.e., which pair of selected radar-seismic tomograms can be easily brought into coincidence with the target parameter. The MSE in feed-forward ANNs is primarily dependent on the data and it can lead the strategy of

the next step in a recursively learning feature selection machine when using it as weighting parameter for ranking the tomograms (e.g., Van der Baan and Jutten, 2000).

The process of feature selection based on the ANNs can be done in a recursive loop or in a complex recurrent ANN machine (e.g., Bailly and Milgram, 2009; Quan et al., 2014; Yan and Yang, 2015) that internally recouples the output layer to the connectivity of input and hidden layer. The final ranking results from such complex machine are not only depending on the training data, but they are also depending on the chosen machine learning strategy, e.g., how to weight the inputs of the ANNs based on the MSE (i.e. linear, exponential) or output layer. Since it is not known yet, whether our 2D tomograms can be realistically ranked using 1D measurements of the exploration target parameter, we decide to start using a two layer feed forward ANN without automated recursive learning. Such system can be regarded objectively data driven and allows for appraisal of the performance parameters suitability for model ranking. If desired, the MSE can be used in a recursive run of our methodology to manually reweight the contribution of the individual tomograms in the second and later iterations. If desired, such recursive learning strategy can be automated (e.g., Kiranyaz et al., 2009; Miche et al., 2010; Razavi and Tolson, 2011; Jing et al., 2012; Wu et al., 2015; Duan et al., 2015). However, the manual setup allows us to assess the general suitability of an internal performance parameter of the ANN, e.g., the MSE, for feature subset selection and tomogram ranking.

2.3.3 Processing Flow

The flowchart in Figure 2-2 summarizes our processing steps when working towards probabilistic prediction of 2D exploration target parameter distributions and ranking of the tomograms based on the performance of the prediction algorithm. We measured two types of data imaging reality into different sets of observations. The first type is crosshole tomographic traveltime data acquired between two boreholes. Two different traveltime datasets are recorded differing by their physical energy excitation, i.e., electromagnetic radar and mechanic seismic waves. The second type is measured in 1D only at the position of two boreholes. The measured quantity is porosity. The 1D porosity profiles provide sparse information about our target parameter. Porosity is one of the key

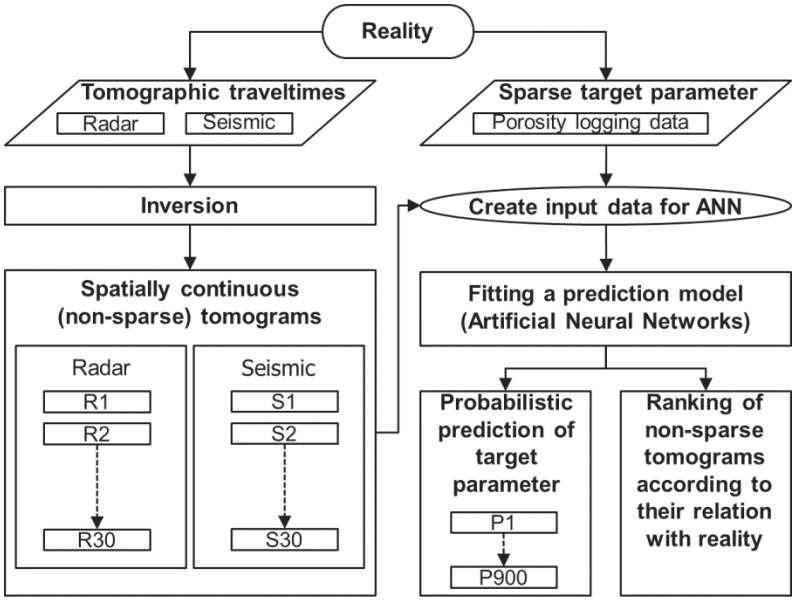

Figure 2-2: Flowchart of the processing workflow for probabilistic prediction of spatially continuous models of sparsely measured target parameters and ranking non-sparse tomograms achieved from inversion or feature construction algorithms.

parameters in hydrological and environmental exploration to understand and simulate fluid flow and storage processes (e.g., water, oil, gas) in the subsurface.

The traveltime datasets carry the information about the physical properties of the ground materials, e.g., radar and seismic wave propagation velocities, and the experiment's geometry. For extracting the ground material information from the spatially continuous traveltimes we follow (Paasche, 2015) to construct 30 mathematically equivalent 2D velocity tomograms from the radar and seismic traveltime datasets, respectively, thus assessing tomographic reconstruction ambiguity. These new features are more suitable for porosity prediction.

To achieve a probabilistic prediction model for our exploration target parameter and for ranking the tomograms, we link the resultant tomograms from the feature construction to the sparse target parameter employing a two layer feed forward ANN. All radar and seismic velocity tomograms are combined pairwise with each other to create the input data for training the ANNs. Based on the 30 radar and 30 seismic tomograms we have 900 different combinations as input scenarios to our ANNs that are all linked to the target parameter. We train the ANNs individually for every input scenario and achieve 900 prediction models, which can be used to generate 900 2D distributions of our target parameter, which is porosity.

Additionally, we calculate the MSE measuring the difference between the output of the ANNs and the measured target parameter. High MSE values indicate a better compliancy among the considered pair of radar and seismic velocity tomograms and the target parameter. We use this information for ranking the radar and seismic tomograms. Since tomograms and target parameter are images of the same reality we hope that tomograms, which can be better brought into coincidence with the target parameter by the ANNs, are the more realistic images of the area between the boreholes. This expectation is also the driving force when striving to constrain geophysical inversion or tomographic model reconstruction by logging data.

2.4 Case Study

2.4.1 Tomographic Data Simulation

In this paper we use a synthetic database for evaluation and illustration of the results of the introduced methodology. The synthetic database has been generated by (Paasche and Tronicke, 2007) as test database for investigating the efficiency of a newly developed joint inversion method. In the way we use it here this database carries three features about the subsurface: first-cycle radar wave traveltime, first-cycle seismic wave traveltime, and sparse porosity information measured in two boreholes.

We begin by describing realistic subsurface conditions using a 2D porosity model that consists of multiple layers representing different lithologies (Figure 2-3a) considered typical for an unconsolidated aquifer. For each layer, a stochastic field with 1.25 cm sample spacing has been generated and superimposed with correlated noise employing a von Kármán auto-covariance function with horizontal and vertical correlation lengths of 120 m and 20 m, respectively, and a moderate raggedness defined by a Hurst number of 0.5. Mean values and standard deviations have been assigned to every layer to achieve a 2D porosity model that could be considered realistic for near-surface sedimentary settings, e.g., gravel, sand, or silt (Figure 2-3b). Using the deterministic transfer functions of Wharton et al. (1980) and Raymer et al. (1980), we convert the porosity model into 2D distributions of radar and seismic wave velocities, respectively (Figures 2-3c, and d)

$$\sqrt{\varepsilon} = \varphi \times \left(\sqrt{\varepsilon_f} - \sqrt{\varepsilon_m}\right) + \sqrt{\varepsilon_m}, \qquad (2\text{-}1)$$

$$v_r = \frac{c}{\sqrt{\varepsilon}}, \quad (2\text{-}2)$$

$$v_p = (1-\varphi) \times v_m + \varphi \times v_f. \quad (2\text{-}3)$$

φ,v,and ε denote porosity, velocity and dielectric permittivity. c denotes the velocity of an electromagnetic wave in air, and the subscripts *r, p, f* and *m* refer to radar, seismic p-wave, pore fluid and dry matrix material, respectively. Table 2-1 lists the values used for the various parameters and each layer. Note that equation 2-3 was developed initially for sandstones, but for the chosen parameters the resulting velocity range can also be regarded as realistic for unconsolidated clastic sediments.

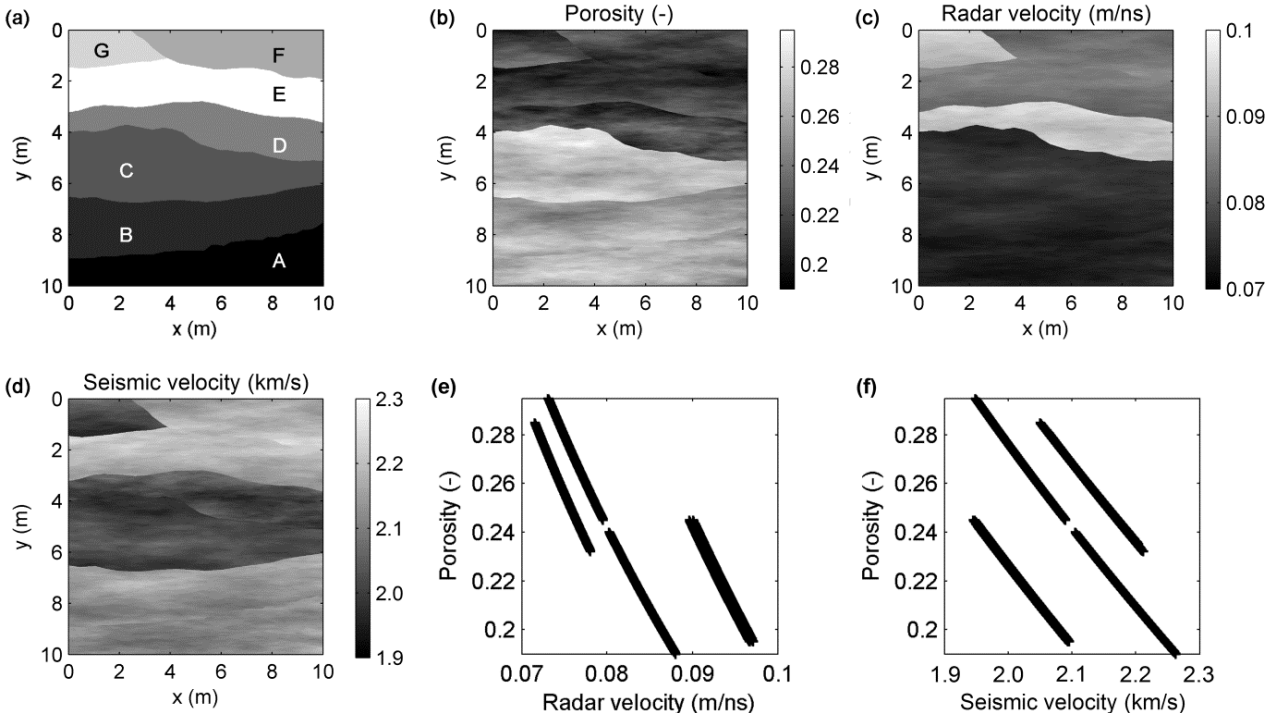

Figure 2-3: Original synthetic models representing ground truth. (a) Layered ground assumed. (b) Porosity variability in the ground. (c) Radar velocity and (d) seismic velocity are obtained from traditional deterministic petrophysical transfer functions used to convert porosity into physical parameters. (e) and (f) scatter plots of the models given in (b) and (c), and (b) and (d), respectively.

The resultant 2D velocity distributions are considered ground truth in our synthetic experiment and we refer to them as the original models. Since we assume a complex subsurface with petrophysical relationships that are facies dependent, the global interrelations for the entire 2D model area between porosity and radar and seismic velocities are non-unique and can thus not be described by a single petrophysical deterministic transfer function (Figures 2-3e and 2-3f). Note, in tomographic field

experiments the true parameter distribution for radar and seismic wave velocities are unknown and can be smeared by noise contaminating the logging data as well as by tomographic reconstruction ambiguity.

Table 2-1: Parameters used in equations 2-1, 2-2,and 2-3. All layers A-G are considered to consist of sandy or gravelly saturated sediments; Layers A and B are considered slightly consolidated, the pore fluid in layers D and G comprises a non-aqueous phase liquid component in addition to water.

Symbol	Values for layers [A, B, C, D, E, F, G] (see Figure 3a)	Description
ε_m	[5.3, 5.3, 4.6, 4.6, 4.6, 4.6, 4.6]	Relative dielectric permittivity of dry matrix material
ε_f	[80, 80, 80, 51, 80, 80, 52]	Relative dielectric permittivity of pore fluid
C	0.3 m/ns	Velocity of an electromagnetic wave in air
v_m	[3140, 3120, 3000, 2810, 3000, 3000, 2810] m/s	P-wave velocity of dry matrix material
v_f	[1560, 1600, 1550, 1400, 1550, 1550, 1410] m/s	P-wave velocity of pore fluid

For simulating radar and seismic cross-borehole tomographic surveys, we assume boreholes to be present at the left and right model edges. To generate tomographic datasets we place sources and receivers in the left and right boreholes, respectively (Figure 2-1). The uppermost sources and receivers in the boreholes are located at 0 m depth with additional sources and receivers being placed along the boreholes with a vertical spacing of 0.25 m. For all source-receiver combinations and the original models in Figures 2-3c and d, radar and seismic traveltime data are determined using finite difference solutions of the electromagnetic and acoustic wave equations followed by picking the times of first energy arrivals. The resultant radar and seismic wave traveltime datasets are shown in Figures 2-4a and 2-4b, respectively, after adding Gaussian random noise to the simulated traveltimes. These datasets are the input for the feature construction step, in order to achieve velocity tomograms of the ground.

2.4.2 Feature Construction By Tomographic Reconstruction

We tomographically reconstruct ensembles of radar and seismic velocity tomograms from the radar and seismic traveltime datasets fitting the underlying datasets equally well. The internal tomographic reconstruction performance parameter used to

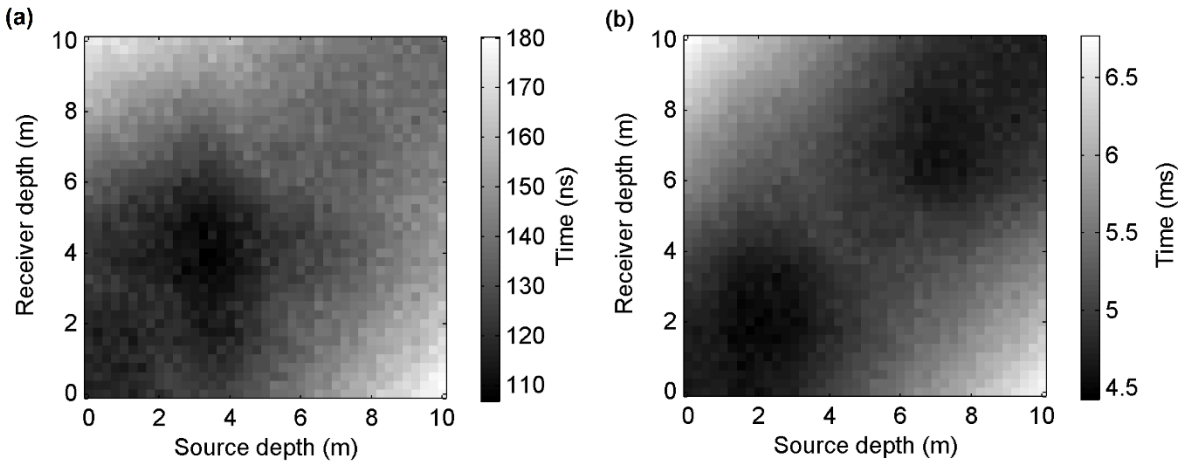

Figure 2-4: Traveltimes of the (a) radar and (b) seismic tomographic datasets.

measure the misfit between the data and the forward response of the achieved tomograms is the root mean squared (rms) error. We reinitialized the inversion 30 times to find 30 independent velocity tomograms for each dataset. All found tomograms fit the underlying dataset within the noise level added to the synthetic traveltimes. The rms errors of the 30 final radar velocity models vary between 2.2378 and 2.2501 ns and correspond to the mean noise level of 2.2511 ns. The rms errors of the 30 final seismic velocity models vary between 0.042567 and 0.042799 ms, which corresponds to the mean noise value of 0.042800 ms.

The final ensemble of the 30 radar velocity tomograms is shown in Figure 2-5. All tomograms are reconstructions of the original radar wave velocity model (Figure 2-3c). When comparing the tomograms in Figure 2-5 with the true model we see that all tomograms correctly indicate regions of higher velocities in the upper half of the tomographic plane. The regions of high velocities in the top left corner of the tomographic plane are also captured by all tomograms. However, the achieved tomograms differ in their ability to reconstruct the region of intermediate velocity in the top right part of the tomograms. Some tomograms show here regions of high velocities (e.g., model R22 in Figure 2-5), which does not resemble ground truth. These differences in the tomograms illustrate the ambiguity of the tomographic reconstruction problem due to limited number of observations and noise-contaminated data.

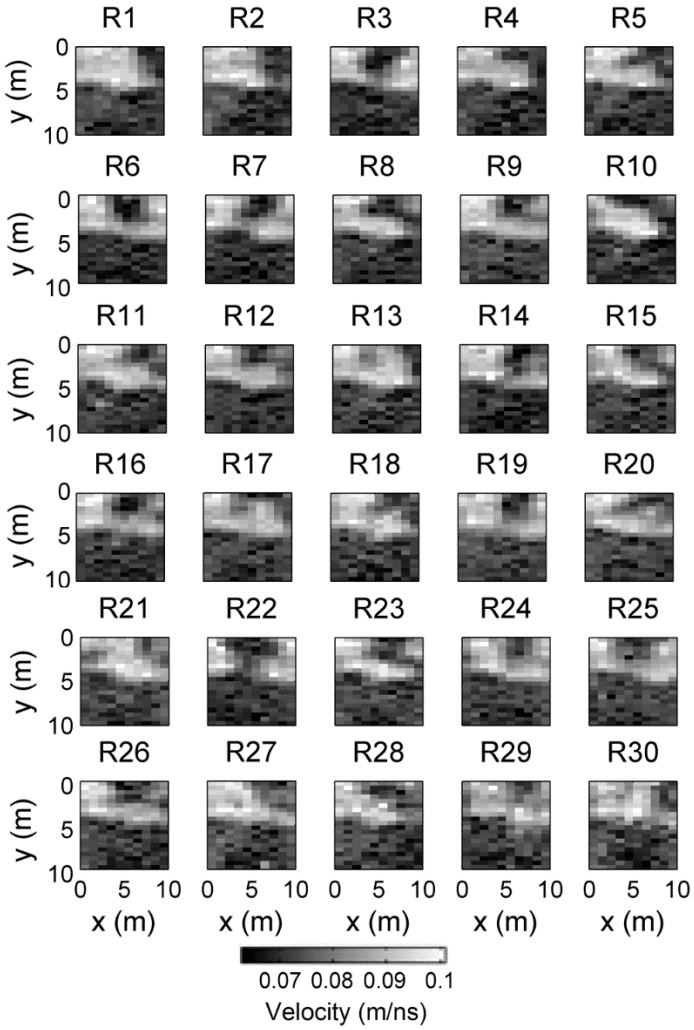

Figure 2-5: 30 tomographic reconstructions of radar wave propagation velocity distributions achieved by fully non-linear self-organizing inversion (Paasche, 2015). The 30 tomograms are achieved by independent inversion runs and fit the underling tomographic dataset equally well.

The final ensemble of the 30 seismic velocity tomograms is shown in Figure 2-6. All tomograms are tomographic reconstructions of the original seismic wave velocity model (Figure 2-3d). When comparing the tomograms in Figure 2-6 with the true velocity model we see that all tomograms again correctly indicate regions of lower velocities in the middle, and higher velocities in the upper and bottom part of the tomographic plane. The high velocities in the top right corner and bottom part, or the low velocity in the middle of the tomographic plane are also captured by all tomograms. However, the constructed tomograms differ in their ability to reconstruct the region of intermediate velocity in the top left part of the tomograms. Some tomograms show in the top left part regions of high velocities (e.g., models S3, and S30 in Figure 2-6), and some tomograms show in the middle bottom regions of low velocities (e.g., models S2, and S4 in Figure 2-6), which

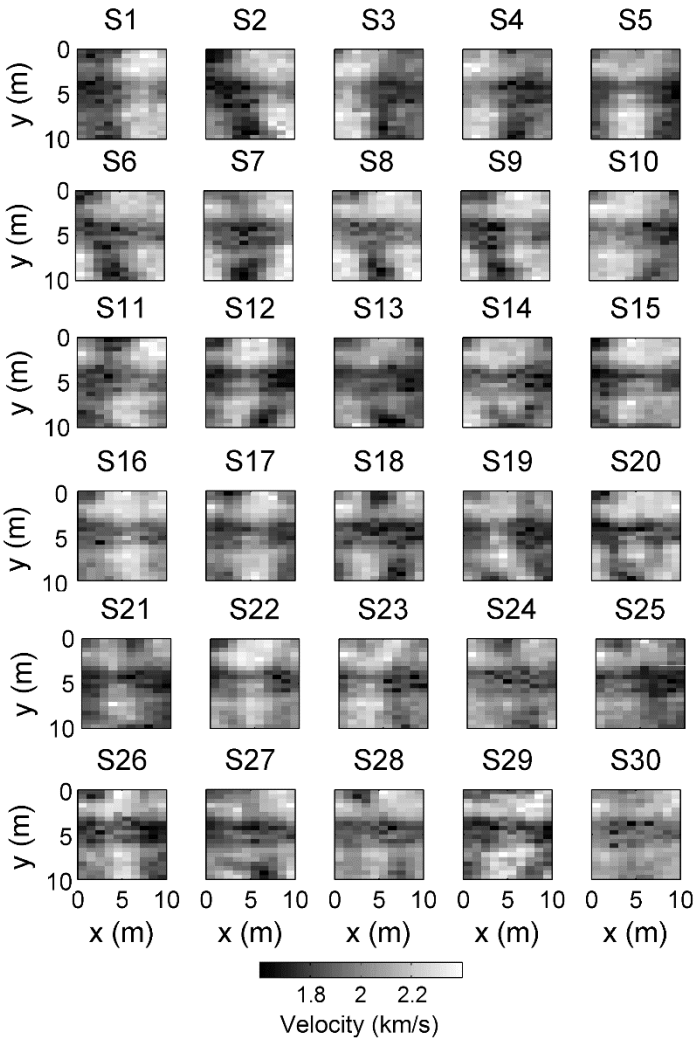

Figure 2-6: The same as in Figure 2-5 but for the seismic tomograms.

does not resemble ground truth. Since the seismic tomograms suffer increased ambiguities in regions that are partly well resolved by all radar tomograms, a combination of tomograms reconstructed from different datasets is desirable when striving to constrain predictions of a target parameter by geophysical tomography, which is in this case porosity.

2.4.3 Generation of Sparse Exploration Target Parameters

In our synthetic experiment we simulate measurements of porosity in boreholes, as they could be done in practice using borehole logging tools. In our database this results in 1D porosity profile data at the left and right model edge. The black lines in Figure 2-7 show the true porosity extracted from the porosity model in Figure 2-3b at the left and right model edges. However, realistic logging tools integrate over a certain sample

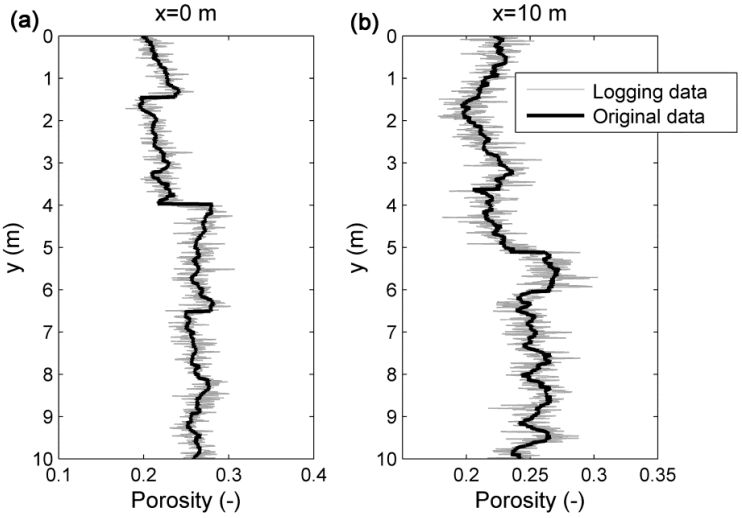

Figure 2-7: Sparse porosity borehole logging data acquired in the boreholes at the (a) left (x=0 m) and (b) right (x=10 m) model edges (see Figure 2-3). Original porosity represents the true information of the ground. Logging porosity represents the modelled response of a realistic borehole porosity logging probe.

volume. We model this by assuming an ellipsoidal sample volume with vertical extension of 50 cm and lateral extension of 25 cm (Knödel et al., 1997). The sensitivity decreases linearly with distance from the measurement point. This procedure results in simulated borehole logging information. To be more realistic, we add random Gaussian noise with 5 % relative error to simulate the contamination of borehole logging data by observational errors. The gray lines in Figure 2-7 show the simulated noisy porosity logging data acquired in the left and right borehole in our synthetic experiment.

2.4.4 Setting up the Artificial Neural Network

For assessing the effect of tomographic ambiguity in the prediction of our target parameters we repeatedly train the ANNs. For providing the training datasets comprising {(input, target)} tuples, one combination of a radar R^i and a seismic S^j tomogram with i = 1, 2, .., 30 and j = 1, 2, .., 30 at the positions of the boreholes forms the input training data, and 1D porosity information at the positions of the boreholes (Figure 2-7) forms the target training data for an ANN prediction model. The reason for considering pairs of two physically different tomograms is the presence of non-unique, facies-dependent parameter inter-relations in our study, which could not adequately addressed in the subsequent inference of 2D porosity scenarios if only one geophysical dataset would be available. At the position of a borehole 20 equally spaced radar and seismic velocity values are available over the depth range from 0 to 10 m.

For assessing the effect of noise contaminating the measured sparse target parameter on the prediction results and the suitability of the ANN performance parameter MSE for feature selection, we repeatedly run our methodology using the original porosity (OP) and the simulated logging porosity (LP) data (Figure 2-7) for training the ANN. This results in two different types of training dataset tuples for the 900 possible tomogram pairs which are RSOP = $\{(\mathrm{R}^{i}_{x,y}\mathrm{S}^{j}_{x,y}, \mathrm{OP}_{x,y,m})\}$ and RSLP = $\{(\mathrm{R}^{i}_{x,y}\mathrm{S}^{j}_{x,y}, \mathrm{LP}_{x,y,m})\}$ with $x \in (0, 1, \ldots, 10)$, $y \in (0.25, 0.75, .., 9.75)$, and m=1,2,..,40.

y defines the center values of half-meter depth intervals with constant radar and seismic velocities. m refers to the 40 equally spaced porosity values that are available per half-meter depth in the left and right boreholes, i.e., for the selected tomographic grid cell defined by y. We randomly divide our data for training, validation, and testing into subsets comprising 70%, 15%, and 15%, respectively. To ensure the utilization of a sufficiently complex ANN capable to offer a well-fitted prediction model we repeatedly train the ANNs for a given training dataset employing different numbers of neurons in the hidden layer. We test 3, 5, 10, 20 and 50 neurons in the hidden layer for the same training dataset. Prediction performance is measured by the MSE which will not be substantially lowered by further increasing the number of neurons in the hidden layer once the ANN offers sufficient complexity for linking the tomograms and the logging data. Figure 2-8 shows the MSE for all 900 tomograms combined with original and noisy logging data and

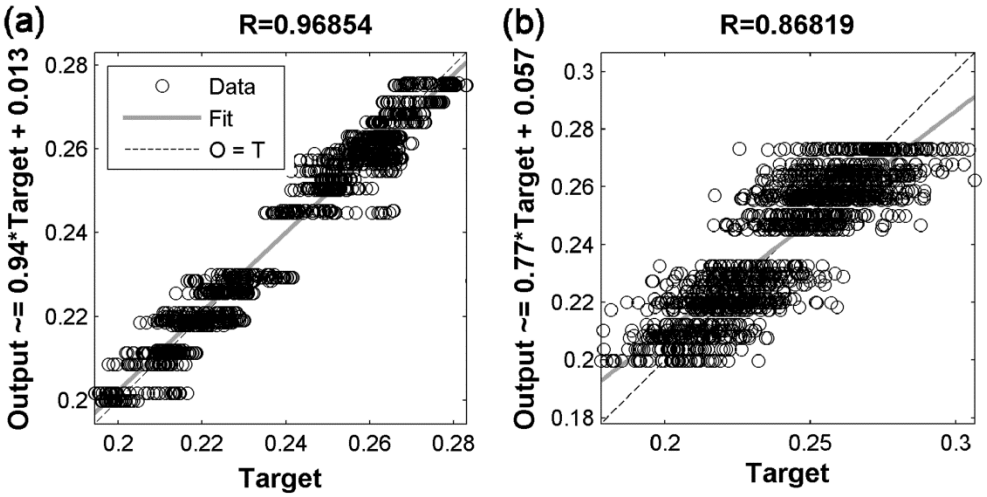

Figure 2-9: Regression results of the training procedure of the ANN when using R30 and S30 tomograms as input and the (a) original porosity (RSOP) and (b) logging porosity (RSLP) as output information. Note, only the radar and seismic velocity information of the left and right model edges has been used for training.

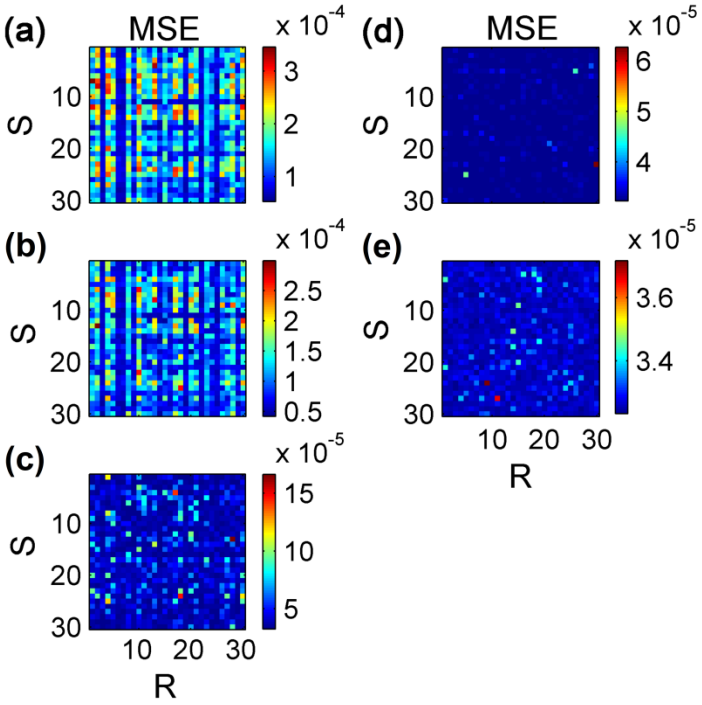

Figure 2-8: MSE from ANN training for all combinations of spatially continuous tomograms. (a)-(e) represent the results of using ANNs with 3,5,10,20 and 50 neurons, respectively.

different number of neurons. Increasing number of neurons allows generally for better ANN performance indicated by lowered MSE for increasing number of neurons. ANNs with only 3 and 5 neurons in the hidden layer show a rather systematic behavior clearly favoring distinct radar or seismic tomograms, which is indicated by stripy pattern in Figures 2-8a and b. For ANNs with 20 and 50 neurons, the stripy pattern of low MSE values is replaced by a rather random pattern. This indicates that the ANN is generally complex enough of fitting a prediction model for any radar-seismic tomogram combination. In turn, this may bear the risk of overfitting the tomogram pairs and the logging data beyond reasonably accuracy limits. Now the success or stopping of the training procedure seems to be of dominating influence on the size of the MSE. For prediction of our target parameter, we choose the solution of the ANNs run with 20 neurons in the hidden layer.

2.5 Results and Discussion

2.5.1 Probabilistic Prediction of 2D Porosity Distributions

Figure 2-9 shows the regression plots for ANNs trained with RSOP and RSLP datasets. Regression coefficients of RSOP=0.96 and RSLP=0.86 indicate a high accuracy of the prediction models found by the ANNs albeit the performance lowered for

the porosity data contaminated by volumetric integration and random noise. According to the regression coefficients our ANNs are for both target parameter datasets able achieving a satisfactory level of performance for our prediction task. Increasing noise components or non-Gaussian noise may result in further decreasing regression coefficients.

Figure 2-10 shows the results of the probabilistic spatially continuous prediction of porosity. We illustrate the prediction uncertainty by showing relative frequency information drawn from the 900 2D porosity scenarios calculated. Figure 2-10a shows the 2D porosity distribution predicted by ANNs using the RSOP training data. The prediction results at the borehole locations (left and right model edges) are highly accurate, since the ANNs have been trained at these positions. Between these two positions the prediction ranges for porosity are broader.

Due to the utilization of a synthetic database, we can evaluate the quality of the predicted porosity distribution. Figure 2-10b shows a comparison between predicted and realistic porosity of the subsurface extracted from the true porosity distribution (Figure 2-3b). The black lines in Figure 2-10b determine the minimum and maximum of the range of the realistic porosity in the region corresponding to the tomographic mesh cell. At the boreholes the predictions are near to reality, which are the positions for training the ANNs. Between the boreholes the predicted porosity ranges exceed but include those of the original porosity model (Figure 2-3b). In most regions, the original porosity ranges are coincident with high relative frequency values of predicted porosities. However, at some regions, (e.g., x=6.5m, y~1m) the original porosity is within the prediction range, but not coincident with high prediction frequency. In this area, most tomographic reconstructions of radar and seismic velocities do not match reality as depicted by the original models (cf. Figures 2-3, 2-5 and 2-6). These tomographic reconstruction errors propagate into the prediction of porosity scenarios. Since some tomographic models are close to reality in these regions, the predicted porosity range is broad enough to include the true porosity range. Figure 2-10b proves that ANNs are capable to offer high quality and accurate prediction models for predicting the porosity based on the radar and seismic tomograms as well as to transduce tomographic reconstruction ambiguity into the probabilistic inference of target parameter distributions. The large ranges of predicted porosity result

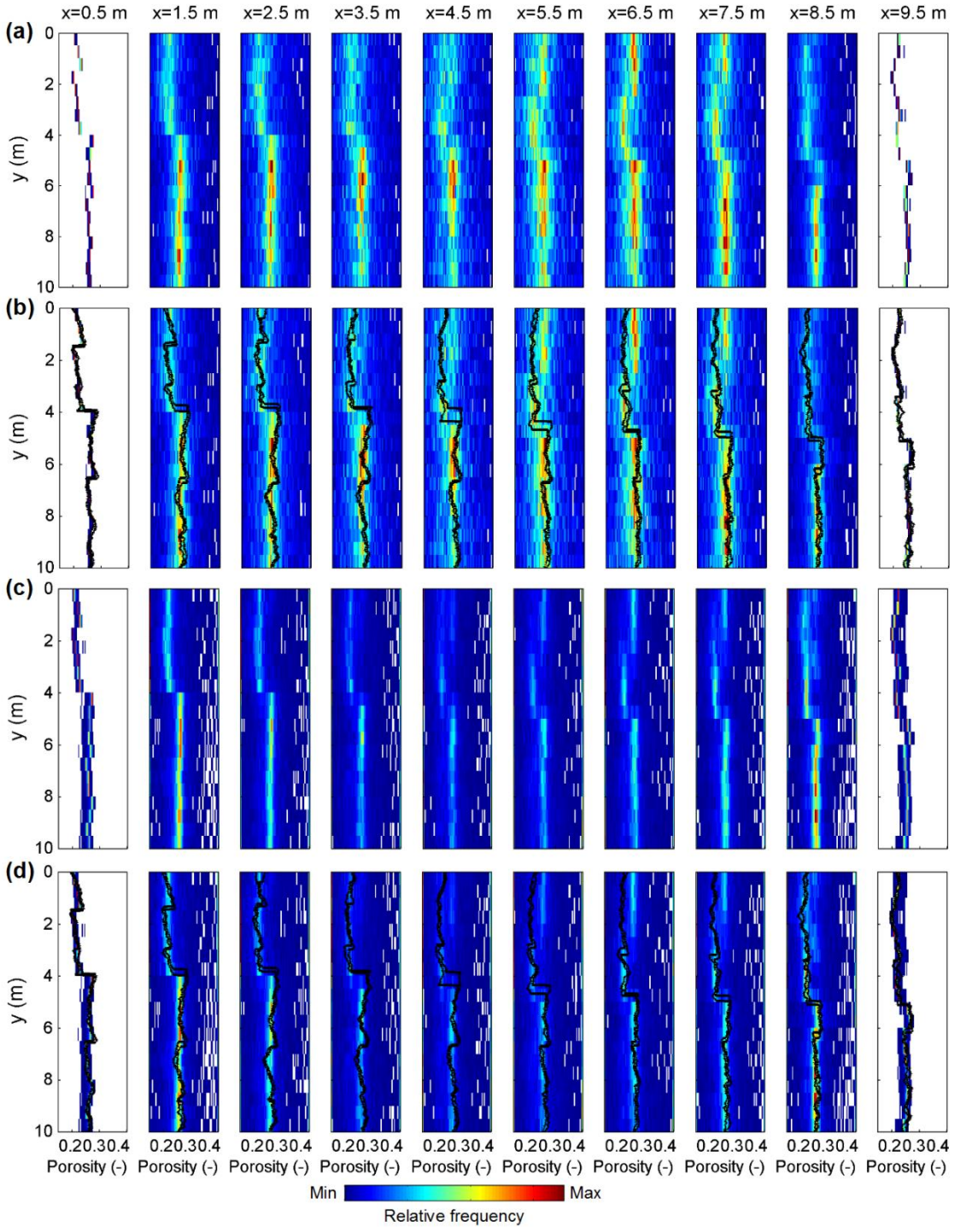

Figure 2-10: Prediction results of spatially continuous 2D porosity models based on the sparse logging data and 900 combinations of spatially continuous tomograms (Figure 2-5 and 6). (a) Relative frequency of porosity prediction from ANN models trained with original porosity. (b) The same as (a) but overlain by minimum and maximum range of true porosity of the ground (see Figure 2-3) shown by dashed black lines. (c) and (d) are analogue to (a) and (b) but for logging porosity instead of original porosity. Predicted porosities outside the displayed range are accumulated at the bins with lowest and highest porosities. They correspond to the ANN models trained with remaining high MSE.

from spatial resolution differences between the tomograms (0.5 m grid cell side lengths in vertical direction) and the logging data (1.25 cm sample distance). Since the ANN strives to learn a perfect prediction model, this may lead for some combinations of radar and seismic models with some logging readings to prediction models that are well fitted but physically not close to reality. When applying such prediction models error propagation may lead to rather extreme porosity values. This effect has also been observed when using a prediction technique based on fuzzy sets rather than ANNs (option 2 in Paasche submitted a). Such effects are inherent to data-driven prediction and inference approaches not taking the uncertainty of individual data samples or tomographic velocity values into account when learning the prediction model.

Figure 2-10c shows the predicted porosity distribution when using the simulated noisy logging data for ANN training. The prediction ranges are broader and less focused then those obtained for the original porosity data (Figure 2-10a). This finding is due to the propagation of an additional data error contributed by the logging data which superimposes with the tomographic ambiguity. In Figure 2-10d the range of the original porosity variability is overlain on the prediction results achieved for the noisy logging data. In regions where the tomograms resemble reality well the prediction result still allows for correct prediction of reasonable porosity values indicated by maximal relative frequency. The prediction result in Figure 2-10c illustrates what can be hoped to achieve when working with field data. Deterministic prediction approaches, or approaches ignoring the tomographic reconstruction ambiguity inherent to geophysical tomography will probably mislead the interpreter by producing prediction results affected by artefacts. In our example, such artefacts are likely to occur in regions where the original porosity range does not coincide with high relative frequencies of predicted porosity values. A first application of the proposed prediction technique to a field dataset confirms the potential of the suggested probabilistic prediction method (Asadi et al., 2016).

2.5.2 Ranking of Tomograms

We rank the radar and seismic tomograms according to the MSE achieved when using only 3 neurons in the hidden layer. We are aware that such a network is probably too simple to achieve good predictions, but Figure 2-8a shows, that some tomographic models can be systematically better linked to the target parameter than others. For example, the radar model R3 achieves always low MSE values regardless of the seismic tomograms it is combined with. These trends hold also for more complex networks (Figures. 2-8b and 2-8c) until the degree of complexity allows for learning input-output relations of very high complexity (Figures 2-8d and 2-8e). For each radar and seismic tomogram we calculate a mean MSE value based on the information in Figure 2-8a. We sort the radar and seismic tomograms according to their MSE (Figure 2-11). Analogue we order the radar and seismic tomograms when using the RSLP input for training (Figure 2-11). Ordering of radar and seismic tomograms differs for the original and noisy logging data indicating that the observational errors of the logging data influence the ANN performance.

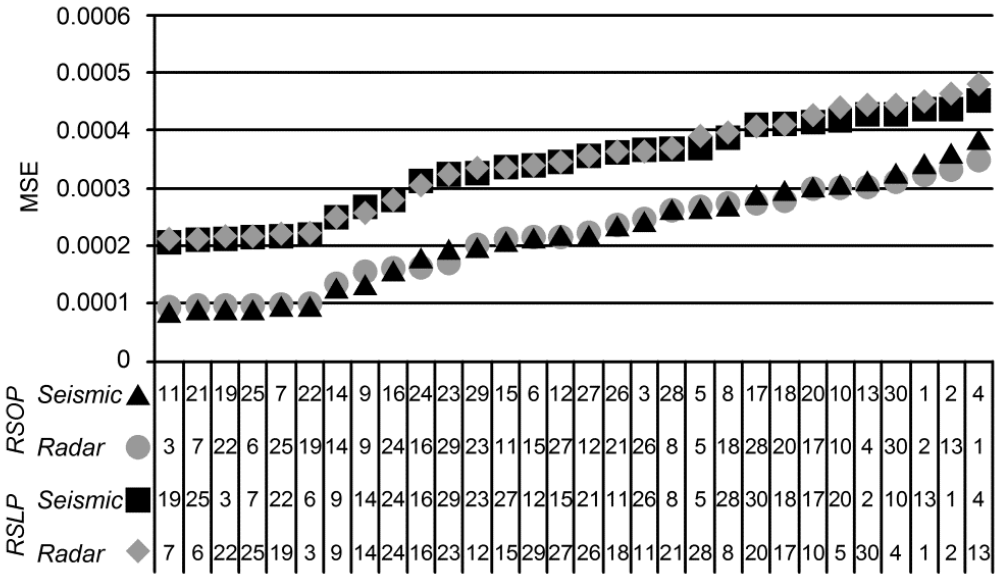

Figure 2-11: Ranking of tomograms (Figure 2-5 and 2-6) according to their relationships with reality. RSOP rank radar and seismic models according to their combination with original porosity for training the ANN models. Likewise RSLP rank radar and seismic velocity tomograms according to their combination with logging porosity for training the ANN models. This ranking has been outcome from ANN with three neurons in the hidden layer.

In our synthetic database the true radar and seismic velocity models are known and we can benchmark the ranking based on the ANN performance by directly comparing the tomographically reconstructed tomograms (Figures 2-5 and 2-6) with the true radar and seismic velocity distributions (Figures 2-3b and 2-3c). Figure 2-12 shows summed squared differences between ground truth information and all radar and seismic tomograms and for both training datasets RSOP and RSLP (The fine sampling of ground truth results in 40 x 80 squared differences per tomographic grid cell, which are summed). Summed squared differences are calculated for the positions of the boreholes (left and right model edges corresponding to the locations used for training the ANNs) and the entire 2D model area. The ordering of the feature number along the abscissa of the plots

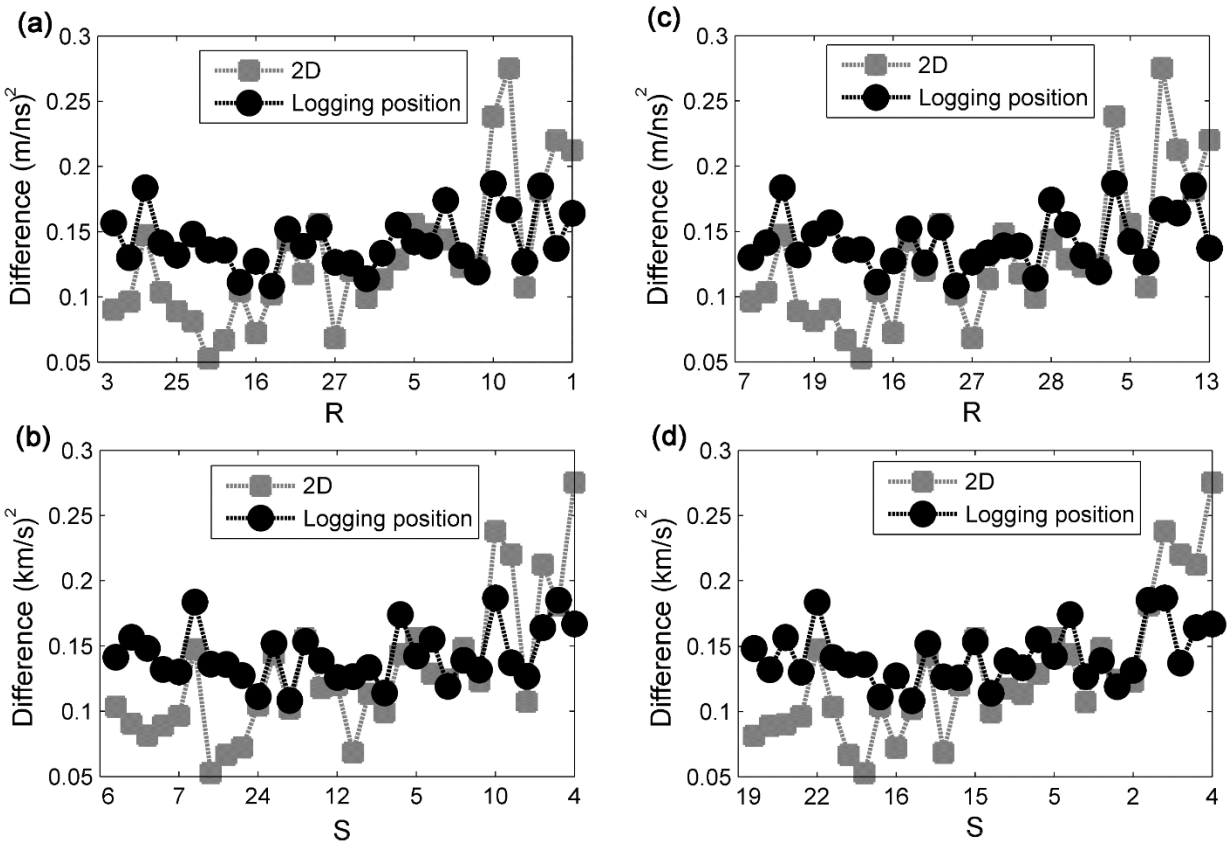

Figure 2-12: Summed squared differences of 30 tomographic radar (R) and seismic (S) velocity tomograms from the true radar and seismic velocity models (Figures 2-3c, and 2-3d). The ordering of the model number (abscissa) corresponds to those proposed by the ANNs trained with three neurons in the hidden layer (Figure 2-11). The black circles illustrate differences at the left and right edges (logging positions) of the tomograms. The gray rectangles illustrate the difference of the entire 2D area. (a) and (b) correspond to the ANN trained with original porosity. (c) and (d) correspond to the ANN trained with logging porosity.

in Figure 2-12 corresponds to the ranking according to the ANNs performance. When analyzing the results in Figure 2-12 it is obvious that the ANN performance does not allow for clear identification of the best tomographic models. For example, R22, which has been ranked 3rd by the ANN performance, offers a relatively poor reconstruction of the true radar velocity model, both at the positions of the boreholes as well as for the entire 2D area (Figure 2-12a). Additionally, some models match reality well at the positions of the boreholes but are poor reconstructions of the true 2D velocity distributions. Interestingly, this case occurs more frequently for models ranked low by the ANN performance. Generally, models ranked low by the ANN performance suffer increased chances to be indeed poor reconstructions of the real 2D velocity models, albeit some exceptions may exist. Tomograms ranked in the upper half by the ANN performance have increased chances to be slightly more realistic tomographic reconstructions of the reality than others. A distinct identification of the best tomogram cannot be made based on the ANN performance.

When striving to recursively learn the prediction by individually ranking radar and seismic tomograms the computational effort will clearly increase. However, tomogram ranking based on ANN performance does not allow for certain reduction of the effects of poor tomographic reconstructions, since the prediction will be dominated by a very limited number of tomographic models, but a certain risk remains that some poor tomograms will be given high weights. This could probably result in even worse prediction than achieved when using the entire tomographic ensemble equally. Hence, we judge the possibility to rank mathematically equivalent tomograms based on their linkage to sparsely measured target parameters as quantitatively limited and consequently we do not repeat our 2D porosity prediction using recursively learning ANNs. This finding may also have consequences for geophysical inversion strategies considering logging data as constraints during the tomographic model reconstruction. Found geophysical tomograms will likely match the logging data where they are available. However, if the lateral spacing of the logging data clearly exceeds the lateral resolution of the tomographic data it is not granted that the found model will outperform inversion results in the entire tomographic reconstruction area achieved without considering the logging data in the inversion procedure.

2.6 Conclusions

We have employed static two-layer feed forward ANNs for 2D probabilistic prediction of sparsely measured Earth properties constrained by ill-posed geophysical tomographic imaging. By using 900 pairs of collocated, physically different radar and seismic tomograms, which fit the underlying datasets equally we were able to transduce tomographic reconstruction ambiguity into the prediction of a target parameter, which is of higher relevance to hydrologic and engineering exploration tasks than the tomographically imaged parameters. The prediction performance of the methodology has been illustrated using a realistic but synthetic database allowing for optimal performance evaluation of the suggested methodology. Prediction performance was found to be excellent, and can be applied to any combination of geophysical tomograms and target parameters since at no point critical assumptions about the involved parameters or the expected relations between the considered datasets and parameters are made. It is even applicable to datasets where different facies dependent petrophysical relations are present. In such situations, it is essential to consider at least two physically different tomograms when constraining the prediction of the target parameter. When combining our approach with fully non-linear (globally searching) geophysical tomographic imaging this methodology can deliver objective and purely data-driven probabilistic predictions of target parameter distributions, which are essentially required when striving to assess, quantify and minimize risks in subsurface exploration and utilization.

We evaluated, whether the performance of the ANNs training, measured by an MSE, can be used to rank the equivalent geophysical tomograms. Fundamental idea of this approach is that tomograms as well as sparse information about an exploration target parameter are images of the same reality and must therefore be compliant. In our synthetic database we could analyze this question which would be practically impossible when working with field data. A rather qualitative statement about the closeness of the tomograms to reality can be made based on the ranking results achieved by ANN training performance, i.e., tomograms ranked low suffer an increased risk of being poor reconstructions of reality. However, outliers from this rule may exist and therefore question the benefits from utilization of recurrent ANNs striving to learn which tomograms may be particularly useful for prediction based on the available database. Such approach

would build the prediction of target parameter distributions on a few models of high importance, but facing the risk that eventually a poor tomographic model will be considered with high weights, which leaves doubts on the chances to achieve better predictions when using recurrent ANNs instead of the simple feed-forward ANNs used in this study.

Chapter 3

Spatially Continuous Probabilistic Prediction of Sparsely Measured Ground Properties Constrained by ill-posed Tomographic Imaging Considering Data Uncertainty and Resolution

Abduljabbar Asadi, Peter Dietrich, and Hendrik Paasche
Manuscript published in Geophysics, 2017

3.1 Abstract

Probabilistic prediction of 2D or 3D distributions of sparsely measured borehole or direct push logging data can contribute to solving hydrological, petroleum, or engineering exploration tasks. We employ and improve a recently developed workflow constrained by ill-posed geophysical tomography to achieve 2D probabilistic predictions of geotechnical exploration target parameters that could only be measured by 1D borehole or direct push logging. We use artificial neural networks (ANNs) to find the optimal prediction models between ensembles of equivalent geophysical tomograms and the sparsely measured logging data. During the training phase of ANNs we consider four different training strategies taking into account the logging data uncertainty and geophysical tomographic ambiguity to avoid data overfitting of the ANNs. Thus, we successfully transform the logging data uncertainty and geophysical tomographic reconstruction ambiguity as well as differences in spatial resolution of logging and tomographic models into the probabilistic 2D prediction of our target parameters in a data-driven manner, which allows application of our methodology to any combination of geophysical tomograms and hydrologic, petroleum or engineering target parameters solely measured in boreholes. To illustrate our workflow, we use an available field dataset collected at a field site South of

Berlin, Germany, to characterize near-subsurface sedimentary deposits. In this example we employ cross-borehole tomographic radar-wave velocity, P-wave velocity, and S-wave velocity models to constrain the prediction of tip resistance, sleeve friction, and dielectric permittivity as target parameters.

Keywords: Probabilistic prediction, data uncertainty, geophysical tomograms, artificial neural networks.

3.2 Introduction

Geophysical tomographic datasets offer valuable information about the internal composition of the ground in two or three dimensions (e.g., Moorkamp et al., 2016). Such datasets uniquely image physical parameter variations, e.g., radar-wave velocity, seismic P-wave velocity, or S-wave velocity, in a spatially continuous manner. For solving many near-surface hydrological and engineering exploration tasks, when a detailed characterization of the subsurface is required, utilization of geophysical tomographic datasets is essential. Traditionally, hydrological or engineering exploration target parameters (e.g., tip resistance, sleeve friction, dielectric permittivity) are measured by 1D exploration techniques, such as borehole and direct push logging (e.g., Lunne et al., 1997; Rubin and Hubbard, 2005). Estimation of a 2D or 3D image of such parameters with traditional geotechnical or hydrological exploration techniques is costly and time consuming. Based on the relation between geophysical tomograms and measured target parameters at the position of the boreholes, the 2D or 3D geophysical tomograms can be converted into a 2D or 3D image of the desired target parameters. There are numerous examples, where geophysical tomography is used to estimate 2D or 3D distributions of target parameters for geotechnical ground characterization (Yamamoto, 2001; Angioni et al., 2003; Rumpf and Tronicke, 2014), hydrological characterization (Hubbard et al., 2001; Binley et al., 2001; Tronicke and Holliger, 2005; Paasche et al., 2006; Dubreil-Boisclair et al., 2011; Ruggeri et al., 2013).

Unfortunately, geophysical tomographic datasets suffer ambiguity due to limited number of observations and measurement errors. Traditionally, deterministic tomographic reconstruction techniques relying on regularized local-search optimization (e.g., Aster et al., 2005) are employed to generate a single geophysical tomographic model. Such approaches do not allow for realistic and quantitative ambiguity appraisal inherent to the model generation. Recently fully nonlinear optimization methods (Sen and Stoffa, 2013), have been employed to explore the model space in more detail and reconstruct ensembles of geophysical tomograms fitting the underlying datasets equally well. Thus, the tomographic ambiguity is represented by a number of equally plausible

geophysical tomograms. These tomogram ensembles can be used to assess the tomographic ambiguity in a realistic manner.

While all resultant tomograms are mathematically equivalent answers to the geophysical tomographic reconstruction problem, they may resemble the internal composition of the ground to variable degrees. Due to the spatial resolution of geophysical tomographic datasets tomograms offer gross averaged physical quantities (on the scale of meters or tens of meters), while measurements of the target parameters offer spatial resolution of a few centimetres or decimetres, but only in the vertical direction. Additionally, like all experimental datasets, measurements of the target parameters by borehole or direct push logging are affected by a variable amount of measurement errors. Therefore, the most important challenge when using geophysical tomographic datasets in hydrological or engineering exploration is to link the 2D or 3D physical tomograms with the target parameters of interest to predict a spatially continuous model of the target parameters. Consequently, for realistic predictions tomographic ambiguity, logging data errors and the difference in spatial resolution must be taken into account.

Numerous approaches are available to link geophysical tomograms to hydrological or engineering target parameters. Traditional techniques rely exclusively on geophysical data for quantitative estimation of the target parameters. This group includes diverse empirical, theoretical or semi-empirical deterministic transfer functions (Archie, 1942; Gassmann, 1951; Wyllie et al., 1956; Topp et al., 1980; Yamamoto, 2001; Angioni et al., 2003) to convert one physical parameter into the desired hydraulic or engineering target parameters. Unfortunately the relations between physical and exploration target parameters are often non-linear, non-unique, and usually not exactly known (Schön, 1998). Classical transfer functions cannot cope with non-uniqueness in the parameter relations and require knowledge about the relations between the available physical and the desired target parameters across different scales.

Recently, statistical or geo-statistical frameworks have been proposed which allow for improved incorporation of uncertain and non-unique parameter relations based on statistical analysis methods, e.g., Bayesian inference (Ezzedine et al., 1999; Hubbard et al., 2001; Chen et al., 2001; Bosch et al., 2010; Boisclair et al., 2011; Ruggeri et al., 2013),

fuzzy systems (Paasche et al., 2006), or conditional stochastic simulations (Tronicke and Holliger, 2005; Dafflon et al., 2009). Usually, they require some measured information about the target parameter, and link deterministically derived geophysical tomograms with the target parameter thus not incorporating tomographic reconstruction ambiguity and logging data errors in their results.

Very recently, ensembles of equivalent tomograms have been taken into account by Rumpf and Tronicke (2014) and Asadi et al. (2016) for constraining the probabilistic inference of spatially continuous 2D predictions of exploration target parameter distributions. Rumpf and Tronicke (2014) employ Alternating Conditional Expectation (Breiman and Friedman, 1985) to link ensembles of 125 radar-wave velocity, seismic P-wave velocity and S-wave velocity tomograms with sparsely measured exploration target parameters, i.e., sleeve friction and effective grain size. In their tomographic reconstruction they rely on the concept of a layered ground and illustrate their prediction uncertainty by mean and median values in combination with percentile ranges. They do not consider the measurement errors of logging data in their prediction model.

Asadi et al., (2016) show a methodology for 2D probabilistic prediction of target parameters based on ensembles of radar-wave velocity and seismic velocity tomograms, and two layer feed-forward artificial neural networks (ANN; Hornik, 1991; Van der Baan and Jutten, 2000). Based on the synthetic dataset their results show that ANNs are able to determine very well the unknown relation between geophysical tomograms and exploration target parameters. Logging data errors are not considered in their prediction model.

In this paper we present the first application of the approach of Asadi et al. (2016) to measured datasets recorded by Linder et al. (2010). We show the probabilistic prediction of 2D tip resistance, sleeve friction, and dielectric permittivity as target parameters based on ensembles of equivalent radar-wave velocity, seismic P-wave velocity, and S-wave velocity tomograms. Furthermore, we extend the approach of Asadi et al. (2016) in a way that also estimated or, if available, measured logging data errors can be considered in the 2D probabilistic target parameter inference as well as difference in the spatial resolution of the tomograms and the logging data.

3.3 Methodology

3.3.1 Artificial Neural Networks (ANNs)

ANNs are powerful and well-studied Machine Learning tools for finding non-linear prediction models (Haykin, 2008; Ban and Chang, 2013). They have been applied to a variety of problems in the geophysical domain particularly in processing of seismic reflection data (Van der Baan and Jutten, 2000; Poulton, 2002; Leite and de Souza Filho, 2009; Leite and Vidal, 2011). Asadi et al., (2016) used feed-forward ANN for porosity prediction based on attributes derived from radar and seismic velocity tomography, and borehole logging data. In this paper we use and improve the approach of Asadi et al., (2016) using a two layer feed-forward ANN (Figure 3-1) for geophysical parameter prediction. Our ANN is composed of interconnected neurons placed in different layers known as input, hidden and output layers. Neurons are the processing units of an ANN and their functionality is related to the layer which they are participating in. Any connection between neurons is evaluated by a weight coefficient w which determines the importance of this connection in the ANN.

Neurons in the input layer assay to prepare a vector of input data (e.g., [a1, a2, ..., an], which are here radar-wave velocity, P-wave velocity, and S-wave velocity. The operation of the hidden layer is based on sets of input information, actual weight coefficients of inputs, an activation function, and a bias parameter. The activation function

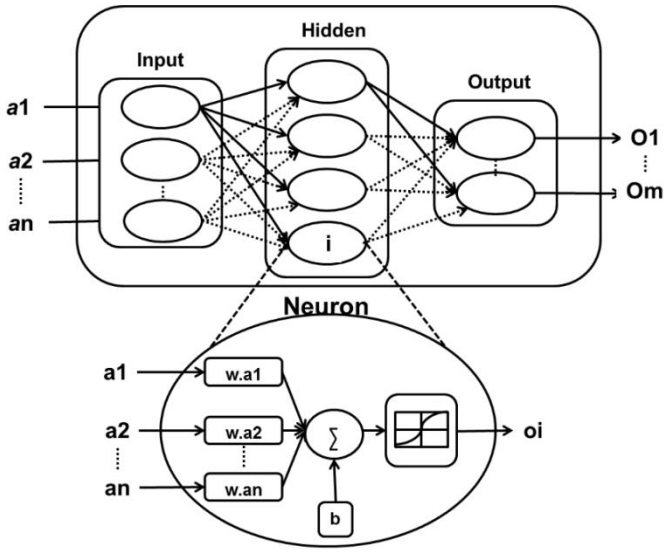

Figure 3-1: Structure of artificial neural networks (ANNs). ANNs are consisting of three interconnected layers. The input layer prepares data for feeding the ANN. The operation of hidden layer is based on sets of input information, weights of inputs w and bias parameter b. Neurons in this layer form a feedforward network with sigmoid formation. The operation of the output layer has been determined by the hidden layer and is connected to the results of the ANN training step.

defines the output of a neuron given an input or set of inputs. Different activation functions exist, e.g., linear, sigmoid or Gaussian. A combination of two layer feed forward ANN with sigmoid function in the hidden layer, and linear function in the output layer can be trained to approximate any function (Beale et al., 1992). The sigmoid function for input d is defined by

$$sig(d) = \frac{1}{1+e^{-d}} \quad . \tag{3-1}$$

The result of *sig*(d) is a number between 0 and 1, and it acts as transfer functions in each neuron of the hidden layer. The output oi of the i^{th} neuron in the hidden layer is

$$o_i = sig\left(\sum_{j=1}^{n} w_{ij} * a_j + b\right) . \tag{3-2}$$

n is the number of observations in the vector of input data [a1, a2, ..., an], and b is the bias parameter that shifts (together with all w_{ij}) the activation function (in this case sigmoid) to the left or right for finding the best fit to the target parameters. In this paper tip resistance, sleeve friction, and dielectric permittivity are our target parameters. The operation of the output layer is determined by the hidden layer and is connected to the results of the ANN in the training phase. A linear function acts as transfer function in the output layer for preparing the results of ANNs.

Three steps are necessary when employing ANN: training or learning, validation, and testing. During the training process based on minimizing the performance parameters ANNs try to discover the best weight coefficient of each connection to find the optimal fit between inputs and outputs of the ANN. In this paper we use two different performance measures, mean squared error (MSE) and weighted mean squared error (WMSE). During training the performance measure is computed in order to evaluate the accuracy of the trained ANN. If $\{(a_1,t_1),(a_2,t_2),\ldots,(a_N,t_N)\}$ be a set of training tuples, where $a_i \subset A$ a vector of input attributes, and $t_i \subset T$ a vector of target parameter, the MSE and WMSE are defined as

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(o_i - t_i)^2 \tag{3-3}$$

and

$$\text{WMSE} = \frac{1}{N}\sum_{i=1}^{N} e_i \ (o_i - t_i)^2 \quad . \tag{3-4}$$

N is the number of tuples in the training dataset (Beale et al.,1992). e_i determines the weights of the related training tuples for ANN and is mathematically considered to be the inverse of the square of the standard deviations of measurement errors (e.g., Tarantola, 1978). When using the MSE, logging errors, and tomographic uncertainty, i.e., model parameter or grid cell uncertainty, are ignored during the training step of the ANNs. If measurement errors shall be considered during the training step of the ANNs the MSE must be replaced with the WMSE. In this case we define e*i* by cumulated relative errors of the logging data and the tomographic uncertainty. This allows ANNs to prevent over-fitting input and target parameters, i.e., by adjusting the ANN to a degree where it also explains even the error present in the training datasets. Our choice here deviates from mathematical theory. However, in practice standard deviations and the assumption of uncorrelated noise may be error descriptions of limited representativeness in cases were residuals or observations are non-normally distributed. Here, conservative but robust estimates, e.g., ranges, could practically replace standard deviations albeit resulting in usually larger estimates of relative errors. Since we follow such a conservative error estimation strategy here we do not square the already rather large relative errors before using them in equation 3-4. However, if preferred, our approach can easily be adjusted to classical statistical theory, e.g., by using standard deviation, or other robust error measures, such as quartile ranges.

Note, since the true relationships between physical parameters imaged by geophysical tomography and the logging data providing information about the target parameter are scale dependent, spatially variable, related to measurement setups (e.g., static vs dynamic, or frequency dependencies) and usually non-unique and unknown, the prediction models learned by the ANNs are inherently site specific. It is not possible to quantify site and data-specific effects of scale differences, or spatial resolution differences on the learned prediction model. This is the reason why we do not recommend transferring a prediction model learned by an ANN at a site to other datasets acquired at different field sites, even if the local geology is generally comparable.

3.4 Processing Flow

Figure 3-2 outlines the processing flow when working towards probabilistic prediction of 2D tip resistance, sleeve friction, and dielectric permittivity constrained by ensembles of radar-wave velocity, P-wave velocity, and S-wave velocity tomograms. Two types of data imaging reality into sets of observations are available. The first type is crosshole tomographic travel-time data acquired between two adjacent boreholes. Three different geophysical travel-time datasets are available differing in their source energy, i.e., by exciting electromagnetic radar-waves, seismic P-waves or S-waves. The second type is measured in 1D by direct push probes as exploration target parameters. The measured quantities are tip resistance, sleeve friction, and dielectric permittivity, which serve as sparse information about our target parameters. For extracting the ground material information from the travel-time dataset we apply a global-search inversion approach (Paasche, 2015). This results in ensembles of q 2D velocity tomograms for the radar-wave travel-time, P-wave travel-time, and S-wave travel-time datasets, respectively. Each of the q tomograms fits the underlying traveltime dataset equally well,

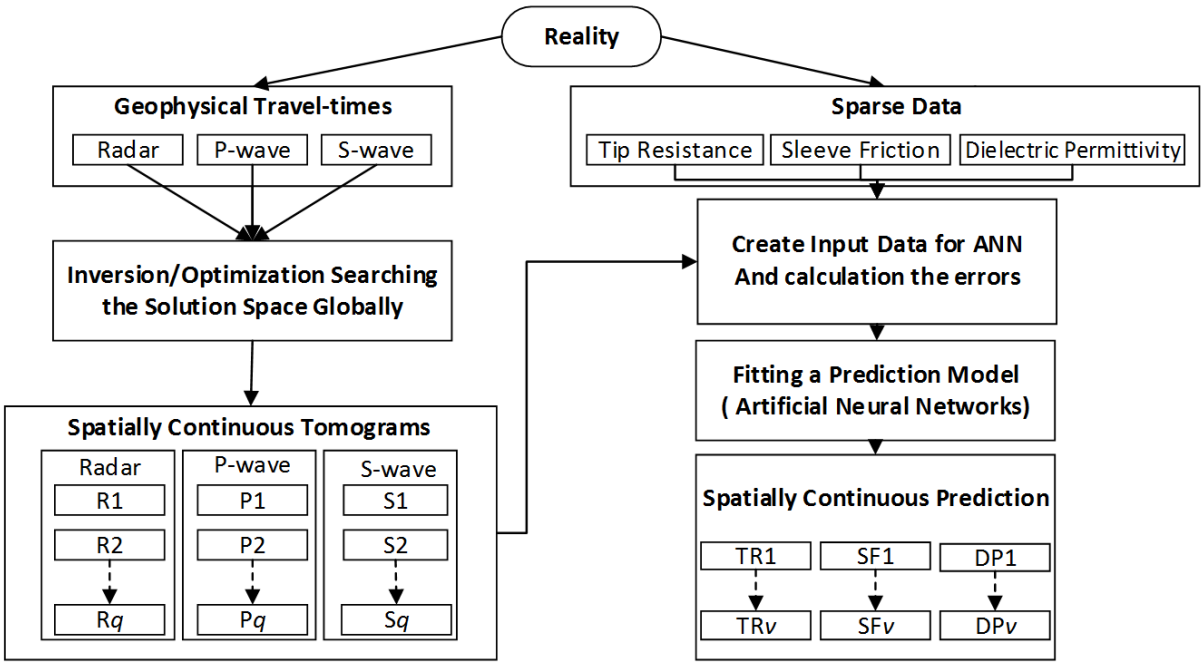

Figure 3-2: Processing workflow to probabilistically predicting spatially continuous models for sparsely measured target parameters and geophysical tomograms achieved from fully non-linear inversion. Based on the training strategy, v determines the number of prediction models resulted from ANNs.

albeit the imaged velocity distribution may be different for each tomogram. The ensemble of tomograms fitting the underlying dataset equally well provides discrete information about the ambiguity of the tomographic reconstruction problem.

For achieving probabilistic prediction models of our exploration target parameters, we link the available tomograms to the sparse target parameters employing two layer feed-forward ANNs. Predictions are independently and repeatedly done using the MSE and WMSE as performance measure for training the ANNs. To ensure an equal contribution of every tomogram, all radar-wave velocity, P-wave velocity, and S-wave velocity tomograms are combined with each other to achieve different sets of the input data for training the ANNs. Per tomographic grid cell at the measurement position of 1D target parameters, u observations have been measured for each target parameter. Based on the q radar-wave velocity, q P-wave velocity, and q S-wave velocity tomograms we have q^3 different combinations of input scenarios to our ANN that are all linked to the u observations of a target parameter. This can give us v different scenarios according to the combination strategy of inputs with targets. We train the ANNs individually for every input scenario and achieve v prediction models, which can be used to generate 2D probabilistic distribution scenarios of tip resistance, sleeve friction, and dielectric permittivity simultaneously.

3.5 The Database

3.5.1 The Field Site

We use a dataset measured by Linder et al., (2010), which has been acquired on a field site located 30 km south of Berlin sustained by the German Federal Institute for Materials Research and Testing (BAM) (Niederleithinger, 2009). Three PVC-cased boreholes with an inner diameter of 80 mm have been used reaching down to depths of approximately 17 m. The local geology is primarily composed of glacial and glaciofluvial sands and gravels. Drillings show that the near-subsurface in this area consists of thin top soil layer followed by layers of medium, partly silty sands and fine gravels with interbedded thin layers of medium gravel and organic material in depth below ~8 m. The ground water table was approximately 3 m below surface at the time of measurement.

The first borehole used in the experiment at x = 0 m is considered to define the left edge of the tomographic plane, the second and third boreholes are placed at 5.01 m and 10.96 m distance from the first borehole when following the tomographic plane. A magnetic deviation logging tool was used for measuring the borehole trajectory, which indicated that the borehole trajectories were deviating no more than 4 cm from verticality at 16 m depth. Thus, when inverting the travel times from crosshole experiments we consider vertical boreholes.

3.6 Data Acquisition

3.6.1 Tomography

Georadar data were acquired operating the borehole receiver antenna in borehole two located at x = 5.01 m in our tomographic plane. The transmitter borehole antenna was operated in boreholes 1 and 3. Nominal centre frequency of the transmitted signal was 100 MHz. The acquisition parameter and signal characteristic, e.g., source and receiver spacing, sample interval, dominant frequency, and dominant wave length were 0.25 m, 0.04 ns, ~ 60 MHz, and ~ 1 m, respectively. First signal onsets in the recorded shot gathers have been determined using an automated picker (Tronicke, 2007) and the quality of the picked travel-times has been controlled manually.

P-wave seismic energy was generated in the water saturated zone by a conventional sparker source and a 24-channel hydrophone string was used for recording seismic energy, i.e., the wave train at the receivers. Source and receiver spacing were 0.25 m covering a depth range of 4.5 m to 16 m. Dominant wave length of ~2.5 m relay on dominant frequency of ~750 Hz and average velocity value of the respective parameters. Picking consistency and data quality have been checked using source-receiver pick images (Harris et al., 1995).

S-wave seismic energy was achieved by an electrodynamic borehole impactor source which generates horizontally polarized SH-waves. At every shot location shots with 180˚ opposite excitation direction were fired to ensure later on a reliable identification of S-wave first on sets. Shear wave energy was recorded using two five-component borehole geophones. The acquisition parameter and signal characteristic, e.g., source

and receiver spacing, sample interval, dominant frequency, and dominant wave length for the S-wave data were 0.5 m, 0.021 ms, ~ 200 Hz, and ~ 1.3 m, respectively. Similar to the P-wave after checking quality and consistency the S-wave travel time was obtained for inversion (Linder et al., 2010).

3.6.2 Sparse Logging Data

Our methodology requires us to have at least sparse knowledge about the exploration target parameters, e.g., in a borehole or at direct push positions within the tomographic model area. Tip resistance, sleeve friction, and dielectric permittivity of the ground are our target parameters. For measuring the target parameters direct push experiments were carried out at two selected locations between the boreholes but in the inter-borehole planes. For measuring tip resistance and sleeve friction cone penetration test have been done. A standard piezocone probe with 4.4 cm diameters, controlled force and constant speed of 2 cm/s was pushed into the underground. Readings were recorded with 1 cm vertical spacing. Figures 3-3a, and 3-3b show the measured tip resistance and sleeve friction, respectively. The black line shows the measured target parameters at x=2.75 m, whilst the gray line shows the measured target parameters at x=8.0 m. Measured data show similar dynamics in the depth but an offset exist between them. The higher tip resistance and sleeve friction (Figure 3-3a, and 3-3b) are observed beloow y=7.0 m, and above y=14 m. Furthermore, at a frequency of 30 MHz the dielectric

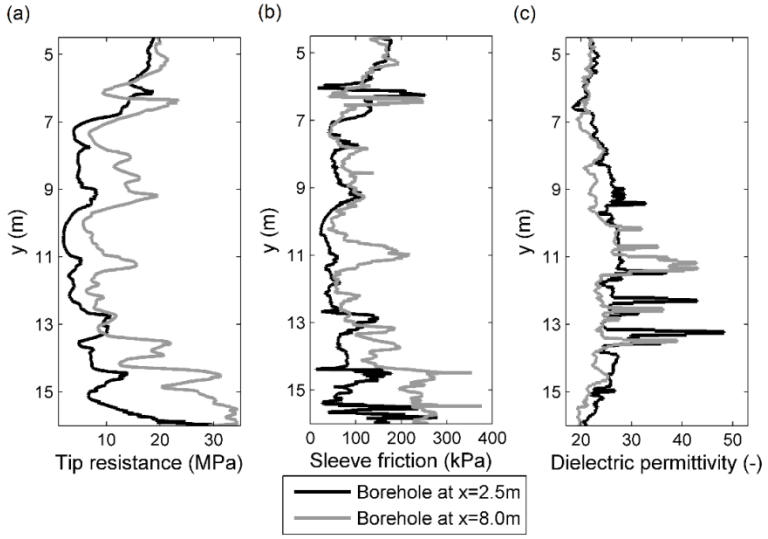

Figure 3-3: Target parameter logging data acquired by direct push technology at x=2.75 m and x=8.0 m for (a) tip resistance, (b) sleeve friction, and (c) dielectric permittivity.

permittivity was measured using a soil moisture probe (Linder et al., 2010). Figure 3-3c shows the dielectric permittivity at x= 2.75 m and x= 8.0 m.

3.7 Processing

3.7.1 Tomography

For 2D tomographic reconstruction of radar-wave velocity, P-wave velocity, and S-wave velocity models we use an inversion approach searching the model space globally (Paasche, 2015). The inversion performance parameter used to measure the misfit between the data and the forward model response is the root mean squared (rms) error. We reinitialized the inversion 30 times to find 30 equivalent and independent geophysical tomograms for each dataset. All found tomograms fit the underlying dataset equally well. The rms errors of the 30 final radar-wave velocity, P-wave velocity, and S-wave velocity models vary between 0.98 and 1.0 ns, 0.04181 and 0.04199 ms, 0.886 and 0.899 ms, respectively.

The final ensemble of 30 equivalent radar-wave velocity tomograms is shown in Figure 3-4. All models show regions of lower velocities in the center of the model reconstruction area. High velocities are found at depth around 16 m and between 4.5 and

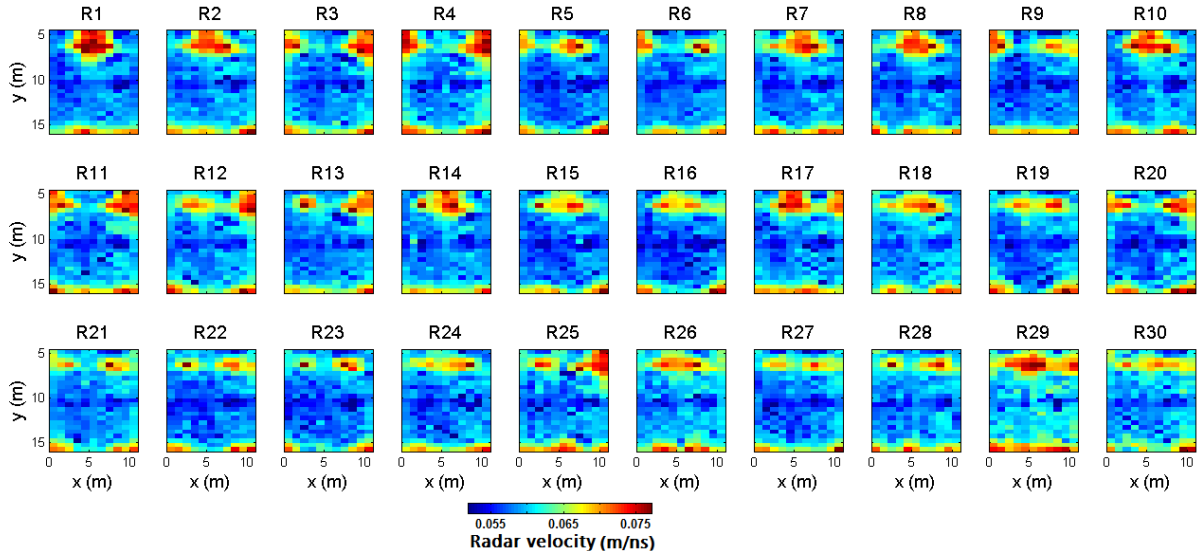

Figure 3-4: 30 tomographic reconstructions of radar-wave propagation velocity achieved by fully non-linear (global-search) inversion. The 30 tomograms are achieved by independent inversion runs and fit the underling tomographic dataset equally well.

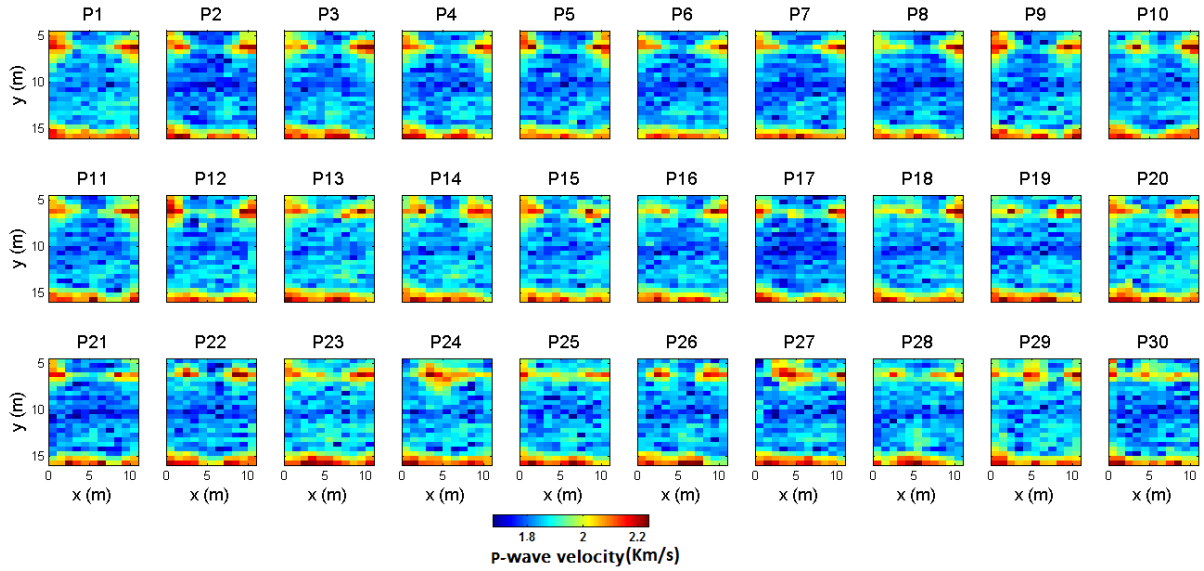

Figure 3- 5: The same as in Figure 3- 4 but for P-wave velocity.

8 m depth (e.g., models R12 and R29 in Figure 3-4). The solutions diverge increasingly towards the upper and lower model edge, whereas the velocities in the more central parts of the tomograms are well defined. Divergence is maximal for model parameters in the vicinity of the boreholes, which reflects the typical imaging capabilities of a crosshole tomographic dataset.

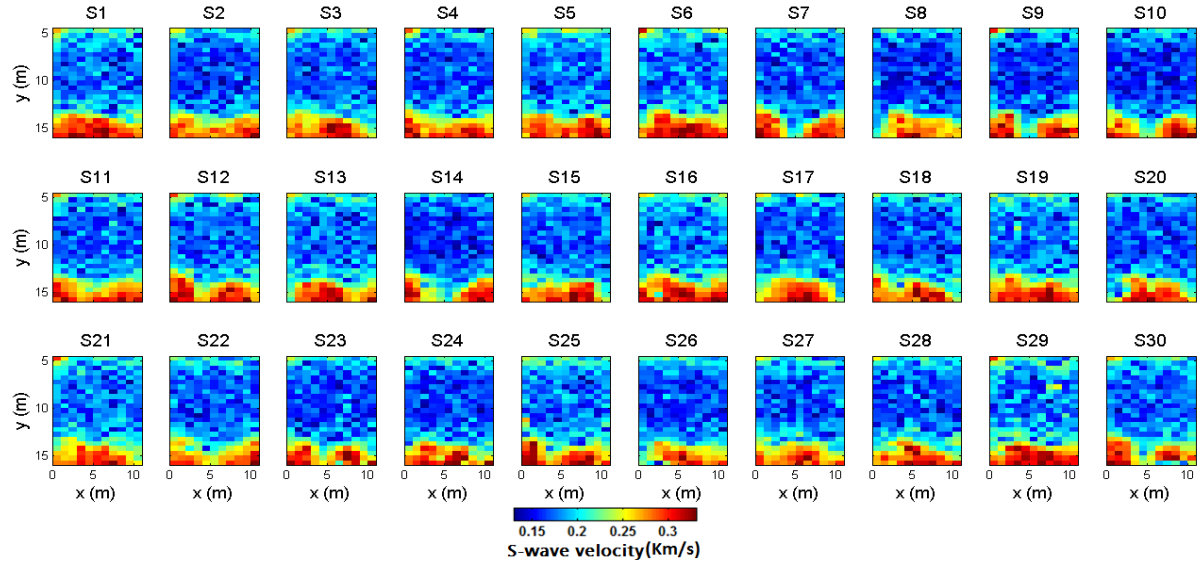

Figure 3- 6: The same as in Figure 3- 4 but for S-wave velocity.

Figure 3-5 shows the final ensemble of the 30 equivalent seismic P-wave velocity tomograms. The high velocity in the bottom part and low velocity in the middle part is captured by all tomograms, whereas significant differences exist for the high-velocity regions at 6-7 m depth.

Figure 3-6 shows the final ensemble of the 30 equivalent seismic S-wave velocity tomograms. The increased velocities at the bottom of the mode l reconstruction area, as well as the low velocities in the middle of the tomographic plane are captured by all tomograms.

3.7.2 Tomographic Uncertainty

The differences in the tomograms in Figures 3-4, 3-5, and 3-6 illustrate the ambiguity of the tomographic reconstruction problem due to limited number of observations and noisy data. All tomograms in Figures 3-4, 3-5, and 3-6 are considered as equivalently acceptable solutions of our inverse problems for radar-wave, P-wave and S-wave tomographic velocity reconstruction, respectively. For achieving a more realistic prediction model and to fit the data to an acceptable level but not beyond, the uncertainty from geophysical tomograms should be propagated in the prediction results.

The two locations for measured target parameters by direct push are at x=2.75 m (black lines in Figure 3-3), and x=8.0 m (gray lines in Figure 3-3). For incorporating model parameter uncertainty derived from equivalent geophysical tomograms in the ANN training, we take cells from these equivalent tomograms into account at the positions where direct push logs are coincident with tomogram grid cells. The measured direct push log at x= 2.75 m coincides with grid cells laterally centered at x= 2.5 m. Then we link the measured target parameter at x= 2.75 m to grid cells centered at x= 2.5 m. Also, The direct push log at x= 8.0 m is between cells laterally centered at x= 7.5 and 8.5 m. We link the measured target parameter to related cells at x= 7.5 and 8.5 m in the tomogram grid domain. Therefore, velocity values of grid cells located at x= 2.5, 7.5, and 8.5 m will participate in the training phase of the ANNs. We determine the uncertainty of the i-th radar-wave velocity tomogram (${}^{R}\Delta_{x,y,i}$) out of the set of k=30 equivalent radar-wave velocity tomograms for all grid cells C at x= 2.5, 7.5, and 8.5 m, and y=5 m, …, 15.5 m as

$$^{R}\Delta_{x,y,i} = \frac{(\max[Cx,y,k]_{k=1}^{30} - \min[Cx,y,k]_{k=1}^{30})}{\mathrm{C}x,y,i} \quad . \qquad (3\text{-}5)$$

In an analogue fashion, equation 3-5 is used to calculate the tomogram uncertainty of P-wave ($^{P}\Delta$), and S-wave ($^{S}\Delta$) velocity tomograms. Note, the range in the nominator of equation 5 could be replaced by other measures of variance, e.g., standard deviation, interquartile range, or others. Since the distribution of velocities for each grid cell does not match a normal distribution for each grid cell, we do not use the standard deviation here. However, choosing the range in equation 3-5 may result in over-accounting for the contribution of outliers, which could be seen of following a very careful and pessimistic weighting strategy striving to avoid over-fitting of observations in any case.

3.7.3 Logging Data Uncertainty

Assessing errors in direct push logging data is difficult, because no repeated measurements in undisturbed ground are possible. However, interpretation or measurement uncertainty of direct push logs should be taken into account when training the ANNs. All logging datasets (Figure 3-3) comprise different unknown errors, e.g., random or systematic. For assessing logging data uncertainty, we create a low pass zero-phase digital filter (Dutoit and Marques, 2010) with the window length adjusted to the estimated size of the sample volume. Here, we assume a vertical filter length of 50 cm and 25 cm for the CPT and dielectric permittivity logs, respectively. Fundamental assumption of this filtering is that anomalies with vertical extension less than the vertical extent of the sample volume cannot be imaged clearly. Hence, we assume that differences between the filtered log and the original log will largely be related to random measurements error or changing coupling conditions of the probe. The noise components of the logging data are computed by subtracting the filtered log from the original logging data. In Figure 3-7 we illustrate this step for tip resistance. In Figure 3-7a the black line

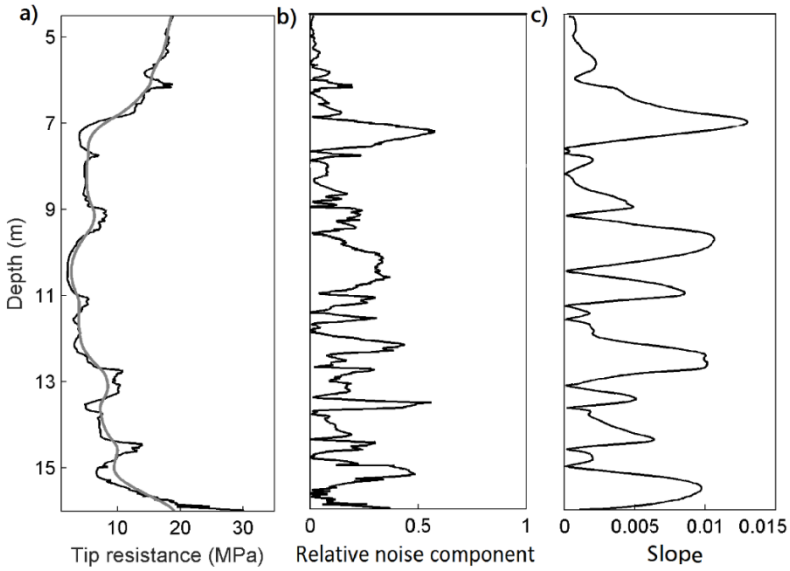

Figure 3-7: Tip resistance error calculation based on a low pass zero-phase digital filter. The black line in Figure 3-7a determines the measured tip resistance; the gray line shows the filtered log. In Figure 3-7b the relative difference between measured and filtered tip resistance are shown which is considered as data noise component. Figure 3-7c determines the slope based on the low pass zero-phase digital filter in Figure 3-7a.

depicts tip resistance logging data and the gray line shows the zero-phase filtered log. Then, we calculate the absolute difference between logging data and filtered log for assessing the relative noise component σ (Figure 3-7b).

Furthermore, the flat areas in the logging data or areas of low gradient are our desirable regions for learning the prediction model by ANN linking tomograms and logging data. Wherever the measured target parameters show high gradients this may indicate the presence of boundaries in the ground. Due to the finite sample volume of direct push probes their response integrates over a certain subsurface volume. From the log alone it is not possible to judge, whether subsurface boundaries are sharp discontinuities or gradually changing. For avoiding the effect of such uncertainty in each point P of measured target parameters we calculate the first absolute derivative between related neighboured readings in the log (i.e., ρ_i=(| P_i-P_{i-1}|+| P_i-P_{i+1}|)/(2*P_i)) (Figure 3-7c). The logging data error ϑ for reading i is then estimated as

$$\vartheta_i = \sigma_i + \rho_i \quad . \tag{3-6}$$

3.8 Setting up the ANNs

3.8.1 Combination of Tomograms for ANNs

We train the ANNs by providing {(input, target)} tuples as training dataset like

$$\{(\text{input, target})\} = \{(R^i_{x,y} P^j_{x,y} S^k_{x,y}, L_{x,y})\} \quad , \tag{3-7}$$

with x= 2.5, 7.5, and 8.5 m, y= 5,…, 15 m, and i,j,k= 1,…,30. For incorporating tomographic ambiguity we create tuples for all possible combinations of radar-wave velocity (R), P-wave velocity (P), and S-wave (S) velocity tomograms. There exist 27 000 (= 30^3) different combinations from {R^1, R^2… R^{30}}, {P^1, P^2… P^{30}}, with {S^1, S^2…S^{30}}. L determines the measured logging data which is either tip resistance, sleeve friction, or dielectric permittivity in our study. The measured logging data at x= 2.75 m are considered representative for grid cells centred at x= 2.5 m. Also the measured logging data at x= 8.0 m are considered representative for grid cells centred at x= 7.5 m and 8.5 m. Per one tomographic grid cell 50 observations of the target parameters are available.

3.8.2 Training Without Error Incorporation

ANN models which are trained with this strategy incorporate averaging of target parameters per grid cell. In this case L represents a vector of 50 observations of the target parameters per tomographic grid cell. Each combination of tomographic grid cells is 50 times repeated in the training dataset and linked to the 50 different observations of the target parameter per grid cell. Note, that the ANN learns a prediction model averaging over the 50 samples per grid cell. Therefore we repeatedly train 27 000 (= 30^3) different ANN prediction models. In doing so, we use 93 150 000 (= 30^3(= combination of tomograms)*3(= x positions of logs)*23(= number of grid cells per log)*50(= logging samples per grid cell)) tuples in total in the training datasets presented to the ANNs in the training phase. Per trained ANN a training dataset comprises 3 (= x positions of logs) * 23 (= number of grid cells per log) * 50 (= logging samples per grid cell) tuples. This training procedure results in 27 000 2D scenarios of the target parameters. In this strategy the performance parameter for ANNs in the training phase is MSE.

3.8.3 Training With Error Incorporation

In this strategy ANNs are trained based on the WMSE, and the related uncertainty of tomograms Δ and logging data measurements ϑ have been fed to the ANNs simultaneously. So the training dataset will be {(input, target, e)} and again 27 000 model combinations are used in this strategy. We estimate e in equation 3-4 for the i-th tuple like:

$$e_{x,y,i} = 1/({}^{R}\Delta_{x,y,i} + {}^{P}\Delta_{x,y,i} + {}^{S}\Delta_{x,y,i} + \vartheta_i) \quad . \qquad (3\text{-}8)$$

3.8.4 Training With Separate Logging Data

This strategy is based on the separation of individual logging datasets in the training phase. The training datasets are created either by logging data measured at x= 2.75 m, or at x= 8.0 m and independently linked to grid cells of the tomograms in the training tuples. Each training dataset in this strategy can carry information about only one position of logging data. This enables assessment of possible systematic shifts in the logging data when analyzing the prediction results. In this strategy the performance parameter in the training phase is WMSE for training 27 000 ANN models per logging data position. For x = 2.75 m a training dataset comprises 1 * 23 * 50 tuples, and for x = 8.0 m, a training dataset comprises 2 * 23 * 50 tuples.

3.8.5 Training Accounting for Resolution Difference

This strategy does not average the target parameters per grid cell, but considers the logs at x = 2.75 m and x = 8.0 m separately in the training phase. In the training tuples, now, we link each combination of tomographic grid cells to one observation of the target parameter. In this case L is a scalar and represents one observation of the target parameter per tomographic grid cell. To ensure equal contribution of every reading in the logs we first employ the uppermost log reading falling into a grid cell. Then we repeat the ANN training using a training dataset comprising the second-uppermost log reading in the grid cells in the tuples. We repeatedly train 1 350 000 (=30^3* 50) different ANN prediction models. Also in this training strategy ANNs are trained based on the WMSE as performance parameter. With this training strategy we not only consider the tomographic ambiguity but additionally the difference in the spatial resolution between tomograms (0.5 m vertical grid cell length) and logging data (1 cm vertical sample spacing) is transformed into prediction uncertainty. This training procedure results in 1 350 000 2D scenarios of the target parameters.

3.8.6 Selecting the Number of Neurons in the Hidden Layer

To ensure the utilization of sufficiently complex ANNs capable to propose a well-fitted prediction model we use the same strategy as Asadi et al., (2016) for selecting the number of neurons in the hidden layer. ANNs are repeatedly trained for a given training dataset employing different numbers of neurons in the hidden layer. We test 15, 25, 35, 50 and 100 neurons in the hidden layer for the same training dataset. The main criterion for selecting the number of neurons in the hidden layer is the prediction performance measured by the MSE. When increasing the number of neurons in the hidden layer the MSE will not be substantially lowered once the ANN offers sufficient complexity for linking the tomograms and the logging data. Figure 3-8 shows the MSE for all 27 000 tomograms combined with measured target parameters and different number of neurons trained with the averaging strategy. Increasing number of neurons allows generally for better ANN performance indicated by lowered MSE. The rather constant MSE for 50 and 100 neurons indicates that the ANN is complex enough for fitting a prediction model for any input combination. In the following we show the prediction results achieved using 50 neurons in the hidden layer.

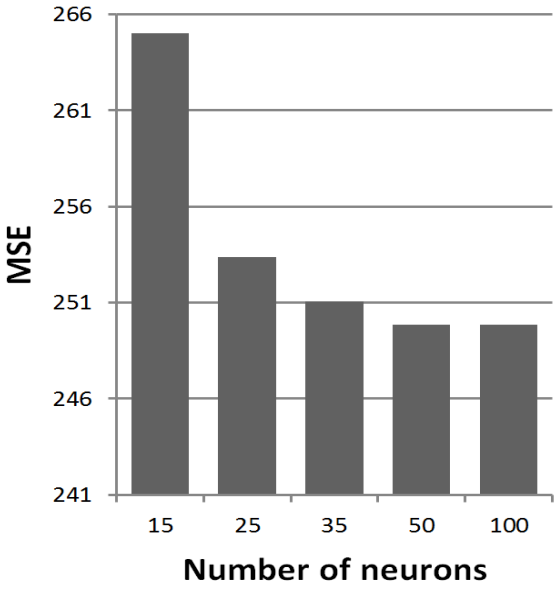

Figure 3-8: MSE from ANN training for all combinations of spatially continuous tomograms with 15, 25, 35, 50, and 100 neurons in the hidden layer. Based on this comparison 50 neurons have been selected as optimal number of neurons.

3.9 Results and Discussion

3.9.1 Prediction Results of ANNs

In each training strategy the achieved {(input, target)} tuples have been separated with 70%, 15%, 15%, for training, validation and test phase, respectively. We use random sampling for creating validation and test sets from our dataset, The evaluation procedure is done by the Neural Network Toolbox (The Mathworks Inc.) based on the MSE or WMSE. Mostly our trained ANN models show regression coefficients of approximately 0.95, 0.96, and 0.85 for tip resistance, sleeve friction, and dielectric permittivity, respectively, which indicate a high accuracy of the prediction models found by the ANNs. According to the regression coefficients our ANNs are able to achieve a satisfactory level of performance for prediction of the desired target parameters.

3.9.2 Training Without Error Incorporation

Figure 3-9 shows the results of the probabilistic spatially continuous prediction of tip resistance q_c sleeve friction f_s, and dielectric permittivity ε_r, respectively, trained with averaging of target parameter observations per grid cells and MSE as performance parameter without error incorporation. We illustrate the prediction uncertainty by showing relative frequency information drawn from the calculated 27 000 2D prediction scenarios. At the logging position (i.e., x= 2.5, 7.5, and 8.5 m in Figure 3-9) where the training has been performed the learned prediction models offer highly accurate prediction. In the rest of the 2D area the prediction ranges for the target parameters are significantly broader and largely unfocused. This is due to simple error propagation, i.e., tomographic and logging data uncertainty are ignored in the ANN training strategy. When applying the over-fitted prediction model to these regions this result in partly poor estimates of the target parameter, which can even take on values outside physically based limitations, e.g., tip resistance below zero MPa. Such values should be excluded from further processing or interpretation. Hence, it is important to account for tomographic and logging data uncertainties when learning the prediction models by ANN training.

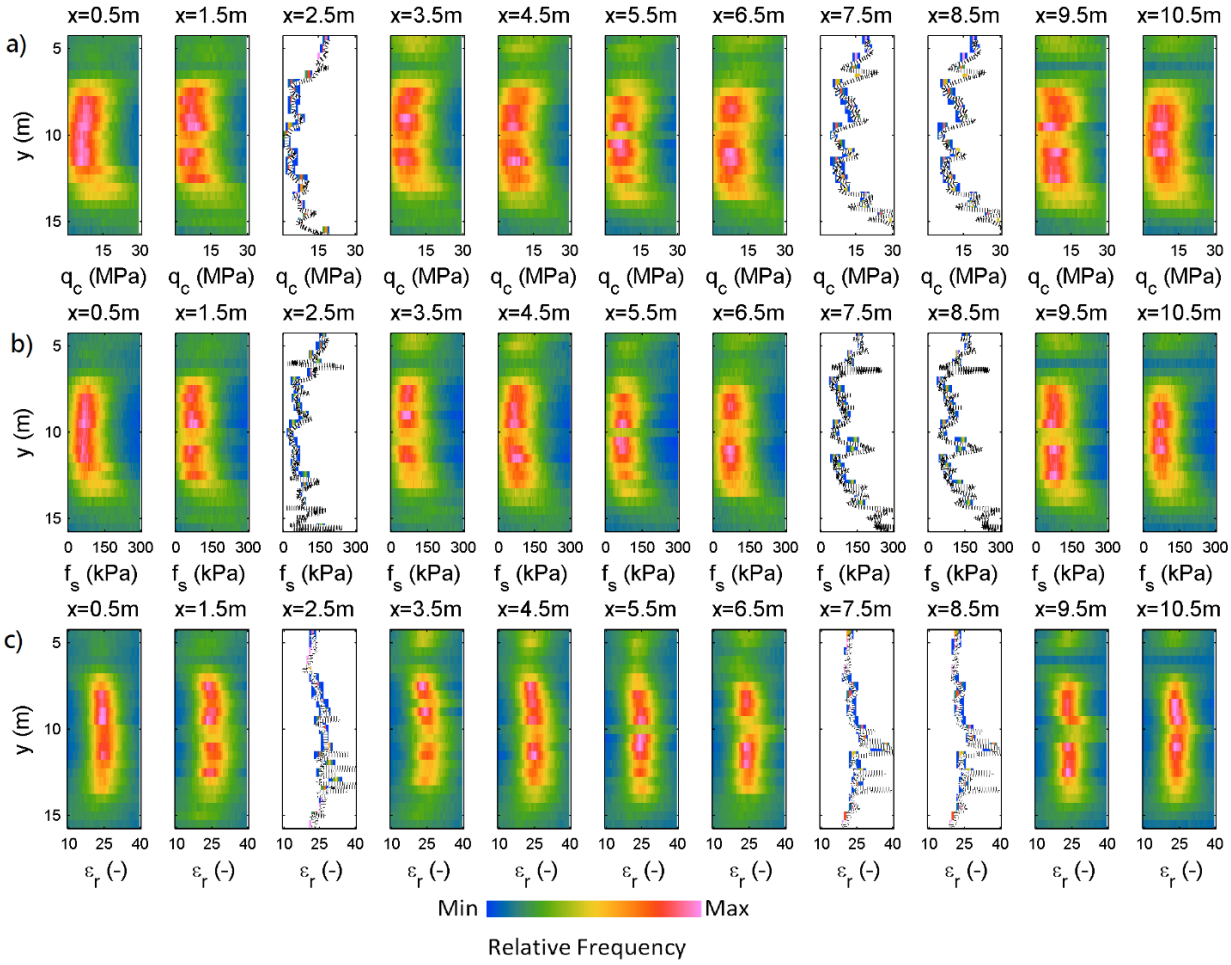

Figure 3-9: Prediction results of spatially continuous 2D target parameters based on the sparse logging data and 27,000 combinations of spatially continuous tomograms (Figure 3-4, 3-5 and 3-6) shown as histographic plot. 2D probabilistic prediction plots show (a) tip resistance, (b) sleeve friction, and (c) dielectric permittivity prediction. The dotted white lines show the measured logging data of target parameters (Figure 3) that are used for training the ANN. Red colors correspond to high relative frequencies. Blue colors correspond to low relative frequencies.

3.9.3 Training With Error Incorporation

To account for tomographic and logging data uncertainties we repeat our predictions based on the WMSE. Figure 3-10 shows the prediction results for training with error incorporation. They scatter now at the position of the logging data and the prediction range is broader at these positions than in Figure 3-9. But in the rest of the 2D area prediction models are able to offer sharper and more focused ranges for predictions of the target parameters. However, for tip resistance and sleeve friction, at the logging data

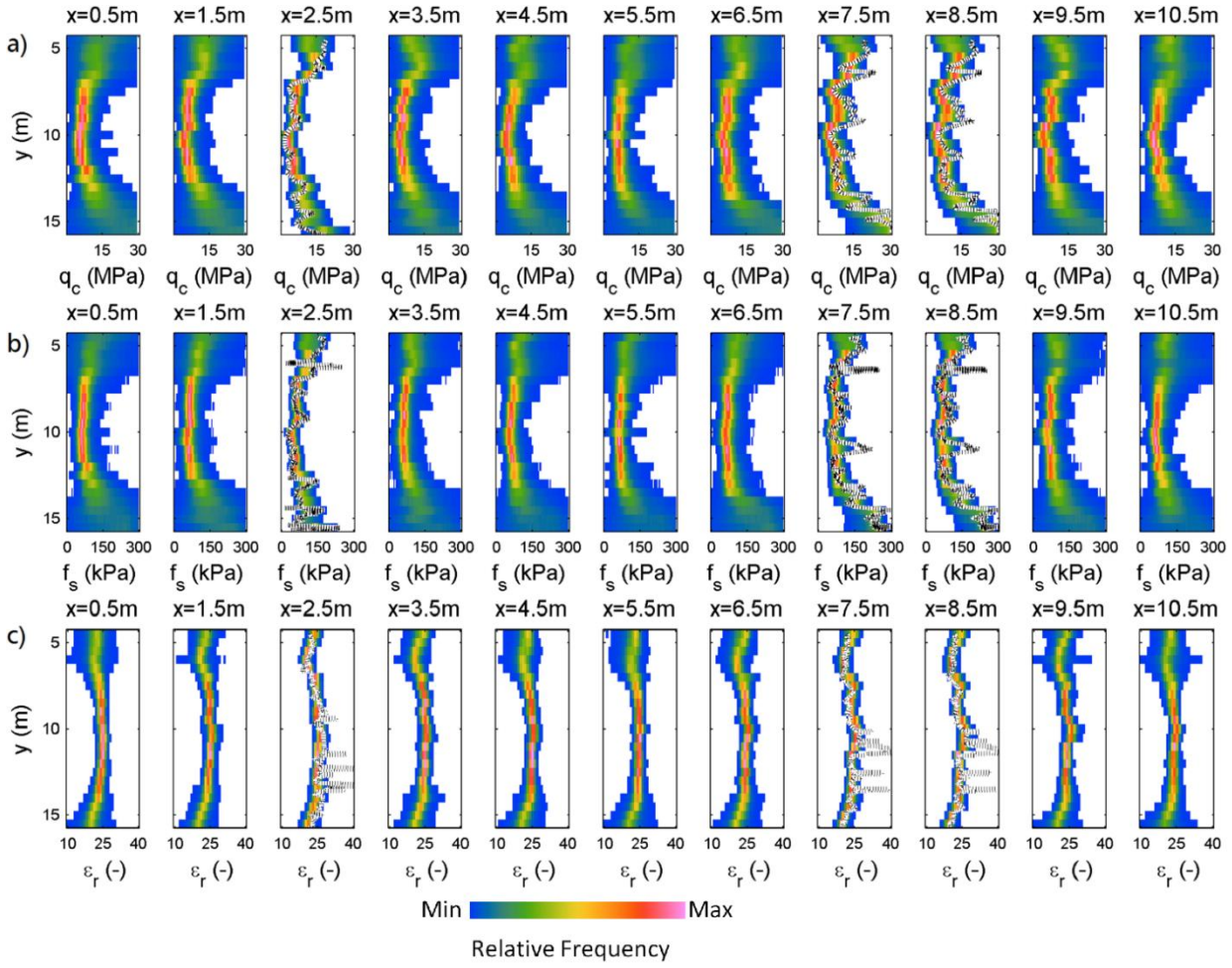

Figure 3-10: The same as in Figure 3-9, but trained considering tomographic and logging uncertainty when training the ANN.

positions the most likely prediction results indicated by highest relative frequencies are systematically lower or higher than the measured logging data. This indicates that the two logs are not fully fitting together when they are integrated in one training dataset for training the ANNs, i.e., both logs provide conflicting information in the training set when linked to the tomograms. In such case, ANNs take the average of both measured loggings data when they are in one training dataset. If the considered tomograms would exhibit systematic lateral velocity changes, the ANNs could have been able to fit both logs. It is difficult to decide, whether the logs acquired at x= 2.75 m and x= 8.0 m carry a systematic shift, the tomographic imaging missed lateral velocity variations, or petrophysical relations between tomograms and logs are spatially highly variable. When comparing the tip

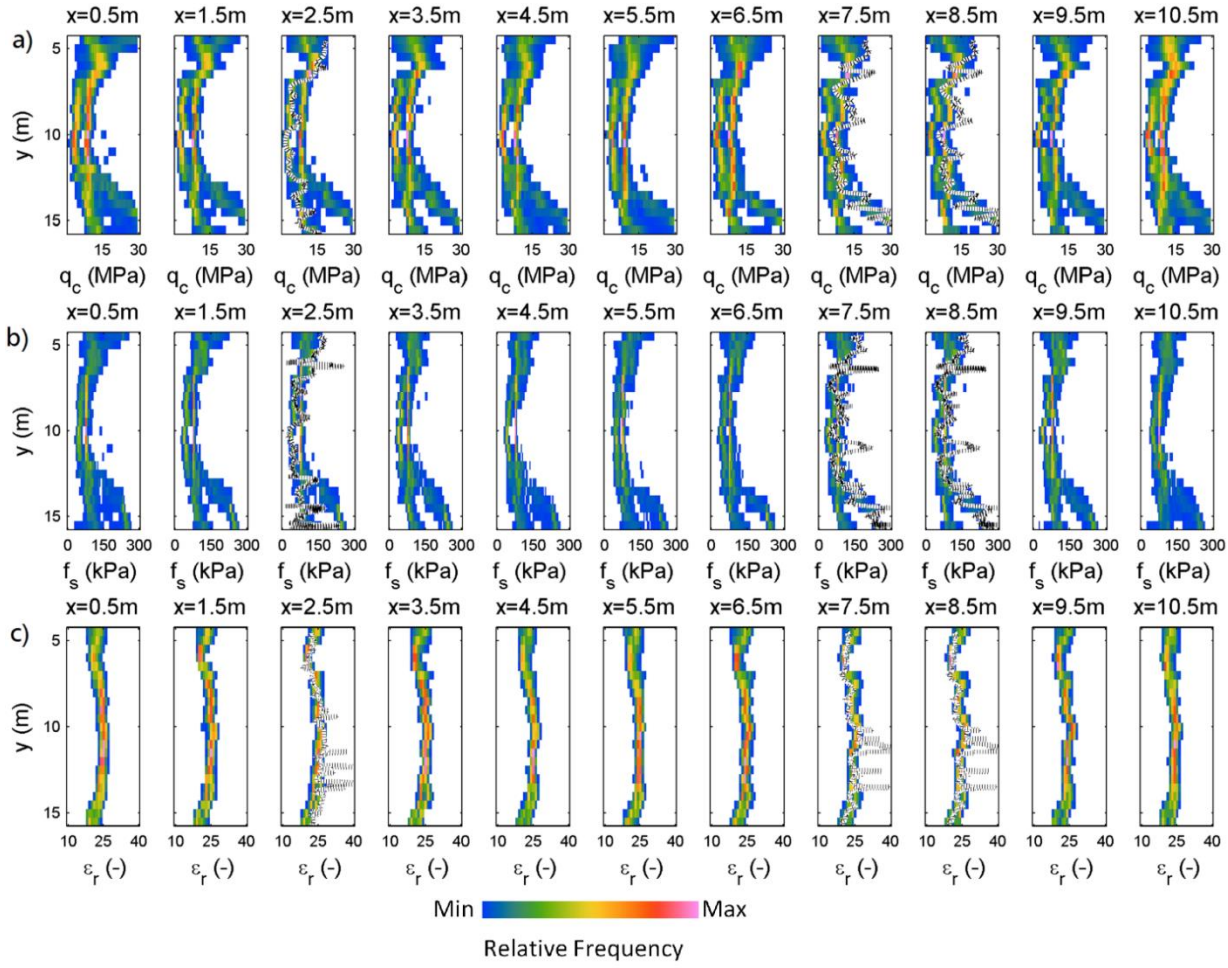

Figure 3-11: The same as in Figure 3-10, but now training was individually performed for logs at x =2.75 m and x = 8.0 m.

resistance and sleeve friction logs acquired at x= 2.75 m and x= 8.0 m (Figure 3-3), it appears that both logs systematically differ at depths below 7 m. Tip resistance and sleeve friction are recorded during the same cone penetration test. The dielectric logs are acquired using a different probe and here the high relative frequencies coincide better with the major trends of the logs (Figure 3-10c). Hence we repeat our prediction training the ANNs individually for the logging data acquired at different positions.

3.9.4 Training With Separate Logging Data

Figure 3-11 shows the prediction results when training individually for the logs from two locations. In this case the computational time will be increased. The result differs from those previously achieved (c.f. Figure 3-9 and 3-10). For tip resistance and sleeve friction

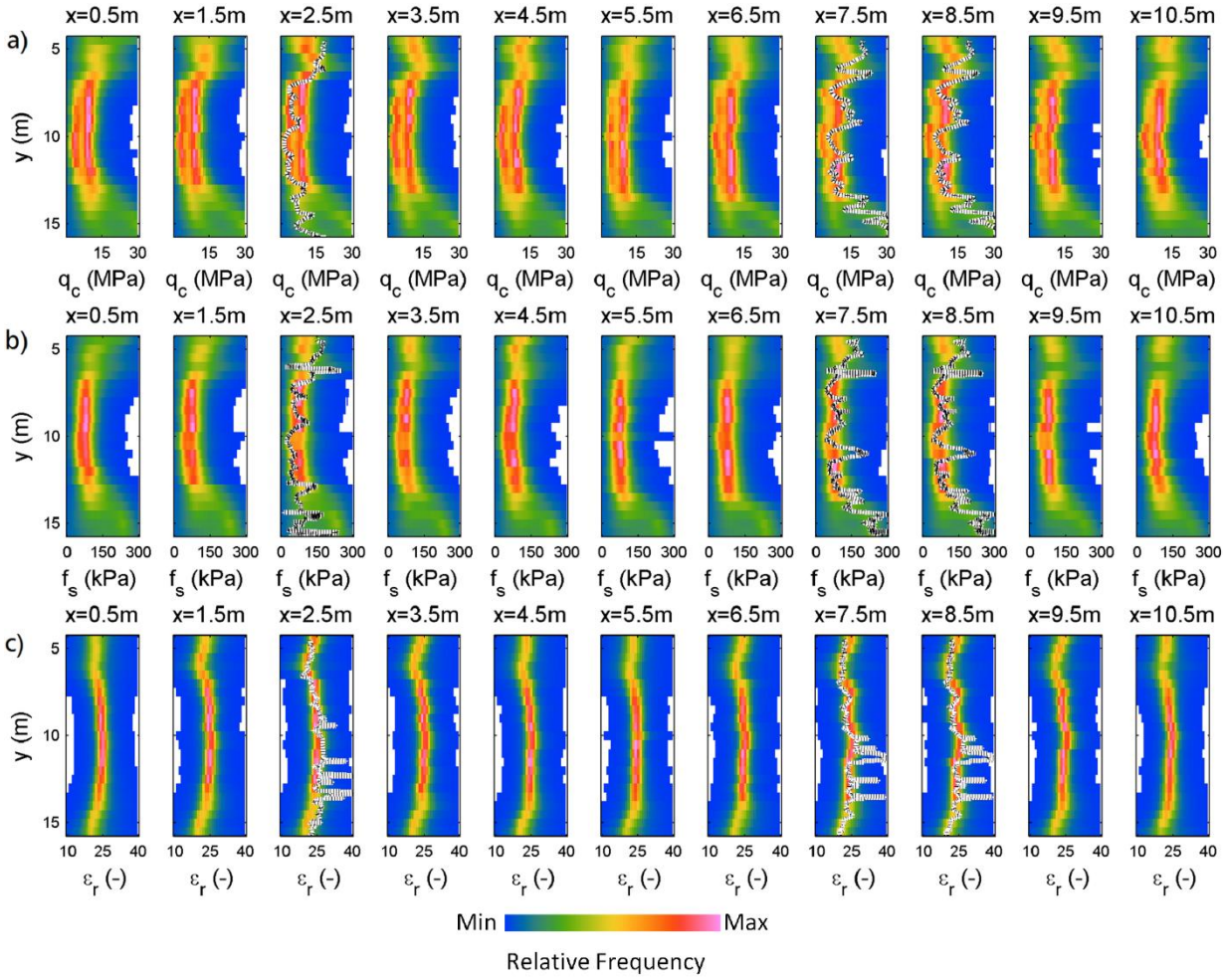

Figure 3-12: The same as in Figure 3- 11 but, now repeatedly considering individual samples from the logging datasets per grid cell when training the ANN, rather than averaging over all samples corresponding to a grid cell**.**

the prediction results show clearly a bi-modal distribution indicating that the logs contain a systematic difference, with regard to the velocity variations in the tomographic data. The prediction ranges are now narrow and highly focused. At each position the logging data can be linked very well to the velocity variability in the tomograms. However, a number of small-scale anomalies present in the logging data exceeds the prediction range.

3.9.5 Taking Resolution Differences into Account

For the prediction results in Figures 3-9, 3-10, and 3-11 the employed training schemes are not able to consider measured small scale variability of the target parameters in their prediction results. In the training phase ANNs try to average the target

parameter observations per gird cell in a least square sense. We repeat our prediction using the training strategy without averaging. The results of this training strategy are shown in Figure 3-12. Prediction ranges increase now significantly, but are large enough to incorporate all logging data. Regions of high relative frequency are slightly more smeared and particularly for sleeve friction prediction the bi-modal character is smeared out.

3.9.6 Comparative Discussion

In Figure 3-13 we compare tip resistance predictions at x = 6.5 m drawn from the Figures 9-12. The panels in Figure 3-13 illustrate results achieved with the four different strategies for training the ANNs (Figure 3-13a without error incorporation, Figure 3-13b with error incorporation, Figure 3-13c error incorporation and separate logging datasets for training, Figure 3-13d additionally accounting for resolution/sampling difference between tomograms and logs). Summarizing, accounting for logging data and tomographic uncertainty reduced the prediction ranges and sharpened the region of high relative frequencies depicting most likely tip resistance values. If the measured logs cannot be brought into full coincidence with the available tomograms by the ANN, either due to lateral differences not captured by the available tomographic datasets, laterally changing petrophysical links between imaged physical parameters and the quantities measured in the logs, or due to systematic shifts when acquiring the logging data we recommend that each training dataset should be created with only one log considered (Figure 3-13c). This will result in bi- or multi-modal distributions if two or more logs are available, but it is regarded as the most honest and conservative approach. Logs should only be shifted, e.g., mean-value calibrated, if there is evidence that the bi-modal prediction results are caused by systematic acquisition errors, e.g., calibration errors. However, in most situations, it will be impossible to decide about this issue with high certainty. Particularly, when striving to achieve good most-likely estimates of the target parameter, the results shown in Figures 13b and c appear optimal. However, when striving to predict plausible ranges of highest or lowest possible target parameter values, it is essential to even incorporate differences in resolution/sampling between tomographic grid cell sizes and logging data sample interval. Hence, when the prediction objective is

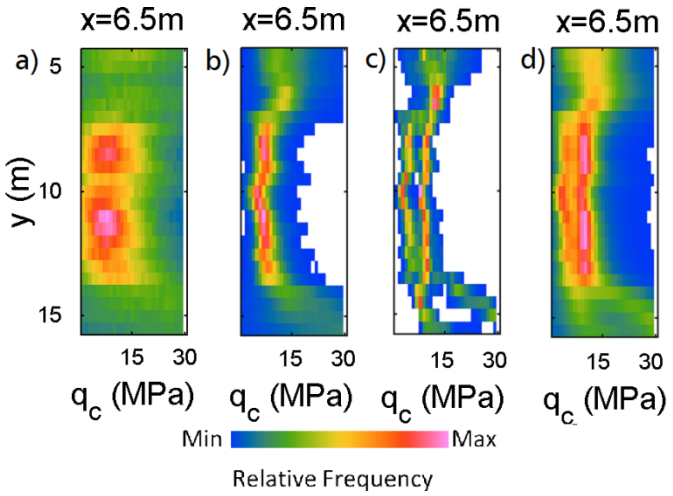

Figure 3-13: Example of comparison of tip resistance predictions at x= 6.5m drawn from Figures 3-9, 3-10, 3-11, and 3-12. Trained (a) without error incorporation, (b) with error incorporation, (c) separate logging data, and (d) accounting for resolution difference between logs and tomograms.

rather directed towards a conservative assessment of realistic and data-driven prediction ranges, training of the ANNs should be done with a strategy avoiding the averaging of target parameters per grid cell (Figure 3-13d).

3.9.7 Transferability and Outreach of Results

When employing ANNs for the probabilistic prediction of 2D or even 3D target parameter distributions no limitations are present with regard to (i) the number of considered tomographic datasets, (ii) the size of the ensembles of equivalent tomograms, and (iii) the nature of the target parameter measured in boreholes. Our analyses in this work are limited to crosshole tomographic acquisition schemes, but since tomographic ambiguity is considered, it will also be transferable to other tomographic setups. However, when considering tomographic datasets acquired solely from the Earth surface, ambiguity is usually expected to increase systematically with depth. In such cases, the prediction model learned by the ANN may be dominated by near-surface information. Speculating, modifications of the presented approach, e.g., by additionally incorporating depth as information layer in the ANN training, may help learning prediction models that evenly apply to strongly depth-dependent parameter interrelations.

The probabilistic inference may be of particular interest in exploration or interpretation projects, where decisions on future actions require a critical and quantitative risk assessment. Here, it can substantially help to use a data-driven inference technique translating tomographic ambiguity, scale or resolution differences, and observational

uncertainties into prediction uncertainty suitable to define realistic upper and lower bounds for the expected underground states.

In near-surface applications, mobile crosshole tomography relying on temporary installations realized by direct-push or sonic-drill technology (Paasche et al., 2013) can replace the classical cross-borehole tomography making the permanent installation of boreholes obsolete. Such highly mobile crosshole tomography allows for on-site adaptation of the tomographic acquisition setup. The resultant tomograms can be linked to exploration target parameters, e.g., acquired by CPT or other 1D exploration techniques as discussed here, for realizing a highly flexible, site-adapted, probabilistic exploration for hydrological or engineering tasks.

3.10 Conclusions

We have employed static two-layer feed forward ANNs to establish a link between geophysical tomographic images and target parameters solely measured in the boreholes or by direct push technology. Based on this relation we are able to calculate 2D probabilistic predictions of the target parameters which are of higher relevance to hydrologic, petroleum, and engineering exploration tasks than the tomographically imaged parameters. The used geophysical tomograms resulted from fully non-linear inversion generating ensembles of geophysical tomograms that fit the underlying datasets equally well. We have tested different learning strategies when training the ANNs. It is important to incorporate uncertainties from tomographic imaging and the target logging data in the ANN training to avoid overfitting the training data offered to the ANN. Depending on the chosen training strategy, our method results in focused probabilistic predictions with small ranges suitable to assess the most likely values of the target parameters in the 2D tomographic plane. Alternatively, the ANNs can be trained such that even small-scale anomalies beyond the spatial resolution of the tomograms are considered, which results in broad and rather conservative prediction ranges. This methodology can be applied to any combination of geophysical tomograms and geotechnical or hydrological logging data. Tomographic ambiguity, logging data

uncertainty, and difference in spatial resolution between tomograms and logging data can be transduced into the probabilistic prediction of the target parameters.

Prediction performance was found to be excellent, and can be applied to any combination of geophysical tomograms and target parameters since at no point critical assumptions about the involved parameters or the expected relations between the considered datasets and parameters are made. When combining this approach with fully non-linear geophysical tomographic imaging, this combination can deliver objective and purely data-driven probabilistic predictions of the target parameter distributions, which are essentially required when striving to assess, quantify and minimize risks in resource exploration and utilization. We believe taking the uncertainty of tomograms and logging data ambiguity into account for probabilistic prediction of target parameters can help in a variety of geophysical applications to analyze and identify complex parameter relation which cannot be described by more conventional and often linear models.

Chapter 4

Conceptual Developments for Clustering Mapped Data Emanating From Technical Sensors and Subjective Insights of Human Experts

4.1 Abstract

When exploring the ground, earth scientists frequently map the available observations in individual but collocated thematic images, e.g. geological, hydrological, magnetic, electrical conductivity or radiometric maps. Scientists in these fields try to analyse these datasets (i.e., map or image information) according to their subjective beliefs or experience. In recent years, attempts have been made in various fields of earth sciences towards rapid, automated, and objective information extraction from spatially continuous disparate datasets based on powerful statistical analyses, machine learning, or data mining techniques (i.e., clustering, classification, regression, etc.). One of the important tasks in earth observation data analysis is the integration and segmentation of multiple thematic images, e.g., by cluster analysis, such as fuzzy c-means or k-means. Traditional workflows for map integration and segmentation are able to offer rapid and automated clustering results of the multi-parameter geophysical datasets only considering the technical or objective data which are measured by sensors or some technical devices. However, currently there is no way for an intelligent combination and utilization of (partly) subjective and technical information in such cluster analyses, e.g., by considering the information provided by pre-classified geological or soil maps. We are going to discuss conceptual ideas inspired by data mining for integrating and clustering multi-parameter

geospatial datasets while paying attention to the subjective and technical acquisition procedure. The explained approach in this chapter may potentially allow for multi-parameter integration and cluster analysis according to the technical or objective information and additional consideration of knowledge provided by human experts. We belief that such integration can strongly help to solve the problem of handling noisy data and unusual structures in geospatial data analysis. We illustrate critical aspects of our conceptual ideas using small synthetic datasets illustrating problems and potential when clustering data emanating from technical sensors and subjective insights or expectations of a human expert. Furthermore, we apply our idea to a real world datasets containing subjective and technical data of the area to offer segmentation or clustering of the considered domain.

4.2 Introduction

Technological developments in ground-borne and air-borne passive and active sensing technologies provide new opportunities for getting information about the ground. These advances result in the acquisition of big databases comprising complementary datasets imaging ground properties and conditions. Often such datasets are presented as highly sophisticated and eye-catching images. The analysis, integration and interpretation of such multi-parameter databases is still a time-consuming challenge and extraction of patterns and assignment of a meaning to them is often a non-trivial task. The success of processing and interpretation is usually related to the experience of the interpreter or scientist and bears at least a partly subjective component. In the last decades, attempts have been made in various fields of earth sciences towards rapid, automated and objective information extraction from spatially continuous disparate datasets. Statistical pattern recognition and data mining tools (i.e., image processing, clustering, or classification) have proven valuable for largely automated and rapid information extraction from multi-parameter spatial datasets (e.g., Leung, 2010; Fischer and Getis, 2013).

One of the vital tasks in dealing with multi-parameter air-borne and space-borne spatial datasets is clustering or segmenting these datasets to achieve maps or images outlining dominant units of similar ground material composition. Clustering is a generic term for a wide variety of powerful data mining algorithms grouping large sets of multi-parameter data into several segments or clusters based on some similarity or distance measures calculated between the data points. Such clustering can be done by crisp or fuzzy cluster analysis. Crisp cluster algorithms provide information about the cluster a data point is most reliably assigned to, whereas fuzzy cluster algorithms follow the concept of partial cluster membership and quantify the assignment of a data point to each cluster by a cluster membership measure and provide information about the statistical significance of the assignment of a data point to a certain cluster (e.g., Höppner et al., 1999; Kaufman and Rousseeuw, 2009). Multiple families of cluster analyses algorithms exist, e.g., hierarchical, partitioning, density-based, model-based and spectral clustering. Out of these, partitioning cluster algorithms have been widely used for analysis of earth

observation databases. For example, crisp cluster analysis has been used to recognize statistically significant structures or clusters in air-borne geophysical datasets (e.g., Peschel, 1973; Lanne, 1986; Pires and Harthill, 1989; and Martelet et al., 2006), and geological and soil degradation mapping (e.g., Eberle, 1993; and Anderson-Mayes, 2002). Furthermore, fuzzy cluster analysis has been employed routinely for segmentation of remotely sensed data (e.g., Du and Lee, 1996; Ahn et al., 1999; Shi et al., 2003), analysis of rock magnetic and geochemical parameters of different materials (e.g., Kruiver et al., 1999; Urbat et al., 2000; Knab et al., 2001), for data-driven soil clustering (e.g., de Bruin and Stein, 1998; Bragato, 2004, Schröter et al., 2015), geophysical and geological mapping (e.g., Paasche et al., 2006; Paasche et al., 2007; Paasche and Eberle, 2009), and soil moisture mapping (e.g., Schröter et al., 2015).

Despite the emergence of new spatial data measurement technologies noise, or outliers, are ubiquitous in technical or objective datasets (Taylor, 1982; Zhang et al., 2003; Han et al., 2011). In addition to random noise, systematic anthropogenic or environmental effects may contaminate measured data, e.g., farmers driving lanes on arable land or increasing cloudiness or dustiness, respectively. As the volume of noise increases in the databases, the cost of mining and evaluating will also increase. Most clustering algorithms are vulnerable to noise and outliers and they may offer unusual segments or clustering results when there is noise in the datasets. Penalizing a clustering algorithm using a spatial domain may increase clustering robustness regarding random noise (e.g., Pham, 2001). However, all new developments in the earth map integration based on clustering (crisp or fuzzy clustering) usually only consider technical, i.e., measured by technical system, or objective spatial datasets which inherently carry noise. Therefore auxiliary information about the considered domain and careful choice of clustering methodology and preparatory processing must be made if the data in the application contain a large amount of noise (Han et al., 2001).

For proving clustering results of spatial geoscientific databases the integrated segmented maps can be compared to subjective insight of geoscientific experts as an auxiliary information about the considered domain. Alternatively, additional ground truthing measurements may be carried out striving to characterize the found segments.

However, sparse ground truthing and human experts may see or overlook some structures, properties or situations of the domain which the technical sensors are not able to catch or have caught, respectively. In such comparisons (partly) subjective belief (i.e., in the form of a soil map or a geological map) about the domain will be shown as an additional map or dataset which highlights the structures which are recognizable by expert scientists. In such map, the only information which is important from a pattern analysis point of view is the boundary between segments which expert scientist draw for separating the structures of the domain.

To our knowledge there is no sophisticated method available allowing the integrated clustering of subjective and objective spatial data (i.e., emanating from technical sensing, such as satellite imagery) in an intelligent and logical way during the data processing stage. In this chapter, we are going to test and discuss conceptual ideas inspired by data mining and machine learning techniques such as boundary detection, graph theory, sampling, and clustering for integrating and clustering spatial datasets, while paying attention to their subjective (humanly beliefs about the domain) and technical acquisition origin. We employ boundary detection techniques and graph theory to weight subjective and objective datasets based on their importance or accuracy before calculating similarity measures between the data points striving towards data clustering. We begin by explaining our processing flow. This is followed by applying it to a synthetic and a field database. Finally, we explain and discuss the results achieved for both databases.

4.3 Methodology

4.3.1 The K-means Clustering

The k-means clustering algorithm (MacQueen, 1976; Kaufman and Rousseeuw, 2009) has become a popular tool for partitioning of multi-variate databases and is independently applied in different scientific fields (i.e., computer sciences, image processing, environmental earth sciences, etc.). This algorithm is one of the most widely used unsupervised clustering algorithms that generates a specific and disjoint number of clusters from data points stored in the datasets. The main reasons for the popularity of the k-means cluster algorithm are ease of implementation, simplicity, efficiency,

robustness, and empirical success. Let P={$\mathbf{p}_1$, …,$\mathbf{p}_n$} be a set of position vectors specifying *n* points in a *d*-dimensional space. This should be partitioned into sets S_j with *j* = 1, …, *k*. *k* specifies the number of clusters. Each set defines a cluster described by its cluster center position $\mathbf{c}_j$ in the *d*-dimensional space. The k-means algorithm finds clusters such that the sum of squared errors *SSE* between the cluster centers, described by the mean value of all data points belonging to a cluster, and the data points belonging to this cluster is minimized by

$$SSE=\arg\min_{S}\sum_{j=1}^{k}\sum_{\mathbf{p}_i\epsilon S_j}\left\|\mathbf{p}_i-\mathbf{c}_j\right\|^2 \quad . \qquad (4\text{-}1)$$

$\left\|\mathbf{p}_i-\mathbf{c}_j\right\|^2$ is the distance between data point $\mathbf{p}_i$ and the centroid $\mathbf{c}_j$, which is known as Euclidean distance (Danielsson,1980). Euclidean distance computations are not invariant to linear transformations, which makes the clustering results dependent on most standard data scaling or normalization techniques. It analyses the entire database globally and does not consider spatial information about data point arrangement in the map domain, which makes it sensitive to measurement errors.

k-means cluster analysis could be directly applied to a multivariate database, e.g., when scaling the different physical quantities mapped by each dataset along orthogonal axes (e.g., Paasche and Eberle, 2009). Clustering would then take place in the spanned physical parameter space, but only the physically measured values at every location in the mapped database would form the information base considered for clustering. Further information present in the mapped images, such as texture or boundary information, cannot be considered in a simple k-means cluster analysis application, if applying the algorithm directly to the database. Nevertheless, it would often be desirable to do the cluster analysis on different image properties, particularly, since human contemplators consider image texture, boundary information and measured values when analysing a mapped database. Preparatory processing is required to extract information about boundary, measured value, and/or texture information, described by different features or attributes. It must be integrated paying attention to the nature and characteristics of the individual considered datasets when integrating the extracted information prior to clustering.

4.3.2 Graph Theory and Shortest Path

Graphs are mathematical structures to model pairwise relations between data points. A graph is made of connected nodes (or vertexes). Nodes represent data points and the connections between them are called edges. Edge weights illustrate distance or similarity between connected nodes. Here, we only consider undirected graphs, i.e., connectivity and connection weights are independent from direction. Each graph G= (V, E) comprises a set of data points V and edges E. One of the important tasks with relation to graph theory is the shortest path problem (Dijkstra, 1959), i.e., finding a path between two nodes (or data points) in a graph such that the sum of the edge weights of its constituent edges along the path is minimized. When two nodes have a similarity (or dissimilarity) with each other then they are adjacent and both are incident to a common edge. A single shortest path is a sequence of nodes Q= ($v_1, v_2, \ldots, v_n$) ϵ V. v_i is adjacent to v_{i+1} and $1 \leq i \leq n$. Q is a path from node v_1 to v_n with length n-1. When e_i be the weight of the edge between nodes v_i and v_{i+1},than the shortest path from node v_1 to v_n is the minimum of $\sum_{i=1}^{n-1} e_i$ (Dijkstra, 1959).

In graph theory, a shortest path from a source node to all other nodes in a graph is called shortest path tree. The all pairs shortest path problem determines the task when we have to find the shortest paths between every pair of all nodes in a graph. There are different algorithms to find single shortest paths, shortest path trees, and all shortest paths in the graph (i.e., Dijkstra's algorithm (Dijkstra, 1959), A* search algorithm (Goldberg and Harrelson, 2005), Floyd–Warshall algorithm (Burfield, 2013), and etc.). Dijkstra's algorithm (Dijkstra, 1959) is one of the oldest and the most popular algorithm to find different types of paths.

4.3.3 Boundary Detection

Numerous image processing techniques identify features in images or maps that are relevant for estimation and extraction of objects or structures in images. Boundaries are one of the important properties which go along with significant local changes in the intensity of the image (Davis, 1975). Figure 4-1 shows different types of boundaries in images or maps. A boundary can be either a step function, where the intensity of an image

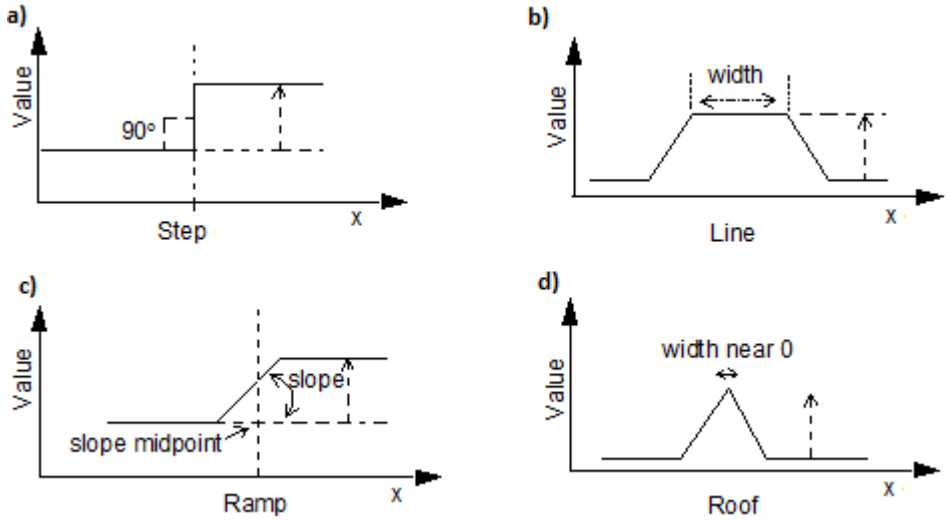

Figure 4-1: Different types of boundaries in an image (a) step (b) line (c) ramp (d) roof.

abruptly changes from one value on one side to a different value on the other side, a line, where the intensity of an image abruptly changes but then returns quickly to the starting value. When intensity is not instantaneous and occurs over a finite distance the step boundaries will be converted to ramp boundaries and line boundaries will be converted to roof boundaries.

Different boundary detection algorithms have been introduced to produce a set of boundaries from an image (Senthilkumaran and Rajesh, 2009). A boundary detection solution should address three criteria. First of all, the boundary detection technique should accurately catch all possible and different types of boundaries. Second, the boundary point should be localized on the center of the boundaries, and third, the image noise possibly should not create false boundaries (Davis, 1975, Senthilkumaran and Rajesh, 2009). One of the useful boundary detection techniques, which consider these three criteria, is the Canny boundary detection algorithm developed by Canny (1986). For extracting the boundaries the Canny boundary detection algorithm follows five steps:

1. Apply Gaussian filter to remove the noise boundary and smooth the image
2. Calculate the intensity gradient of the image
3. Apply non-maximum supervision to extract boundaries on the center and get rid of spurious response
4. Use a threshold to determine all potential boundaries

5. Finalize the detected boundaries by removing all other boundaries that are weak or not connected to strong boundary using information of eight connected neighbour pixels or mapped data points

When applying a Gaussian moving window filter the noise in the image will be smoothed out. It further smooth the boundaries as high-frequency feature in a way that increases the possibility of missing weak boundaries and emphasizes the strong boundary points in the results.

Another type of boundary detection which can be applied to a color image or color-coded map is introduced by Martin et al., (2004), who determine a function $Pb(x, y, \theta)$ that predicts the posterior probability of a boundary at each image pixel (x, y) with orientation θ by measuring the difference in color, local image brightness, and texture channels (Arbelaez et al., 2011). The basic idea of the *Pb* boundary detector is the computation of an oriented gradient signal $G(x, y, \theta)$ from an intensity map m. This procedure proceeds by placing a circular disc at location or data point (x, y) and split the area around this position into two independent half-discs (i.e., h_1 and h_2) by a diameter at angle θ. For each half-disc, the histogram of the intensity values or color of the pixels will be calculated. At location (x, y), the gradient magnitude G is defined by the χ^2 distance between the two half-disc histograms h_1 and h_2:

$$\chi^2(h_1,h_2)=\frac{1}{2}\sum_i \frac{(h1(i)-h2(i))^2}{h1(i)+h2(i)} \quad . \quad (4\text{-}2)$$

The χ^2 distance for the pixels show the gradient magnitude of the 2D image or map for each pixel and highlight the boundary of the images (Martin et al., 2004; Arbelaez et al., 2011).

4.3.4 Sampling Strategy

In statistics and data analysis, sampling is the technique of selecting a subset of individual points from a statistical data point for the estimation of the characteristics or properties of the whole population (Israel, 1992; Marshall, 1996). Two important advantages of applying sampling to a population for data analysis are that the cost and

running time of the algorithms is lower than measuring the whole population. The sampling procedure has several important steps (Israel, 1992):

- determining the population of interest
- defining a sampling rule by a set of items or events which are possible to measure
- selecting a sampling method for selecting a subset of items or events from the population
- selecting the sample size
- sampling or collecting data from the population based on the desired sampling method

In the last decades different sampling strategies have been introduced with regard to the nature, quality, accuracy, or cost of the chosen sampling strategy (Israel, 1992; Esfahani and Dougherty, 2013). A popular sampling technique is the systematic sampling method (Esfahani and Dougherty, 2013), which relies on arranging and ordering the population according to some rules and conditions. Then, the subsets or elements will be selected at regular intervals from the ordered list. Systematic sampling involves a random start point and then proceeds with the selection of every k^{th} (= population size/sample size) element from then forward.

4.3.5 Processing Flow

Figure 4-2 outlines the processing flow when working towards integration and clustering of multi-parameter spatial databases. Two types of datasets that image the reality into sets of observations are available. The first type is named as objective datasets (i.e., hyper spectral, EMI, or radar data), which are measured by some technical sensors or devices. Such datasets inherently carry noise and different types of boundaries (i.e., step, line, ramp, or roof). The second type is subjective (i.e., soil maps, or geological maps), at least partly, created by humans or expert scientists that can be either correct

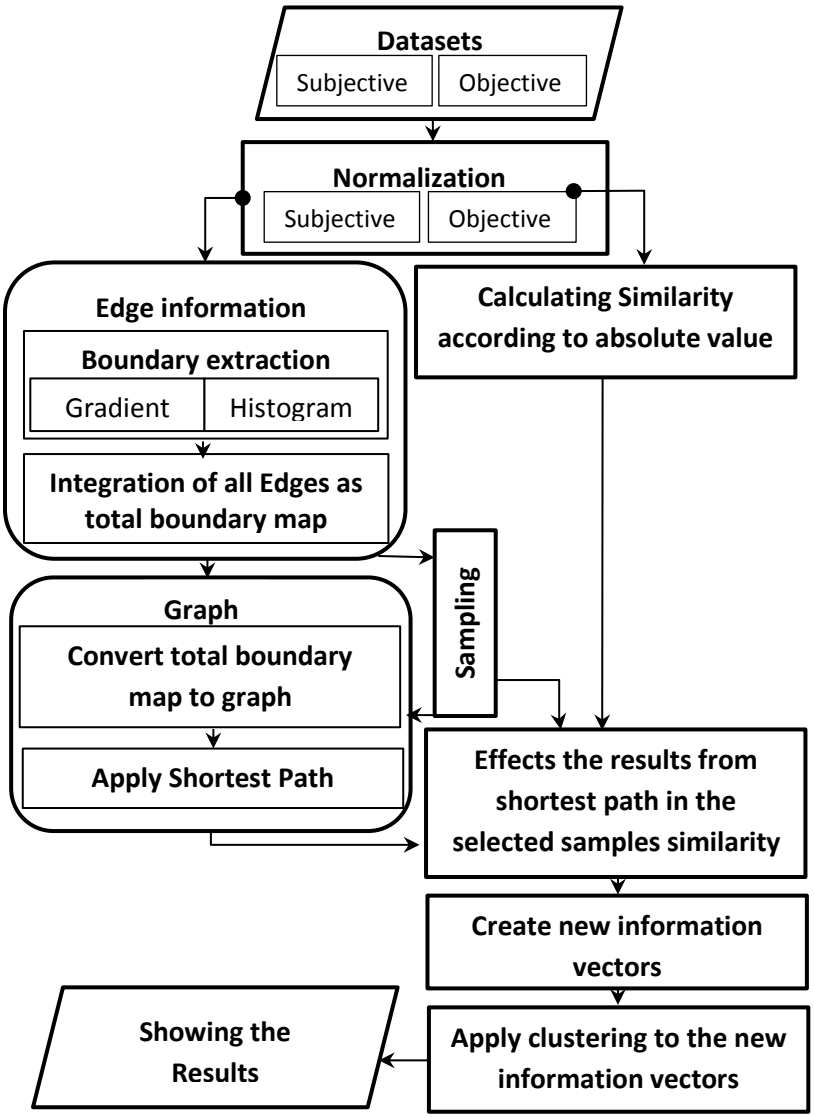

Figure 4-2: Processing workflow to integration and segmentation of subjective and objective datasets. After normalizing the datasets two processing branches will be followed in the workflow. In the right part the similarity only based on the objective data will be calculated. Simultaneously in the left part the boundary of the subjective and objective maps will be extracted. Then, the new information vector based on results of these two branches will be calculated. This new information vector will be considered as new dataset and will be the input for the clustering algorithm.

or incorrect based on the experience of the scientist. Such datasets carry only step boundaries or show sharply contoured structures.

In the next step normalization will be applied to each dataset to have parameters varying in the same scale range. One normalization method which is usually well known for rescaling data is scaling all numeric variables in the range [0, 1] by

$$p_{new} = \frac{p_d - min_d}{max_d - min_d} \quad , \quad (4\text{-}3)$$

where P_d is a value of data point p in a d-dimensions dataset. After normalization, two processing branches will be applied to the datasets (see Figure 4-2) focusing on either boundary information or the measured physical values. Here, a third branch could potentially be added considering image texture information. In the branch for calculating the similarity between all pixels of a datasets or image, we use only the objective datasets by applying Gaussian similarity for a data point p with d dimention formulated as:

$$Sim(p_i,p_j) = exp(-||p_i - p_j||^2 / 2\Omega^2) \quad . \quad (4\text{-}4)$$

In the branch for extracting the boundary information all subjective and objective datasets will be used. The spatial subjective and technical datasets can be seen as an image of the map domain and each data point in a dataset defines a pixel in the image. Therefore, a boundary point is a data point in a dataset or a pixel in an image with coordinates [x,y] where a significant local intensity change occurs. Different types of boundary detection can be applied to the datasets. In the subjective map because only the boundary points are important a method which be able to extract sharp boundary is necessary (i.e., Canny boundary detection). But in the technical map the gradients over the whole area should be extracted to cover all structure's changes in the map. Here, for subjective datasets we apply a gradient based method like Canny and for objective datasets we use the χ^2 distance, which are both able to extract boundaries from a different type of map or image. In the next step, all extracted boundaries from subjective and objective datasets will be integrated into one boundary map referred to as total boundary map (TBM). When there are m independent boundary subjective maps (BS) and n independent boundary objective maps (BO), then the value of pixel i,j in TBM is equal to:

$$TBM(i,j) = \sum_m (\alpha_m BS(i,j) * \sum_n \sigma_n BO(i,j)) + \sum_n \sigma_n BO(i,j)) \quad (4\text{-}5)$$

The first term of this equation,$\sum_m(\alpha_m BS(i,j) * \sum_n \sigma_n BO(i,j))$ explains when the boundary of the subjective map can be participated in the TBM if it is supported at-least by one objective map. The second part$\sum_n \sigma_n BO(i,j))$ controls that the boundary of an objective map can be directly participated in the TBM. α and σ are tuning parameters to weight the subjective and objective maps, respectively.

Next, we convert the 2D total boundary map to a graph, where each node in the graph represents a pixel from the 2D map domain and each node is connected to the eight connected neighbor pixels of the 2D map domain. The weight of the edge in the graph between two nodes is the sum of the related pixel values in the total boundary map TBM. Therefore, two pixels or nodes in the graph that have a low-weight connection are in a flat map area or, in other words, there is no boundary between them. Contrary, when two pixels or nodes have a high weight connection it shows that they are not in a flat map area and there is probably a boundary between them or, in other words, they are likely in different classes/clusters. When applying an all shortest path algorithm to the boundary graph, the relation between all pixels in the 2D map or the nodes in the graph will be clarified based on the boundary between them. If the edge of two nodes is part of a longer shortest path between two nodes then they are maybe in the same class and are related to each other, else they are not in the same class.

When the graph is big, the shortest path algorithm will be costly and timely expensive. For avoiding this problem, sampling techniques can be applied to the graph. Based on the total boundary map we determine a systematic sampling method to select a subset of pixels in the maps or nodes in the graph. We arrange all points of the total boundary map in a list from minimum value to maximum value. Then, from this ordered list a sample with size *s* can be selected by considering every k^{th} point from the list. Now, a shortest path algorithm can be applied to the graph with only calculating all shortest paths and path from these *s* nodes to all nodes in the graph. This results in two matrices (all shortest path and all path length) with *s* rows and *n* columns for each matrix where *n* is the total number of pixel in the map or all nodes in the graph. In the next step the same *s* rows from the similarity matrix resultant from the objective datasets will be selected which results in a Sim_{s*n} matrix. Then, based on the selected subset from the similarity

matrices (Sim_{s*n}), all shortest paths (SP_{s*n}), and all path lengths (PL_{s*n}) matrixes we create a new information vector IV_{s*n}

$$IV_{s*n} = \frac{Sim_{s*n}}{SP_{s*n}+PL_{s*n}} \quad , \quad (4\text{-}6)$$

with n samples or data points with *s* attributes or variable layers which carry the information about colors and boundary information of points in the map. The denominator of equation 5 decrees the similarity of the point when there is a boundary between them or a long path is passed between them. This new information vector will be the input for clustering. Different clustering algorithms can be applied to this new information vector. When the number of cluster is *k* the k-means clustering results in a *1*n* vector with values between 1 to k for each cell in the vector, which shows the cluster for a related data point in the 2D map.

4.4 The Datasets

4.4.1 The Synthetic Datasets

A synthetic example dataset is used to illustrate the efficiency and performance of the introduced method in this chapter. Three equally normalized maps in Figure 4-3 provide 2D information about the variability of properties in an exemplary survey region. The modeled database comprises one subjective map (Figure 4-3a) and two objective maps (Figure 4-3b and 4-3c). These maps are independent parameters which provide information about the same survey region. We consider 6 different structure in the whole

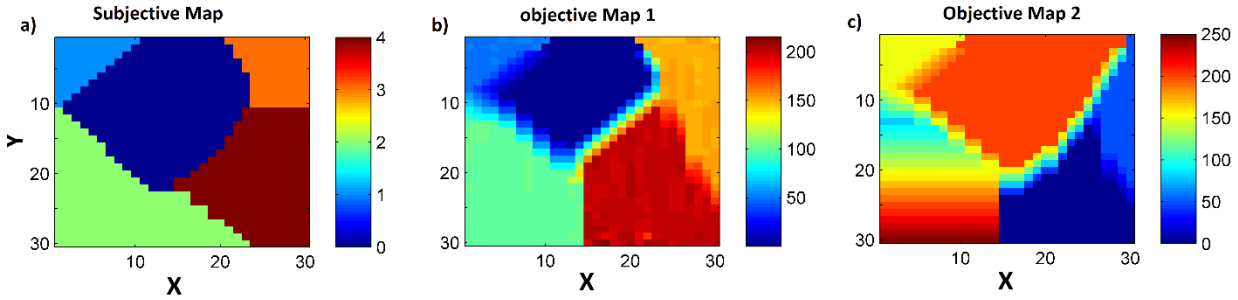

Figure 4-3: Synthetic datasets for a 30*30 2D domain with (a) subjective map, (b) and (c) objective technical maps. The subjective map shows structures mostly the same as in the technical maps, but with step boundaries and noise free. The technical maps show structures with different boundaries (i.e., step, line, roof, and ramp), noise, and anomalies from anthropogenic effects.

area. In each map there is some information which is not captured by another one. While the subjective map exhibits five zones characterized by step boundaries and noise-free information (see Figure 4-3a), the technical maps display different boundaries (i.e. step, ramp, or line) and partly different zones with some anomalies and environmental or technical noise superimposed, which are not captured by the subjective map (compare Figure 4-3a with 4-3b and 4-3c). These datasets are scaled to the interval [0 1].

4.4.2 The Field Datasets

The real-world datasets are measured at the Schäfertal research site (11°03¢E, 51°39¢N) located in the Lower Harz Mountains in central Germany (150 km southwest of Berlin) that is a small low-mountain catchment and is part of the long-term Earth observation network TERENO (Zacharias et al., 2011). Surface topography is V-shaped with a first-order stream in the valley and with gentle to moderate slopes (up to 20%) on both sides of the stream (Schröter et al., 2015). The site is determined by distinct landforms i.e., north-facing slope and south-facing slope, both intensively used for agriculture, the valley bottom with pasture or meadow, and topographic depressional areas disrupting the slopes on both sides of the stream (Schröter et al., 2015). The catchment is underlain by shale and Devonian greywacke covered by a complex of periglacial layers with different fractions of rock fragments and silt (Altermann, 1985). According to the sequence of the cover layer and landscape position different types of

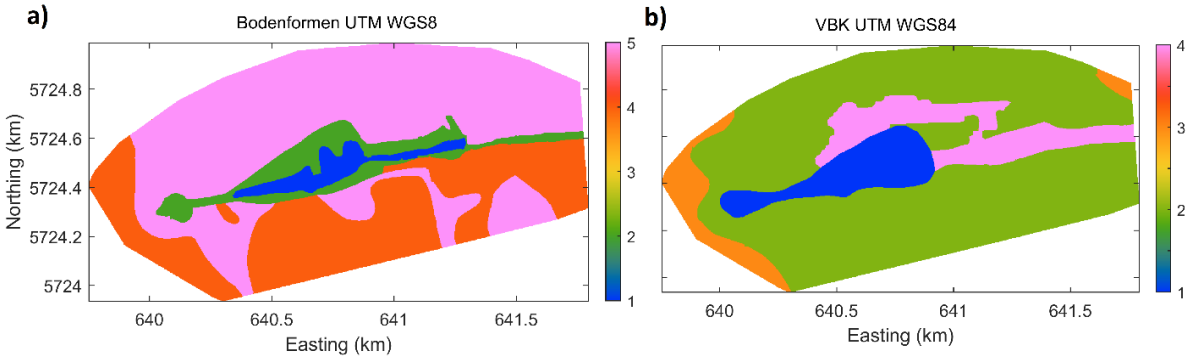

Figure 4-4: Two subjective maps of the Schäfertal catchment created by (a) (Borchardt 1985; Ollesch et al. 2005) and (b) Landesamt für Geologie und Bergwesen Sachsen-Anhalt. They show the structures in this catchments based on observations recorded by scientists.

soils have evolved from this area. The dominant soils comprise peaty Gleysols in the valley bottom, Luvisols and Cambisols on the hillslopes (Borchardt, 1982).

From this catchment subjective and technical datasets are exemplary considered in this study. Figure 4-4 presents two independent types of at least partly subjective soil maps created by different expert scientists to highlight the structure in the area with step boundary and noise free information (Borchardt 1985; Ollesch et al. 2005). Information about potential classification errors or the underlying sampling scheme is not available. Each map shows four segments or structures which are not completely similar. The diversity showed in these maps is related to the experience and beliefs of scientists who are responsible for sampling the ground properties and structures in the catchment. Different experiences can highlight different structures.

Figure 4-5 shows technical maps prepared and used by Schröter et al., (2015). These are relying on topographic information in the Schäfertal catchment obtained from a high-

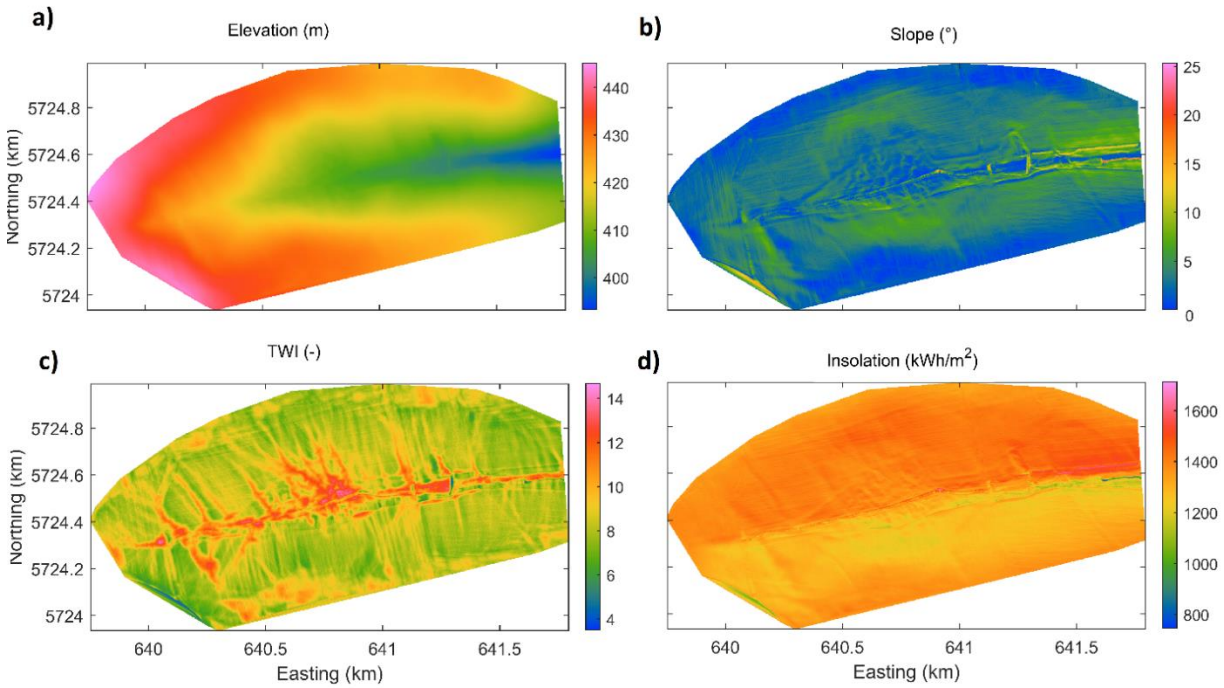

Figure 4-5: Four attributes shown as objective or technical maps; (a) elevation, (b) slope, (c) SAGA wetness index, and (d) annual potential incoming solar radiation derived form a 2-m digital elevation model for the Schäfertal catchment (Schröter et al., 2015).

resolution 1 * 1 m^2 digital elevation model measured by air-borne laser scanner. Figure 4-5a-d present elevation, slope, topographic wetness index (TWI), and total annual incoming solar radiation (TIR), respectively, which are all quantities derived from the degitial elevation model Each map independently contributes information about contextual and local landscape conditions commonly used in the literature (Western et al., 1999; Wilson et al., 2005; Takagi and Lin, 2012, Schröter et al., 2015). The elevation (Figure 4-5a) can be used to describe the gravitational potential energy and landscape position that drives water flow. The slope information (Figure 4-5b) is indicative of the hydraulic gradient which drives near-subsurface and surface fluxes (Western et al., 1999). The TWI (Figure 4-5c) presents zones of surface saturation for a more realistic prediction for cells situated in valley floors with small vertical distance to a channel (Böhner and Selige, 2006). The Insolation (Figure 4-5d) presents annual potential incoming solar radiation derived for the Schäfertal catchment. This technical information (Figure 4-5a - d) are related to hydrological processes controlling the spatial distribution of soil moisture. For example, such information was used by Schröter et al., (2015) in their fuzzy c-means clustering for segmentation of the area to optimize the sampling support for soil moisture measurements in order to predict the spatial soil moisture based on the clustering results.

4.5 Experiments and Results

4.5.1 Application to the Synthetic Dataset

To evaluate the efficiency and performance of our introduced method we apply the processing steps to the synthetic dataset (Figure 4-3). First of all, after normalizing the data based on equation 3, we calculate the similarity for the technical maps (Figure 4-3a and 3b) of the synthetic dataset. The 2D maps are a 30*30 matrix and comprise 900 data points. We use the Gaussian similarity based on equation 4 for calculating point to point similarities, which results in a 900*900 similarity matrix shown in Figure 4-6. In parallel the boundary detection techniques will be applied to the subjective and objective maps (Figure 4-3a-c) to extract the boundary information. For extracting the boundaries of the subjective map (Figure 4-3a) we use Canny's boundary detection. For the objective maps

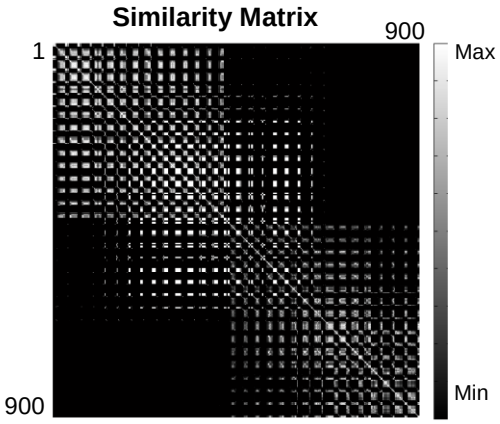

Figure 4-6: 900*900 Gaussian similarity matrix for the data points in the synthetic dataset based on the absolute values in technical maps. The similarity of points is a value between 0 and 1.

we use χ^2 distance to highlight the boundary information shown in Figure 4-7a-c related to the maps in Figure 4-3a-c, respectively.

Using equation 5 we calculate the total boundary information of the boundary maps (shown in Figure 4-7a-c) that results in a 2D total boundary map presented in Figure 4-7d. This map carries the boundary information of all subjective and technical maps. Here, we set the tuned parameters as $\alpha = 1$ and $\sigma = 1$ for weighting and summing the boundary information. In this case we trust completely the subjective and technical maps

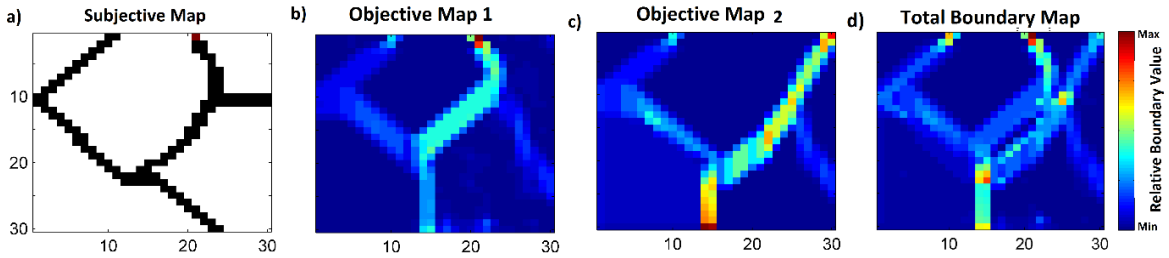

Figure 4-7: Boundary information for the maps in the synthetic datasets. (a) Extracted boundaries based on Canny boundary detection for the subjective map, (b) and (c) extracted boundaries based on the χ^2 distances for technical maps, (d) calculated total boundary map of subjective and objective boundary maps.

and a boundary point of subjective boundary will be accepted if it will be supported by a technical map. For this reason the boundary shown in the range between x=15 - 23 and y= 24 -30 is not shown in the total boundary map because it is not supported by any boundary in the technical maps. In this test we assume that all boundaries of technical maps will contribute to the total boundary map.

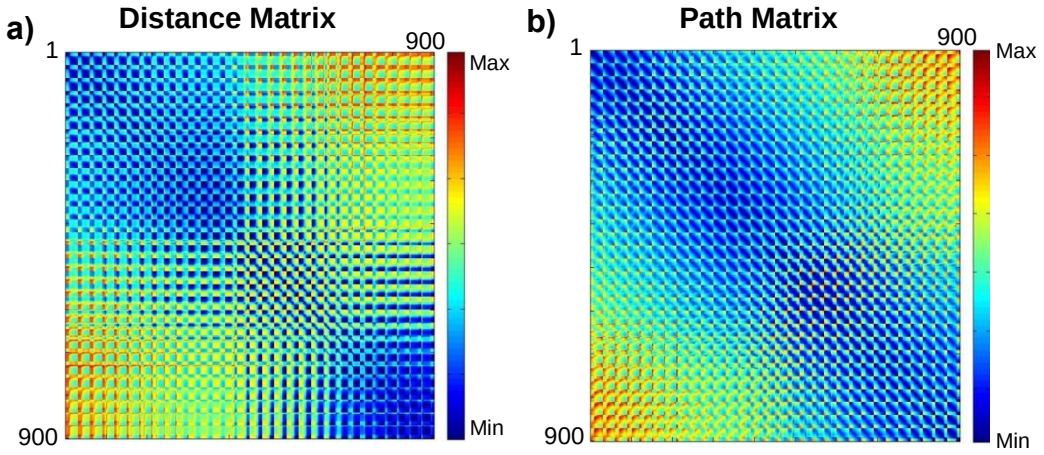

Figure 4-8: 900*900 matrices resultant from Dijkstra algorithm (a) all shortest paths and (b) all path lengths for data points in the synthetic dataset.

In the next step, the total boundary map will be converted into a graph by calculation of all shortest paths and path lengths. In this experiment, because of the low number of data points (900 data points), we take the sample size equal to the number of data points in the 2D map. We calculate all shortest paths between the 900 data points which results in two 900*900 matrices for shortest path and path length, shown in Figure 4-8a and 8b, respectively.

Applying equation 6 to the selected samples from the similarity matrices related to shortest paths and path lengths we define the new information vectors shown in Figure 4.9 for 900 data points (the number of columns) and 900 attributes or variables (the number of rows). This new information vectors will be the input for k-means clustering with considering *k*=6 in the whole area. Figure 4-10 shows the results of k-means

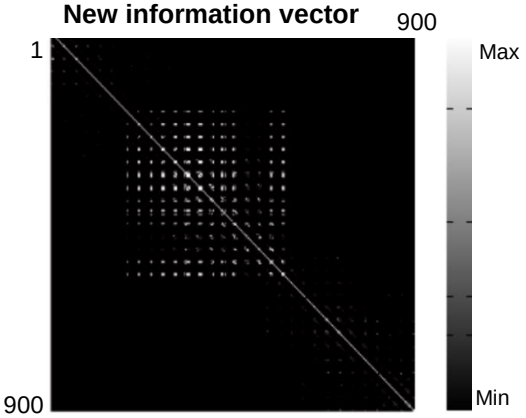

Figure 4-9: 900*900 matrix as new information vector resultant from similarity, shortest paths, and path lengths matrices. The columns show 900 samples or data points and rows show 900 attributes or variable layers. This matrix carries the information about colors and boundary information of points in the map.

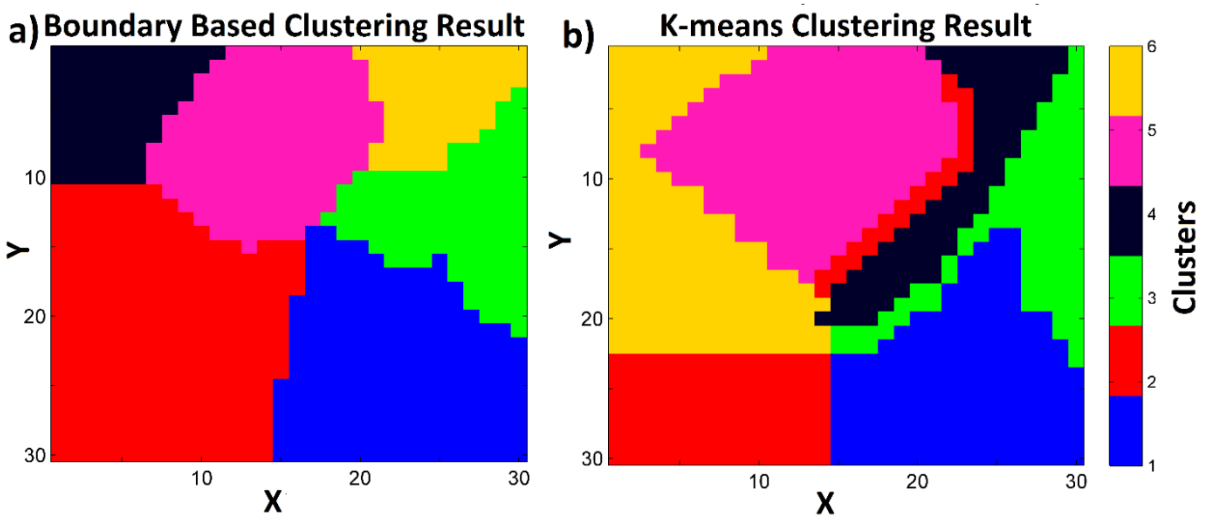

Figure 4-10: k-means clustering results for the synthetic dataset with two strategies. (a) The k-means clustering results based on the new information vector resultant from the introduced strategy in this chapter, (b) the results of clustering without considering the boundary information of the subjective map. 6 cluster are desired in this dataset.

clustering with two strategies. Figure 4-10a presents the k-means clustering results based on the new information vectors resultant from the introduced strategy in this chapter. Figure 4-10b shows the results of clustering without considering the edge information, only taking the measured values (color) of the data points into account. When comparing these results to Figure 4-3a-c, it shows that the introduced strategy in this chapter is able to consider the boundary information of the maps and taking the subjective information into account which improves the clustering results. The results achieved when only using the measured values are shown in Figure 4-10b. This type of clustering has problem in dealing with roof and ramp type boundaries and handling anthropogenic effects or unusual structures in the date sets (i.e., the gradient noise in at left part of Figure 4-3c effects in the clustering resultant as separated clusters in the left part presented in Figure 4-10b).

4.5.2 Application to the Field Dataset

We apply the introduced strategy to the field datasets shown in Figures 4-4 and 4-5. There are 364096 data points in these maps. Calculating the similarity based on the normalized objective maps (shown in Figure 4-5) results in a 364096*364096 similarity

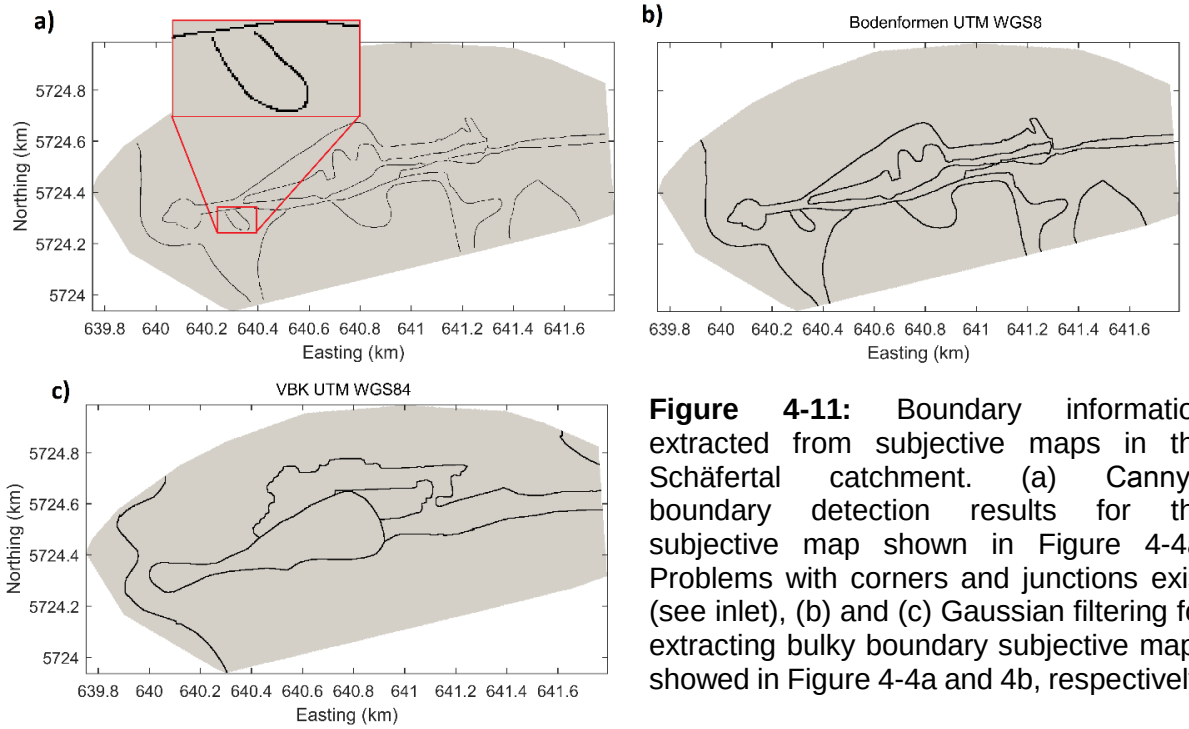

Figure 4-11: Boundary information extracted from subjective maps in the Schäfertal catchment. (a) Canny's boundary detection results for the subjective map shown in Figure 4-4a. Problems with corners and junctions exist (see inlet), (b) and (c) Gaussian filtering for extracting bulky boundary subjective maps showed in Figure 4-4a and 4b, respectively.

matrix. In parallel the boundary detection will be applied to the subjective and objective maps. First, we apply Canny's boundary detection to the subjective maps. The results of this method are shown in Figure 4-11a for the subjective map shown in Figure 4-4a. Figure 4-11a shows that Canny's boundary detection has some problems with corners and junctions (see Figure 4-11a). Therefore, we use Gaussian filtering for solving this problem which results in bulky boundaries different than the sharp edges of Canny's boundary detector but strongly connected. The results of the Gaussian filtering boundary detection are presented in Figure 4-11b and 11c for the subjective maps presented in Figure 4-4a and 4b, respectively. For extracting the boundaries of the objective maps shown in Figure 4-5 we use the χ^2 distance with a radius equal to 10 pixels for the disc and 5 bins for the histogram in each half disc. Because the slope (Figure 4-5b) is the derivative (or boundary) of elevation (Figure 4-5a), we only apply the χ^2 to the TWI and insolation maps (Figure 4-5c and 4-5d). Slope information will directly participate in the boundary detection procedure of objective maps as a boundary information of the elevation map. Figure 4-12a and 4-12b present the results of χ^2 distance for the TWI and insolation datasets (Figure 4-5c and 4-5d), respectively.

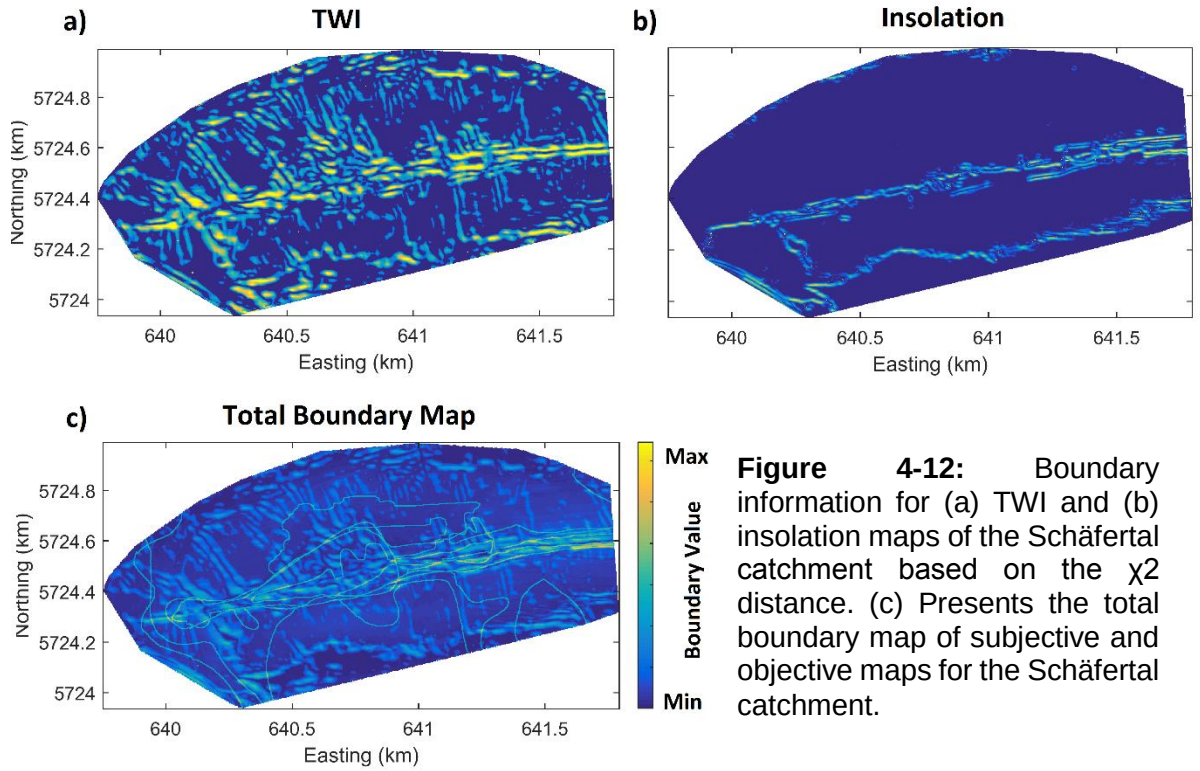

Figure 4-12: Boundary information for (a) TWI and (b) insolation maps of the Schäfertal catchment based on the χ2 distance. (c) Presents the total boundary map of subjective and objective maps for the Schäfertal catchment.

Then, the total boundary map will be calculated based on the weighting of the extracted boundary maps of subjective and objective maps presented in Figure 4-11 a-c and Figure 4-12a-b. When summing the subjective boundaries, they can be weighted based on their accuracy. We assume that the subjective map shown in Figure 4-4a is 90% true and the subjective map shown in Figure 4-4b is 70% true, therefore, we set the tuning parameters α_1 and α_2 equal to 0.9 and 0.7, respectively. We set the tuning parameter σ=1 for all technical maps. Figure 4-12c shows the total boundary map calculated based on the maps shown in Figure 4-11a-c and Figure 4-12a-b by means of equation 5.

In the next step, the total boundary map will be converted to the graphs by calculating all shortest paths and path lengths. Because of the high number of data points (364096 data points) we test the introduced method with a sampling size *s*=1000 data points in the 2D map. Figure 4-13a shows the *s*=1000 sampling points with systematic sampling selection strategy. We calculate all shortest paths from the selected 1000 data points to all other data points in the 2D map which results in $SP_{1000*364096}$ and $PL_{1000*364096}$

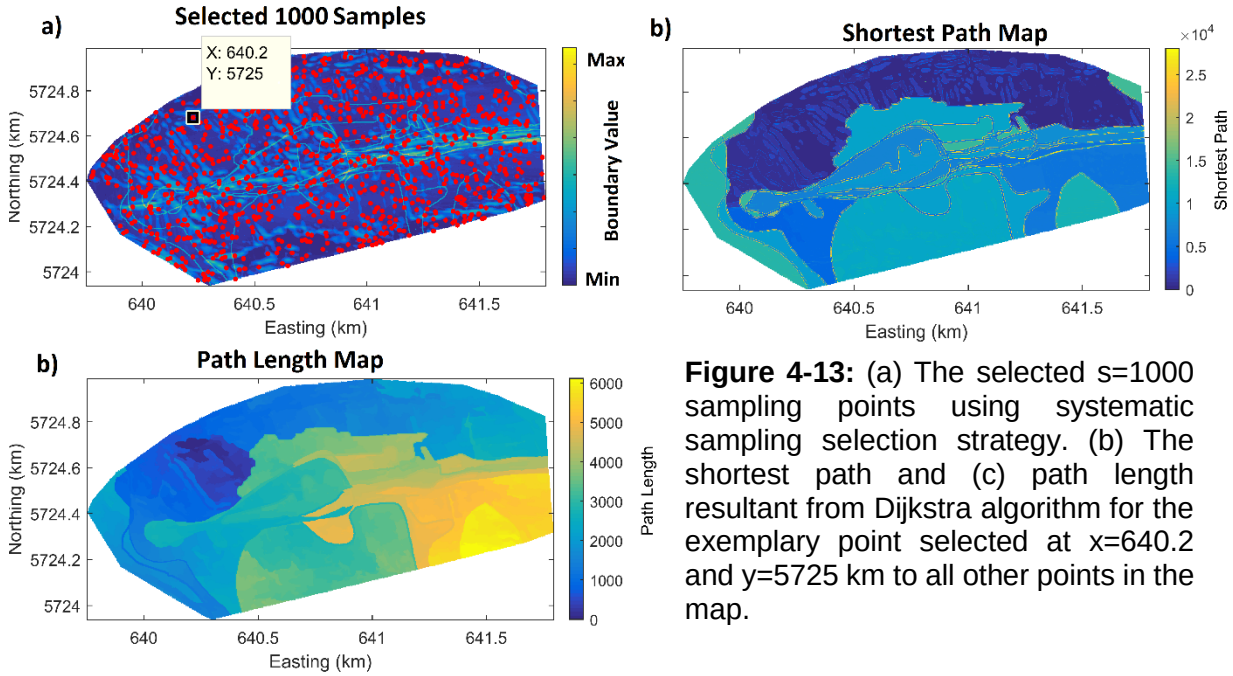

Figure 4-13: (a) The selected s=1000 sampling points using systematic sampling selection strategy. (b) The shortest path and (c) path length resultant from Dijkstra algorithm for the exemplary point selected at x=640.2 and y=5725 km to all other points in the map.

for shortest path and path length matrices, respectively. Figure 4-13b and 4-3c show the shortest path and path length, respectively, for the exemplary point selected in Figure 4-13a at x=640.2 and y=5725 km to all other points in the map.

From the similarity matrix, we extract the similarities of the selected 1000 samples to all other n=364096 data points that results in a matrix of size 1000*364096. Applying equation 6 to the selected samples and the corresponding shortest paths and path lengths we shape the new information vector $IV_{1000*364096}$ with 364096 data points (the number of columns) and 1000 attributes or variables (the number of rows). This new information vector will be the input for k-means clustering with k=30. We use the same number of clusters determined by Schröter et al., (2015) which delineates the major sub-surface zonation.

Figure 4-14 shows the results of k-means clustering following two different strategies. Figure 4-14a presents the k-means clustering results on the new information vector resultant from the newly introduced strategy. Figure 4-14b shows the results of clustering without considering the edge information as done by Schröter et al., (2015). As shown in Figure 4-14b the method used by Schröter et al., (2015) has problems, i.e. by finding highly nested clusters, and it cannot cope with anthropogenic effects, such as

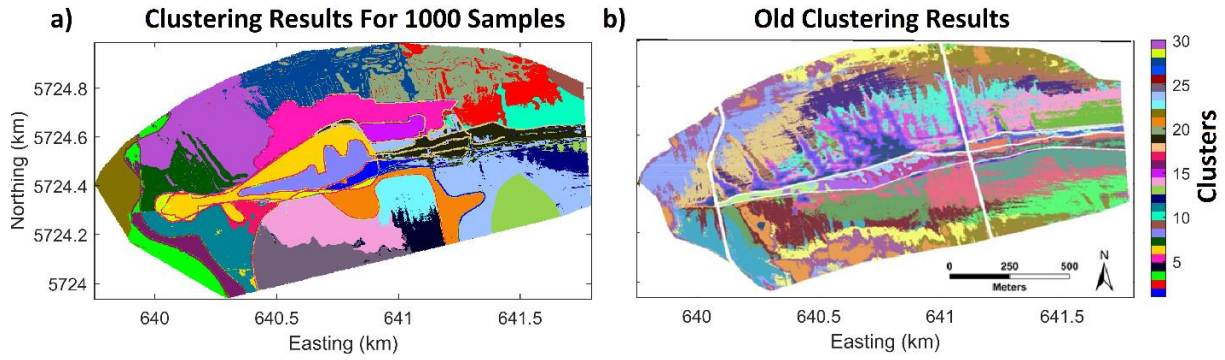

Figure 4-14: The results of clustering the maps of the Schäfertal catchment with two strategies,(a) the k-means clustering results on the new information vector resultant from the introduced strategy in this chapter, (b) presents the results of clustering without considering the boundary information (Schröter et al., 2015). 30 clusters are desired in this catchment, each color determine an independent cluster.

driving lanes, in the data. When taking the subjective maps into account and incorporating boundary information our method is more robust. When there is a big structure in the subjective map the introduced method tries to separate this area based on the edge information and absolute values of the technical maps (see northern part in the map shown in Figure 4-14a). When the size of a structure is small this method try to determine a cluster result based on the subjective information (see the valley part of Figure 4-14a). Because of the optimum number of the samples is not achieved (I select only 1000 samples) in some parts the method shows some anomaly or nested clusters which with finding the optimum sample number this problem can be solved. Another shortcoming of this method is due to the bulky boundaries of the subjective maps, which are presented

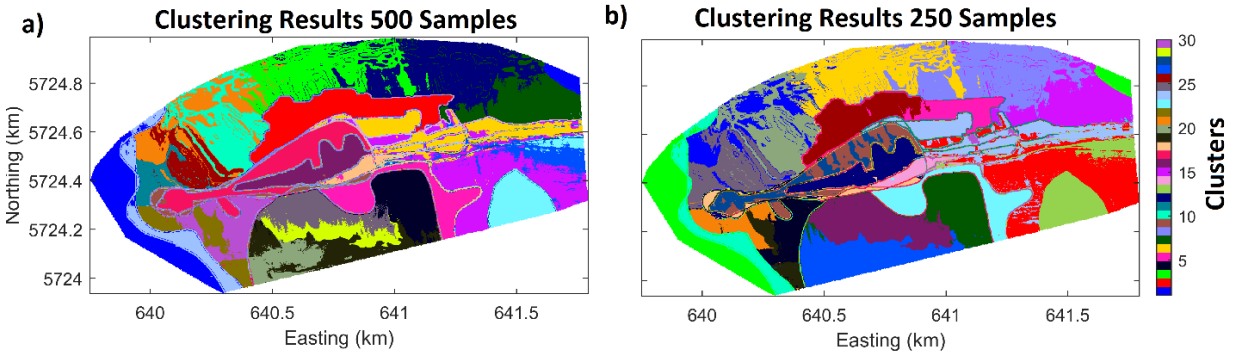

Figure 4-15: The same as in Figure 4-14a, but now with sampling size reduced to (a) 500 and (b) 250 samples.

as an individual class. For testing the stability with regard to the sample size in the introduced method, Figure 4-15a and 4-15b show the clustering results for only 500 and 250 samples, respectively. These samples are a subset of the 1000 samples used before. Comparing Figures 4-14a, 4-15a, and 4-15b reveals that in most parts of the 2D area the introduced method is able to offer stable results for clustering. The differences illustrate the effects of the number of sampling points on the results of this integration method.

4.6 Conclusions

Traditional integration and segmentation methods do not allow taking subjective maps (i.e., soil or geological maps, which inherently carry beliefs of the scientists about the catchments) into account, such that inherent data characteristics are acknowledged. Our straightforward and rapid integration and segmentation method has been designed such that subjective maps can be integrated with objective or technical maps via the boundary information provided by subjective maps. We can weigh the subjective and technical maps relative to each other thus increasing the performance of the integration and segmentation method and matching them to the scientists understanding. At the same time, we experience an increasing ability of the clustering against noise and unusual structures. We have employed boundary detection, graph theory, sampling, and cluster analysis to integrate a multi-parameter spatial database comprising partly noise free subjective and technical datasets that inherently carry noise. The technical datasets need to be normalized prior to the integration procedure. Effects of the chosen normalization procedure may exist (see Appendix H). New information vectors are created based on boundary information and the measured values in the available technical maps. Then a crisp clustering algorithm like k-means enabled the rapid and automatic integration of a segmented integrated map delineating distinguished sub-surface units upon the information provided by each subjective and technical dataset. This methodology can be applied to any combination of the subjective and objective maps to offer a better segmentation of the considered domain. More testing is certainly necessary to show investigate the flexibility, efficiency, applicability, and robustness of the approach, particularly in view of the chosen sampling strategy and data normalization.

Chapter 5
Summary and outlook

5.1 Data Mining Techniques Add Value to Geophysical Tomography and Logging Data

The main core of this thesis was putting forward different objectives towards offering knowledge discovery in environmental datasets. I focused on geophysical tomography which is one important type of environmental datasets that offers valuable and unique information about the internal composition of the ground. In chapter 2 and 3 I had answered the most important challenge when using geophysical tomography in hydrological, environmental or engineering exploration, that was, how to link the tomographically reconstructed physical parameter variations to the aquifer, reservoir or geotechnical target parameters of interest, which are usually different from those imaged by geophysical tomography. One major goal of this thesis which I had presented in chapter 2 was to develop a framework based on artificial neural networks (ANNs) for 2D or 3D probabilistic prediction of sparsely measured Earth properties constrained by ill-posed geophysical tomographic imaging acting here as preprocessing of the measured database. The structure of this framework follows KDD's structure (see Figure1-1). Tomogram inversion is the preprocessing of the available traveltime datasets, in the data mining part the ANN learn the optimal link between tomograms and target parameters, due to the probabilistic nature and error accounting in this method the evaluation part is limited to interpretation, but not critical interaction or rerun of the ANN. First I showed the application of this method based on a realistic but synthetic database that allows for

optimal performance evaluation of the suggested methodology containing different 2D radar and seismic tomograms and 1D porosity as target parameter. The employed static two-layer feed forward ANNs, based on the introduced strategy in chapter 2 of this thesis can successfully fit the underlying datasets equally well in a way that I was able to transduce tomographic reconstruction ambiguity into the prediction of a target parameter. The prediction results of this chapter are of higher relevance to hydrologic and engineering exploration tasks than the tomographically imaged parameters. When combined with fully non-linear (globally searching) geophysical tomographic imaging I demonstrated that this methodology can deliver objective and purely data-driven probabilistic predictions of target parameter distributions, which are essentially required when striving to assess, quantify and minimize risks in subsurface exploration and utilization. A classical user-based tuning or result evaluation interacting with the machine learning or prediction task is here of minor importance since the major uncertainties in the available database, resulting from tomographic reconstruction ambiguity are already taken into account in the probabilistic prediction approach.

Furthermore, in chapter 2 I evaluated, whether the performance of the trained ANNs, measured by MSE, can be used to rank the equivalent geophysical tomograms. Fundamental idea which I was looking at to prove this assumption was that tomograms as well as sparse information about an exploration target parameter are images of the same reality and must therefore be compliant. In our synthetic database I could analyze this question which would be practically impossible when working with field data. A rather qualitative statement about the closeness of the tomograms to reality can be made based on the ranking results achieved by ANN training performance, i.e., tomograms ranked low suffer an increased risk of being poor reconstructions of reality. However, outliers from this rule may exist and therefore question the benefits from utilization of recurrent ANNs striving to learn which tomograms may be particularly useful for prediction based on the available database. I found that such approach would build the prediction of target parameter distributions on a few tomograms of high importance, but facing the risk that eventually a poor tomographic model will be considered with high weights, which leaves doubts on the chances to achieve better predictions when using recurrent ANNs instead

of the simple feed-forward ANNs used in chapter 2 and 3. Based on this finding it seems not promising to develop complex tools for incorporating the information of the logging data in the solution of the tomographic reconstruction problem when following approaches resulting in most-likely or best-fit computations without giving a probabilistic overview about the possible solution range.

5.1.1 Taking Data Uncertainty into Account in the Machine Learning Prediction Part

Another main challenge in the environmental datasets is uncertainty of measured data (e.g., borehole logging data errors) or ambiguity resulted from preprocessed datasets (e.g., tomographic ambiguity of the inversion results). For realistic and objective predictions of geotechnical or hydrological target parameters the uncertainty and differences in spatial resolution must be taken into account. In chapter 3 of this thesis I have shown an application of the method introduced in chapter 2 to solve real-world problems. Furthermore, I have improved this method by considering the uncertainty or variability of the input data when offering a probabilistic prediction of the target parameters. When doing so, it is important to incorporate uncertainties from tomographic imaging and the target logging data in the training phase of the ANN to avoid overfitting the training data by the ANN. Depending on the different training strategy introduced in chapter 3, my introduced method resulted in focused probabilistic predictions with smaller ranges suitable to assess the most likely values of the target parameters in the 2D tomographic plane. I have shown that ANNs can be trained such that even small-scale anomalies beyond the spatial resolution of the tomograms are considered, which resulted in broad and rather conservative prediction ranges, which do not significantly distort the most-likely predictions. I believe that the approach followed in chapter 3, taking the uncertainty of tomograms and logging data ambiguity into account for probabilistic prediction of target parameters, can help in a variety of geophysical applications to analyze and identify complex parameter relations which cannot be described by traditional petrophysical models.

5.1.2 Choosing Optimum Parameters for Artificial Neural Networks

One drawback of my approach for probabilistic prediction of hydrological or geotechnical target parameters introduced in chapters 2 and 3 is the lack of robust criteria for determining the optimum ANN architecture, that means finding the optimum number of neurons in the layers, selecting the best activation function for the hidden layer, and selecting the best training strategy. I used the most common approaches to determine the optimum number of neurons in the hidden layer that start with a very small number of neurons (see chapter 2 or 3) estimating the mean squared error of the prediction results. I had repeated the procedure increasing the number of neurons in the hidden layer. In this case I had selected the optimum number of neuron in the hidden layer so that the related artificial neural network model offer the minimum mean squared error. Another important parameter when setting up the neural networks is selecting the activation function of the neurons in the hidden layer which can be step, linear combination, softmax, or sigmoid functions. I have tried different functions, but the best one was a sigmoid function since the combination of sigmoid function with neurons in one hidden layer ANN was sufficient for all prediction applications. While employing the ANN for probabilistic prediction of geophysical target parameters, I found in chapter 2 and 3 that the results were only weakly dependent on the number of neurons in the hidden layer and the type of activation function. In future works when applying these introduced methods in the distinct datasets, different strategy can be tested for selecting the optimum number of neurons and activation function to prove the training phase and the performance of the achieved neural networks prediction model. In future applications I suggest to follow the strategy introduced by Yuan et al. (2003) for estimating the number of hidden neurons in feed-forward neural networks based on information entropy, or the strategy introuduced by Benardos and Vosniakos (2007) for optimizing feedforward artificial neural network architecture. Furthermore, different training strategy like Bayesian regularization backpropagation (Kay, 1992), resilient backpropagation (Riedmiller and Braun, 1993), and structure like cascade-forward neural network (Fahlman and Lebiere, 1990) or new techniques like deep learning (Schmidhuber, 2015) can be tested for evaluating and finding the best ANN structure.

When incorporating aggregated relative errors in the training of the ANNs (the WMSE measure in chapter 3), I simply summed relative range information. However, weighted MSE computation is usually relying on Gaussian distribution of errors, e.g., standard deviations. In many practical cases we do not know the distribution and use Gaussian assumptions due to their mathematically simple implementation and description. However, improvements, and a more rigorous incorporation of uncertainties, may require developing improved error models moving away from Gaussian assumptions that may allow for more realistic selections than existing approaches. This may probably be a challenging task requiring the development of mathematic error models away from simple statistical ideas, such as standard deviations etc.

5.1.3 Testing with Different Tomograms and Target Parameters

The prediction performance of the introduced methods in chapter 2 and 3 was excellent, and the offered methods in these chapters can be applied to any combination of geophysical tomograms and target parameters since at no point critical assumptions about the involved parameters or the expected relations between the considered datasets and parameters are made. For showing the applicability of the introduced methods more combinations of geophysical tomograms and geotechnical or hydrological logging data should be tested in future works. While my developed approach in chapters 2 and 3 is highly flexible and applicable, it would be a relatively trivial matter to incorporate 3D tomographic datasets to offer 3D probabilistic prediction of the target parameters. However, more interesting would be to employ tomographic datasets solely acquired from the Earth's surface. For such data, tomographic ambiguity increases systematically with depth. Consequently, ANNs would learn the relations between tomograms and logging data primarily at the very near surface, since tomographic ambiguity, and thus the considered errors, will systematically increase with depth. Depending on the specific ground composition in some cases, additional considerations may be necessary to ensure that also tomogram-logging data relations at greater depths will contribute to the learned prediction model.

5.2 Human in the Loop for Mapping, Integration, and Segmentation of Geophysical Datasets

Interactive mining of knowledge or human in the loop at multiple levels of knowledge discovery models are an issue that, when considered during the KDD process, can help to know exactly what can be extracted from a dataset. A fundamental target in earth sciences and environmental study which can get benefit of human in the loop of KDD process is mapping that address various environmental and economic issues, such as mining target identification, soil conservation, or ecosystem management. In chapter 4, I have discussed the conceptual idea of a workflow towards integration and segmentation of environmental datasets considering subjective data of a human or expert scientist in a logical and acceptable way, such that it matches with the experience of a geophysical scientist. For doing this I employed boundary detection, graph theory, selected sampling, and cluster analysis to integrate a multi-parameter geophysical database comprising subjective and technical datasets that inherently carry noise and erroneous information. The obtained clustered multi-physics map projects multi-parameter information emanating from the underlying subjective and technical datasets onto a two-dimensional map. The resultant segmented 2D map can be used to develop optimal sampling schemes including all major segments, i.e., defining the locations of a limited number of sampling locations to monitor near-surface or subsurface catchments, or the earth properties at the small catchment scale. I have shown the efficiency of the introduced method by applying it on a synthetic dataset and a real world problem recorded in the Schäfertal catchment, which is part of the TERENO Harz/Central German Lowland Observatory. In chapter 4, by using the idea of the human in the loop I was able to: (i) take the scientists experiments into account, (ii) determine a way to weighing the subjective and technical maps (iii) increasing the performance of the integration and segmentation method and matching them to the scientists understanding with solving the nested clustering problem and increasing their robustness against noise and unusual structures, and (iv) to add value to the subjective data.

5.2.1 Further Testing the Idea of the Human in the Loop for Integration and Segmentation of Geoscientific Datasets

I have shown that the idea of the tacking the subjective data can prove the integration and segmentation results and it increase the ability of the integration method to work with noise, anomaly and unusual structure. But for proving and showing the ability of this method more field and laboratory analysis and testing is necessary to assign a geological meaning to the achieved zones in the 2D segmented map. Consequently, for getting deeper insight into environmental or subsurface transactions and properties more subjective or technical datasets (i.e., hyperspectral, multispectral datasets, and etc.) should participate in the integration analysis. I recommend that for the Schäfertal catchment more data and more field and laboratory analysis should be done to achieve a better understanding of the detected segments. This should go in line with a more thorough testing of the suggested methodology. For example, within this thesis it was not possible to prove or carefully investigate the impacts of the individual processing settings and selections on the final segmentation. For example this incorporates the investigation of different edge detection features, different similarity measures or graph partitioning techniques. The approach is flexible enough to incorporate additional image texture features, which may improve broaden the analyses on which the segmentation patterns are learned.

Furthermore, in future works the introduced methods in chapters 3 and 4 can be combined to offer a probabilistic prediction of desired target parameters in the Schäfertal catchment. I believe that the combination of these two methods can offer great prediction results for recognizing the distribution of target parameters in this area. When running the ANNs employed in the prediction procedure on fuzzy representations of the segmented input database, subjective and technical maps can be evenly considered in the prediction procedure without the need to teach the ANNs learning the prediction model how to deal with human data in the processing loop. Compared to Schröter et al., (2015), who use a fuzzy segmented map for deterministic interpolation of sparse soil moisture measurements solely considering technical maps, such a combined approach could produce fuzzy segmented maps incorporating additional subjective data and the

probabilistic prediction approach could provide likelihood information for soil moisture all over the catchment.

References

Abrahamsson, M. (2002). *Uncertainty in quantitative risk analysis-characterisation and methods of treatment*. *LUTVDG/TVBB--1024--SE*.

Aggarwal, C. C., & Philip, S. Y. (2009). A survey of uncertain data algorithms and applications. *IEEE Transactions on Knowledge a.nd Data Engineering*, *21*(5), 609-623.

Aggarwal, C. C. (Ed.). (2010). *Managing and mining uncertain data*. Springer Science & Business Media, Vol 35.

Ahn, C. W., Baumgardner, M. F., & Biehl, L. L. (1999). Delineation of soil variability using geostatistics and fuzzy clustering analyses of hyperspectral data: Soil Science Society of America Journal, 63, 142–150

Albalate, A., & Minker, W. (2013). Semi-Supervised and Unsupervised Machine Learning: Novel Strategies. John Wiley & Sons.

Al-Hegami, A. S. (2004). Subjective measures and their role in data mining process. In Proceedings of the 6th International Conference on Cognitive Systems, New Delhi, India.

Altermann, M. (1985). Standortkennzeichnung landwirtschaftlich genutzter Gebiete des östlichen Harzes, Habilitationsschrift. Univ. Rostock, Germany.

Anderson-Mayes, A. M. (2002). Strategies to improve information extraction from multivariate geophysical data suites: Exploration Geophysics, 33, 57–64. doi: 10.1071/EG02057

Angioni, T., Rechtien, R. D., Cardimona, S. J., & Luna, R. (2003). Crosshole seismic tomography and borehole logging for engineering site characterization in Sikeston. Tectonophysics, 368, 119–137.

Ankerst, M. (2001). Human involvement and interactivity of the next generation's data mining tools. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*.

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(5), 898-916.

Archie, G. E. (1942). The electrical resistivity log as an aid in determining some reservoir characteristics. Trans. Americ. Inst. Mineral. Met, 146, 54-62.

Asadi, A., Dietrich, P., & Paasche, H. (2016). 2D probabilistic prediction of sparsely measured geotechnical parameters constrained by tomographic ambiguity and measurement errors. Expanded abstracts of the 78th EAGE Conference and Exhibition, Vienna, 2016. Doi: 10.3997/2214-4609.201601402.

Aster, R.C., Borchers, B., & Thurber, C.H. (2005). Parameter Estimation and Inverse problems. Academic Press.

Bailey, T. C., & Gatrell, A. C. (1995). *Interactive spatial data analysis* (Vol. 413). Essex: Longman Scientific & Technical.

Balling, R. (2003). The maximin fitness function; Multi-objective city and regional planning. In: C. M. Fonseca, P. J. Fleming, E. Zitzler, K. Deb, and L. Thiele, eds., Second international conference on evolutionary multi-criterion optimization, Springer Lecture Notes in Computer Science, vol. 2632, 1 – 15.

Bailly, K. & Milgram, M. (2009). Boosting feature selection for Neural Network based regression. Neural Networks, 22, 748-756.

Ban, J.C. & Chang, C.H. (2013). The learning problem of multi-layer neural networks. Neural Networks, 46, 116-123.

Beale, M. H., Hagan, M. T., & Demuth, H. B. (1992). Neural Network Toolbox™ User's Guide. R2014a ed, 2014.

Behrens, T., Zhu, A. X., Schmidt, K., & Scholten, T. (2010). Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma, 155(3), 175-185.

Benardos, P. G., & Vosniakos, G. C. (2007). Optimizing feedforward artificial neural network sarchitecture. Engineering Applications of Artificial Intelligence, 20(3), 365-382.

Bevington, P. R., & Robinson, D. K. (1992). Data reduction and error analysis for the physical sciences. New York: McGraw-Hill.

Bevington, P. R., & Robinson, D. K. (2003). Data reduction and error analysis. *McGraw-Hill*.

Binley, A., Winship, P., & Middelton, R. (2001). High-resolution characterization of vadose zone dynamics using cross-borehole radar. Water Resources Research, 37, 2639-2652.
Böhner, J., & T. Selige. (2006). Spatial prediction of soil attributes using terrain analysis and climate regionalisation. In: J. Böhner, K.R. McCloy, and J. Strobl, editors, SAGA–Analyses and modelling applications. Göttinger Geographische Abhandlungen, Göttingen, Germany. p. 13–28.
Bodin, T. & Sambridge, M. (2009). Seismic tomography with the reversible jump algorithm. Geophysical Journal International, 178, 1411–1436.
Bodin, T., Sambridge, M., Rawlinson, N, & Arroucau, P. (2012). Transdimensional tomography with unknown data noise. Geophysical Journal International, 189, 1536–1556.
Boisclair, C.D., Gloaguen, E., Marcotte, D., & Giroux, B. (2011). Heterogeneous aquifer characterization from ground-penetrating radar tomography and borehole hydrogeophysical data using nonlinear Bayesian simulations. Geophysics, 76, J13–J25.
Borchardt, D. (1982). Geoökologische Erkundung und hydrologische Analyse von Kleineinzugsgebieten des unteren Mittelgebirgsbereiches, dargestellt am Beispiel von Experimentalgebieten der oberen Selke/ Harz. Petermanns Geogr. Mitt. 482:251–262.
Bosch, M., Mukerji, T., & Gonzalez, F. E. (2010). Seismic inversion for reservoir properties combining statistical rock physics and geostatistics: a review. Geophysics, 75, 165-176.
Bragato, G. (2004). Fuzzy continuous classification and spatial interpolation in conventional soil survey for soil mapping of the lower Piave plain: Geoderma, 118, 1–16.
Breiman, L., & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc, 80, 580-598.
Burfield, C. (2013). Floyd-Warshall Algorithm. Massachusetts Institute of Technology.
Canny, J., 1986, A computational approach to edge detection. IEEE Transactions on pattern analysis and machine intelligence, (6), 679-698.
Cassiani, G., Böhm, G., Vesnaver, A., & Nicolich, R. (1998). A geostatistical framework for incorporating seismic tomography auxiliary data into hydraulic conductivity. Journal of Hydrology, 206, 58–74.
Cawley, G. C, Janacek, G. J., Haylock, M. R., & Dorling, S. R. (2007). Predictive uncertainty in environmental modelling. Neural Networks, 20, 537-549.
Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]. IEEE Transactions on Neural Networks, 20(3), 542-542.
Chen, J., Hubbard, S., & Rubin, Y. (2001). Estimating the hydraulic conductivity at the South Oyster Site from geophysical tomographic data using Bayesian techniques based on the normal linear regression. Water Resources Research, 37, 1603–1613.
Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.
Dafflon, B., Irving, J., & Holliger, K. (2009). Simulated-annealing-based conditional simulation for the local-scale characterization of heterogeneous aquifers. Journal of Applied Geophysics, 68, 60-70.
Danielsson, P. E. (1980). Euclidean distance mapping. Computer Graphics and image processing, 14(3), 227-248
Davis, L. S. (1975). A survey of edge detection techniques. Computer graphics and image processing, 4(3), 248-270.
de Bruin, S., and Stein, A., 1998, Soil-landscape modelling using fuzzy c-means clustering of attribute data derived from digital elevation model (DEM): Geoderma, 83, 17–33. doi: 10.1016/S0016-7061(97)00143-2
Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numerische Mathematik. 1: 269–271. doi:10.1007/BF01386390.
Du, C., and Lee, J. S. (1996). Fuzzy classification of earth terrain covers using complex polarimetric SAR data:International Journal of Remote Sensing, 17, 809–826. doi: 10.1080/01431169608949047
Dubreuil-Boisclair, C.D., Gloaguen, E., Marcotte, D., & Giroux, B. (2011). Heterogeneous aquifer characterization from ground-penetrating radar tomography and borehole hydrogeophysical data using nonlinear Bayesian simulations. Geophysics, 76, J13–J25.
Duan, S., Hu, X., Dong, Z., Wang, L., & Mazumder, P. (2015). Memristor-Based Cellular Nonlinear/Neural Network: Design, Analysis, and Applications. IEEE Transactions on Neural Networks and Learning Systems, 26, 1202-1213.
Dutoit, T., & Marques, F. (2010). Applied Signal Processing: A MATLABTM-based proof of concept. Springer Science & Business Media.

Eberle, D. (1993) Geologic mapping based upon multivariate statistical analysis of airborne geophysical data: International Institute for Aerospace Survey and Earth Sciences (ITC) Journal, Special issue, 173–178

Eberle, D. G., Cole, J., Häuserer, M., and Stettler, E. H. (2005). Combined stochastic and deterministic modelling as an innovative approach to jointly interpret multi-method airborne geophysical datasets: Extended abstract, 9th SAGA Biennial Conference and Exhibition, Cape Town.

Esfahani, M. S., & Dougherty, E. R. (2013). Effect of separate sampling on classification accuracy. Bioinformatics, btt662.

Ezzedine, S., Rubin, Y., & Chen, J. (1999). Bayesian method for hydrogeological site characterization using borehole and geophysical survey data: theory and application to the Lawrence Livermore National. Water Resources Research, 35, 2671–2683.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture.

Fails, J. A., & Olsen Jr, D. R. (2003, January). Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 39-45). ACM.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Foulds, L. R., 2012, Graph theory applications. Springer Science & Business Media.

Frénay, B., Doquire, G., & Verleysen, M. (2013). Is mutual information adequate for feature selection in regression? Neural Networks, 48, 1-7.

Friedel, S. (2003). Resolution, stability and efficiency of resistivity tomography estimated from a generalized invers approach. Geophysical Journal International, 153, 305-316.

Galindo, J. (Ed.). (2005). *Fuzzy Databases: Modeling, Design and Implementation: Modeling, Design and Implementation*. IGI Global.

Ganivada, A., Sankar Ray, S., & Pal, S. K. (2013). Fuzzy rough sets, and a granular neural network for unsupervised feature selection. Neural Networks, 48, 91-108.

Gassmann, F. (1951). Über die Elastizität poröser Medien. Vierteljahresschrift der Naturforsch. Ges., Zürich, 96, 1-22.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR), 38(3), 9.

Gloaguen, E., Chaouteau, M., Marcotte, D., & Chaouis, R. (2001). Estimation of hydraulic conductivity of an unconfined aquifer using cokriging of GPR and Hydrostratigraphic data. Journal of Applied Geophysics, 47, 135-152.

Goldberg, A. V., & Harrelson, C. (2005). Computing the shortest path: A search meets graph theory. In Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 156-165). Society for Industrial and Applied Mathematics.

Gonzalez, R. C., & Thomason, M. G. (1978). *Syntactic pattern recognition: An introduction*.

Haibo, Z.(2014). Error Propagation. Treatise on Geochemistry (Second Edition), Elsevier, Oxford, Pages 33-42, ISBN 9780080983004.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Han, J., Kamber, M., and Tung, A. K. H. (2001). Spatial Clustering Methods in Data Mining: A Survey in Geographic Data Mining and Knowledge Discovery.

Harris, J.M., Nolen-Höeksema, R.C., Langan, R.T., Van Schaack, M., Lazaratos, S.K., & Rector, J.W. (1995). High-resolution crosswell imaging of Texas carbonate reservoirs: part 1-Project summary and interpretation. Geophysics, 60, 667–681.

Haykin, S. (2008). Neural Networks and Learning Machines,Third Edition. Pearson Education, Inc., Upper Saddle River, New Jersey.

Hill, T., Lewicki, P., & Lewicki, P. (2006). Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. StatSoft, Inc.

Höppner, F., F. Klawonn, R. Kruse, and T. Runkler. (1999). Fuzzy cluster analysis: Methods for classification, data analysis and image recognition: John Wiley & Sons, Inc.

Hoffman, F. M., Larson, J. W., Mills, R. T., Brooks, B. G. J., Ganguly, A. R., Hargrove, W. W., ... & Vatsavai, R. R. (2011*). Data mining in earth system science (dmess 2011).* Procedia Computer Science, 4, 1450-1455.

Holliger, K., Tronicke, J., Paasche, H., and Dafflon, B. (2008) Quantitative integration of hydrogeophysical and hydrological data: Geostatistical approaches, in C. J. G. Darnault, ed., Overexploitation and contamination of shared groundwater resources: Springer, 67–82.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop?. *Brain Informatics*, *3*(2), 119-131.
Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. Neural Networks, 4, 251-257,
Hubbard, S., Chen, J., Peterson, J., Majer, E., Williams, K., Swift, D., Mailliox, B., & Rubin, Y. (2001). Hydrogeological characterization of the D.O.E. bacterial transport site in Oyster Virginia using geophysical data. Water Resources Research, 37, 2431–2456.
Israel, G. D. (1992. Determining sample size. University of Florida Cooperative Extension Service, Institute of Food and Agriculture Sciences, EDIS.
Jain, A. K., Mao, J., & Mohiuddin, K.M. (1996). Artificial neural networks: A tutorial. IEEE Computer, 29, 31-44.
Jing, X. (2012). Robust adaptive learning of feedforward neural networks via LMI optimizations. Neural Networks, 31, 33-45.
Kadlec, P., Gabrys, B., & Strandt, S. (2009). *Data-driven soft sensors in the process industry*. Computers & Chemical Engineering, 33(4), 795-814.
Kay M., Neural Computation, Vol. 4, No. 3, 1992, pp. 415–447
Kaufmann, L., and P. J. Rousseeuw, 2009, Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, Vol. 344, Inc.
Kennedy, J., & Eberhart, R. C. (1995). Particle swarm optimization. Proceedings of IEEE International Conference on Neural Networks, Piscataway, 1942-1948.
Khoshdel, H., & Riahi, M. A. (2011). Multi attribute transform and neural network in porosity estimation of an offshore oil field — A case study. Journal of Petroleum Science and Engineering, 78, 740–747.
Kiranyaz, S., Ince, T., & Yildirim, A., G. (2009). Evolutionary artificial neural networks by multi-dimensional particle swarm optimization. Neural Networks, 22, 1448-1462.
Knab, M., Appel, E., and Hoffmann, V., 2001, Separation of the anthropogenic portion of heavy metal contents along a highway by means of magnetic susceptibility and fuzzy c-means cluster analysis: European Journal of Environmental and Engineering Geophysics, 6, 125–140
Knödel, K., Krummel, H., & Lange, G. (1997). Handbuch zur Erkundung des Untergrundes von Deponien und Altlasten. Springer.
Kohavi, R., & John, G. H. (1997). *Wrappers for feature subset selection*. Artificial intelligence, 97(1), 273-324.
Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.
Kruiver, P. P., Kok, Y. S., Dekkers, M. J., Langereis, C. G., and Laj, C., 1999, A pseudo-Thellier relative paleointensity record, and rock magnetic and geochemical parameters in relation to climate during the last 276 kyr in the Azores region: Geophysical Journal International, 136, 757–770. doi: 10.1046/j.1365-246x.1999.00777.x
Kumar, V. (2010). *Discovery of patterns in global earth science data using data mining*. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 2-2). Springer Berlin Heidelberg.
Lanne, E. (1986). Statistical multivariate analysis of airborne geophysical data on the SE border of the central Lapland Greenstone Complex: Geophysical Prospecting, 34, 1111–1128. doi: 10.1111/j.1365- 2478.1986.tb00516.x
Leite, E.P., & de Souza Filho, C.R. (2009). Artificial neural networks applied to mineral potential mapping for copper-gold mineralizations in the Carajás Mineral Province, Brazil. Geophysical Prospecting, 57, 1049-1065.
Leite, E. P., & Vidal, A. C. (2011). 3D porosity prediction from seismic inversion and neural networks. Computers & Geosciences, 37, 1174-1180.
Leray, P., & Gallinari, P. (2002). Feature selection with neural networks. Behaviormetrika, 26, 145-166.
Leung, Y. (2010). Knowledge discovery in spatial data. Heidelberg: Springer.
Li, J., & Narayanan, R. M. (2004). *Integrated spectral and spatial information mining in remote sensing imagery*. IEEE Transactions on Geoscience and Remote Sensing, 42(3), 673-685.
Linder, S., Paasche, H., Tronicke, J., Niederleithinger, E., Vienken, T. (2010). Zonal cooperative inversion of crosshole P-wave, S-wave, and georadar traveltime datasets. Journal of Applied Geophysics, 72, 254-262.

Liu, H., & Motoda, H. (1998). Feature transformation and subset selection. IEEE Intelligent System, 13, 26-28.
Liu, H., Motoda, H. (2001). Feature Extraction, Construction and Selection: A Data Mining Perspective. Boston: Kluwer Academic Publishers. 2nd Printing 2001.
Liu, J., Wilson, A., & Gunning, D. (2014). Workflow-based Human-in-the-Loop Data Analytics. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research* (p. 49). ACM.
Longley, P. A., Fischer, M. M., & Getis, A. (1998). Recent Developments in Spatial Analysis.
Lunne, T., Powell, J.J.M., & Robertson, P.K. (1997), Cone Penetration Testing. Taylor and Francis.
MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281-297, University of California Press, Berkeley, Calif.
Marshall, M. N. (1996). Sampling for qualitative research. Family practice, 13(6), 522-526.
Martelet, G., Truffert, C., Tourlière, B., Ledru, P., and Perrin, J. (2006). Classifying airborne radiometry data with agglomerative hierarchical clustering: a tool for geological mapping in context of rainforest (French Guiana): International Journal of Applied Earth Observation and Geoinformation, 8, 208–223. doi: 10.1016/j.jag.2005.09.003
Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE transactions on pattern analysis and machine intelligence, 26(5), 530-549.
Meeker, W. Q., & Escobar, L. A. (2014). *Statistical methods for reliability data.* John Wiley & Sons.
Miche, Y., Sorjamaa, A., Bas, P., Simula, O., Jutten, C., & Lendasse, A. (2010). OP-ELM: Optimally Pruned Extreme Learning Machine. IEEE Transactions on Neural Networks, 21, 158-162.
Mitchell, M. (1998). An Introduction to Genetic Algorithms. MIT Press.
Moorkamp, M., Lelièvre, P. G., Linde, N., & Khan, A. (2016). Integrated Imaging of the Earth: Theory and Applications. Wiley.
Niederleithinger, E., Wiggenhauser, H., & Taffe, A. (2009). The NDT-CE test and validation center in Horstwalde. Proceedings of NDTCE, 9.
Odeh, I. O. A., McBratney, A. B., & Chittleborough, D. J. (1990). Design of optimal sample spacings for mapping soil using fuzzy-k-means and regionalized variable theory. Geoderma, 47(1-2), 93-122.
Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.
Paasche, H. (submitted a). Translating tomographic ambiguity into the probabilistic inference of hydrologic and engineering target parameters. Geophysics,
Paasche, H. (submitted b). Probabilistic inference of spatially continuous geotechnical parameter fields by means of sparse calibration data and geophysical tomography. Geophysical Prospecting.
Paasche, H. (2015). Fully non-linear self-organizing inversion of cross-borehole tomographic data. Near Surface Geoscience - 21st European Meeting of Environmental and Engineering Geophysics, Italy.
Paasche, H., Tronicke, J., Holliger, K., Green, A. G., & Maurer, H. R. (2006). Integration of diverse physical-property models: Subsurface zonation and petrophysical parameter estimation based on fuzzy c-means cluster analyses. Geophysics, 71, H33–H44.
Paasche, H., Günther, T., Tronicke, J., Green, A. G., Maurer, H., and Holliger, K. (2007). Integrating multi-scale geophysical data for the 3D characterization of an alluvial aquifer: Istanbul, Turkey: 13th European Conference on Environmental and Engineering Geophysics (EAGE Near Surface Geophysics), Expanded Abstracts, 5 p.
Paasche, H., & Tronicke, J. (2007). Cooperative inversion of 2D geophysical dataset: A zonal approach based on fuzzy c-means cluster analysis. Geophysics, 72, 35-39.
Paasche, H., & Eberle, D. (2009). Rapid integration of large airborne geophysical data suites using a fuzzy partitioning cluster algorithm: A tool for geological mapping and mineral exploration targeting. Exploration Geophysics, 40(3), 277-287.
Paasche, H., Eberle, D. (2011). Automated compilation of pseudo-lithology maps from geophysical datasets: A comparison of Gustafson-Kessel and fuzzy c-means cluster algorithms. Exploration Geophysics, 42, 275-285.
Paasche, H., Rumpf, M., Hausmann, J., Fechner, T., Werban, U., Tronicke, J., & Dietrich, P. (2013). Advances in acquisition and processing of near-surface seismic tomographic data for geotechnical site assessment. First Break, 31, 59-65.

Paasche, H. (2015). Fully Non-linear Self-organizing Inversion of Cross-borehole Tomographic data Near Surface Geoscience. 21st European Meeting of Environmental and Engineering Geophysics, Italy.
Paasche, H., Eberle, D., Das, S., Cooper, A., Debba, P., Dietrich, P., Dudeni-Thlone, N., Gläßer, C., Kijko, A., Knobloch, A., Lausch, A., Meyer, U., Smit, A., Stettler, E., Werban, U. (2014). Are Earth Sciences lagging behind in data integration methodologies?. Environ.Earth Sci. 71 (4), 1997 – 2003.
Pengra, D. B., Dillman, L.T. (2009). Notes on Data Analysis and Experimental Uncertainty. University of Washington.
Pham, D. L. (2001). Spatial models for fuzzy clustering. Computer vision and image understanding, 84(2), 285-297.
Peschel, G. (1973). Zur quantitativen komplexen Interpretation gravimetrischer und magnetischer Profile: Zeitschrift für Angewandte Geologie, 19, 287–292
Pires, A. C. B., and Harthill, N. (1989). Statistical analysis of airborne gamma-ray data for geologic mapping purposes: Crixas-Itapaci area, Goiás, Brazil: Geophysics, 54, 1326–1332. doi: 10.1190/1.1442592
Pirkle, F. L., Howell, J. A.,Wecksung, G.W., Duran,B. S., and Stablein, N. K. (1984). An example of cluster analysis applied to a large geologic dataset: Aerial radiometric data from Copper Mountain, Wyoming: Mathematical Geology, 16, 479–498. doi: 10.1007/BF01886328
Poulton, M.M. (2002). Neural networks as an intelligence amplification tool: a review of applications. Geophysics, 67, 979-993.
Ramachandran, R., Conover, H. T., Graves, S. J., & Keiser, K. (2000). *Challenges and solutions to mining earth science data*. In AeroSense 2000 (pp. 259-264). International Society for Optics and Photonics.
Riedmiller, M., & Braun, H.1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm," Proceedings of the IEEE International Conference on Neural Networks. pp. 586–591.
Pham, D. L. (2001). Spatial models for fuzzy clustering. Computer Vision and Image Understanding, 84, 285-297.
Poulton, M.M. (2002). Neural networks as an intelligence amplification tool: a review of applications. Geophysics, 67, 979-993.
Quan, H., Srinivasan, D., & Khosravi, A. (2014). Short-term load and wind power forecasting using neural network-based prediction intervals. IEEE Transactions on Neural Networks and Learning Systems, 25, 303-315.
Raeesi, M., Moradzadeh, A., Doulati Ardejani, F., & Rahimi. M. (2012). Classification and identification of hydrocarbon reservoir lithofacies and their heterogeneity using seismic attributes, logs data and artificial neural networks. Journal of Petroleum Science and Engineering, 82, 151–165.
Raman, K. Research Statement: Machine Learning with Humans in the Loop.
Raymer, D. S., Hunt, E. R., & Gardner, J. S. (1980). An improved sonic transit time-to-porosity transform. Proceeding of SPWLA 21st ann. Meeting, paper P.
Razavi, S., & Tolson, B.A. (2011). A new formulation for feedforward neural networks. IEEE Transactions on Neural Networks, 22, 1588-1598.
Recknagel, F. (2001). Applications of machine learning to ecological modelling. Ecological Modelling, 146(1), 303-310.
Rothrock, L., & Narayanan, S. (2011). *Human-in-the-loop Simulations* (pp. 26-29). Springer.
Roy, L., Sen, M.K., McIntosh, K., Stoffa, P.L., & Nakamura, Y. (2005). Joint inversion of first arrival seismic travel-time and gravity data. Journal of Geophysics and Engineering, 2, 277–289.
Rubin, Y., & Hubbard, S.S., 2005, Hydrogeophysics. Springer.
Ruggeri, P., Irving, J., Gloaguen, E., & Holliger, K. (2013). Regional scale integration of multiresolution hydrological and geophysical data using a two-step Bayesian sequential simulation approach. Geophysical Journal International, 194, 289–303.
Rumpf, M., & Tronicke, J. (2014). Predicting 2D geotechnical parameter fields in near-surface sedimentary environments. Journal of Applied Geophysics, 101, 95–107.
Russ, J. C., & Woods, R. P. (1995). *The image processing handbook*. Journal of Computer Assisted Tomography, 19(6), 979-981.

Sarma, A. D., Benjelloun, O., Halevy, A., & Widom, J. (2006). Working models for uncertain data. In 22nd International Conference on Data Engineering (ICDE'06) (pp. 7-7). IEEE.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.
Schön, J. H. (1998). Physical properties of rocks: Fundamentals and principles of petrophysics. Pergamon Press.
Schröter, I., Paasche, H., Dietrich, P., & Wollschläger, U. (2015). Estimation of Catchment-Scale Soil Moisture Patterns Based on Terrain Data and Sparse TDR Measurements Using a Fuzzy C-Means Clustering Approach. Vadose Zone Journal, 14(11).
Schwarzbach, C., Börner, R.U., & Spitzer, K. (2005). Two-dimensional inversion of direct current resistivity data using a parallel, multi-objective geneteic algorithm. Geophysical Journal International, 162, 685-695.
Senthilkumaran, N., & Rajesh, R. (2009). Edge detection techniques for image segmentation–a survey of soft computing approaches. International journal of recent trends in engineering, 1(2).
Sen, Mrinal K., and Stoffa, Paul L. (2013). Global optimization methods in geophysical inversion. Cambridge University Press.
Seteiono, R., & Liu, H. (1997). Neural-network feature selector. IEEE Transaction on Neural Networks, 8, 354-362.
Shi, H., Shen, Y., and Liu, Z. (2003). Hyperspectral bands reduction based on rough sets and fuzzy c-means clustering: Proceedings of the 20th IEEE Instrumentation and Measurement Technology Conference, Vol. 2, 1053–1056.
Siegel, D. A., Buesseler, K. O., Behrenfeld, M. J., Benitez-Nelson, C. R., Boss, E., Brzezinski, M. A., ... & Perry, M. J. (2016). *Prediction of the Export and Fate of Global Ocean Net Primary Production: The EXPORTS Science Plan*. Frontiers in Marine Science, 3, 22.
Silberschatz, A., Korth, H. F., & Sudarshan, S. (1997). Database system concepts (Vol. 4). New York: McGraw-Hill.
Sondhi, P. (2009). *Feature construction methods: a survey.* sifaka. cs. uiuc. edu, 69, 70-71.
Spate, J., Gibert, K., Sànchez-Marrè, M., Frank, E., Comas, J., Athanasiadis, I., & Letcher, R. (2006). Data Mining as a tool for environmental scientists.
Stanton. J. (2012). An introduction to Data Science. Syracuse University.
Takagi, K., and H.S. Lin. (2012). Changing controls of soil moisture spatial organization in the Shale Hills Catchment. Geoderma 173–174:289–302. doi:10.1016/j.geoderma.2011.11.003
Tarantola, A. (1978), Inverse Problem Theory. Elsevier.
Taylor, J. (1997). *Introduction to error analysis, the study of uncertainties in physical measurements* (Vol. 1).
Topp, G. C., Davis, J. L., & Annan, A. P. (1980). Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. Water Resources Research, 16, 574-582.
Tronicke, J., & Holliger, K. (2005). Quantitative integration of hydrogeophysical data: Conditional geostatistical simulation for characterizing heterogeneous alluvial aquifers. Geophysics, 70, H1–H10.
Tronicke, J. (2007). The influence of high frequency uncorrelated noise on first-break arrival times and crosshole traveltime tomography. Journal of Environmental & Engineering Geophysics, 12(2), 173-184.
Tronicke, J., Paasche, H., & Böniger, U. (2012). Crosshole traveltime tomography using particle swarm optimization: a near-surface field example. Geophysics, 77, R19–R32.
Van der Baan, M., & Jutten, C. (2000). Neural networks in geophysical applications. Geophysics, 65, 1032-1047.
Velis, D. R. (2001). Traveltime inversion for 2D anomaly structures. Geophysics, 66, 1481-1487.
Verikas, A., & Bacauskiene, M. (2002). Feature selection with neural networks. Pattern Recognition Letters, 23, 1323–1335.
Urbat, M., Dekkers, M. J., & Krumsiek, K. (2000). Discharge of hydrothermal fluids through sediment at the Escanaba Trough, Gorda Rich (ODP Leg 169); assessing the effects on the rock magnetic signal: Earth and Planetary Science Letters, 176, 481–494. doi: 10.1016/S0012- 821X(00)00024-8.
Western, A.W., R.B. Grayson, G. Blöschl, G.R. Willgoose, & T.A. McMahon. (1999). Observed spatial organization of soil moisture and its relation to terrain indices. Water Resour. Res. 35:797–810. doi:10.1029/1998WR900065

Wharton, R.P., Hazen, G.A., Rau, R.N., & Best, D.L. (1980). Advancements in electromagnetic propagation logging. SPE 9267 American Institute of Mining Metallurgical and Petroleum Engineers.
Widrow, B., Greenblatt, A., Kim, Y., & Park, D. (2013). The No-Prop algorithm: A new learning algorithm for multilayer neural networks. Neural Networks, 37, 182-188.
Wiley, J., Ltd, S. (1982). Handbook of Measurement Sciences, Vol 1. Edited by P.H. Sydenham.
Wilson, D.J., A.W. Western, and R.B. Grayson. (2005). A terrain and databased method for generating the spatial distribution of soil moisture. Adv. Water Resour. 28:43–54. doi:10.1016/j.advwatres.2004.09.007
Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
Wu, X., Rozycki, P., & Wilamowski, B.M. (2015). A hybrid constructive algorithm for single-layer feedforward networks learning. IEEE Transactions on Neural Networks and Learning Systems, 26, 1659-1668.
Wyllie, M. R. J., Gregory, A. R., & Grander, L. W. (1956). Elastic wave velocities in heterogeneous and porous media. Geophysics, 26, 41-70.
Yamamoto, T. (2001). Imaging the permeability structure within the near-surface sediments by acoustic crosswell tomography. Journal of Applied Geophysics, 47, 1-11.
Yan, H., & Yang, J. (2015). Locality preserving score for joint feature weights learning. Neural Networks. 69, 126-134.
Zacharias, S., Bogena, H., Samaniego, L., Mauder, M., Fuß, R., Pütz, T., Frenzel, M., Schwank, M., Baessler, C., Butterbach-Bahl, K., Bens, O., Borg, E., Brauer, A., Dietrich, P., Hajnsek, I., Helle, G., Kiese, R., Kunstmann, H., Klotz, S., Munch, J.C., Papen, H., Priesack, E., Schmid, H.P., Steinbrecher, R., Rosenbaum, U., Teutsch, G., & Vereecken, H. (2011). A network of terrestrial environmental observatories in Germany. Vadose Zone J. 10:955–973. doi:10.2136/vzj2010.0139
Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5-6), 375-381.
Zhang, W., Lin, X., Pei, J., & Zhang, Y. (2008). Managing uncertain data: Probabilistic approaches. E*valuation*, *1*, 9.

Appendix A
Towards Probabilistic Prediction of Soil Moisture in the Schäfertal Catchment

A.1 Introduction

Probabilistic prediction of soil moisture patterns and their temporal dynamics is an important issue to infer hydrological flux and flow pathways to improve the description and prediction ability of hydrological, ecological, and pedological models. Measurement campaigns offer uncertain information about the target parameter according to limited number of observations or measurement errors. Traditionally, soil moisture prediction models do not offer a probabilistic prediction based on these errors. Quantitative and realistic prediction of uncertainty in the exploration target parameter estimation would offer valuable information for decision taking in hydrological, ecological, and pedological tasks with regard to risk quantification and minimization. In this work based on Artificial Neural Network algorithms I illustrate a data-driven recent attempt towards probabilistic prediction of soil moisture.

A.2. Methodology

One of the most powerful algorithms for prediction are Artificial Neural Networks (ANN) (see chapters 2 and 3). ANNs are composed of an input and an output layer interconnected by a hidden layer of "neurons" which are capable of learning complex interrelations between input data and target prediction parameters. For creating my prediction model I use the same strategy for creating and training the ANNs as in chapter 2.

A.3 Processing

For probabilistic prediction of soil moisture in the Schäfertal catchment I have used the same dataset used by Schröter et al., (2015) (see also chapter 4). They separated the catchment in two distinct parts shown in Figure A-1 as farm land used for agriculture and growing crops as well as grass land.

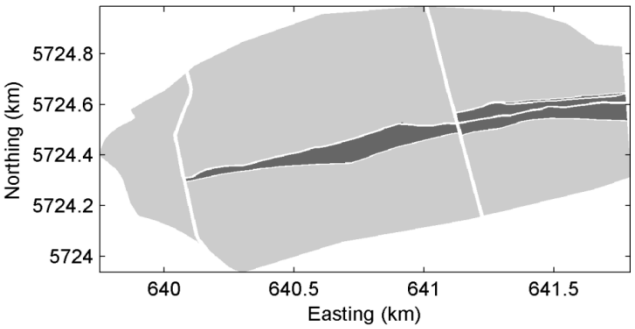

Figure A-1: Land use map created by Schröter et al., (2015). Arable land is depicted by light gray, grassland by dark gray colors.

Schröter et al., (2015), used only topographic information (Figure A-2) which is related to hydrological processes controlling the spatial distribution of soil moisture, particularly at generally wet states. Figure A-2 shows topographic attribute maps prepared and used by Schröter et al., (2015) of the Schäfertal catchment obtained from a high-resolution 2 x 2 m^2 digital elevation model measured by air-borne laser scanning.

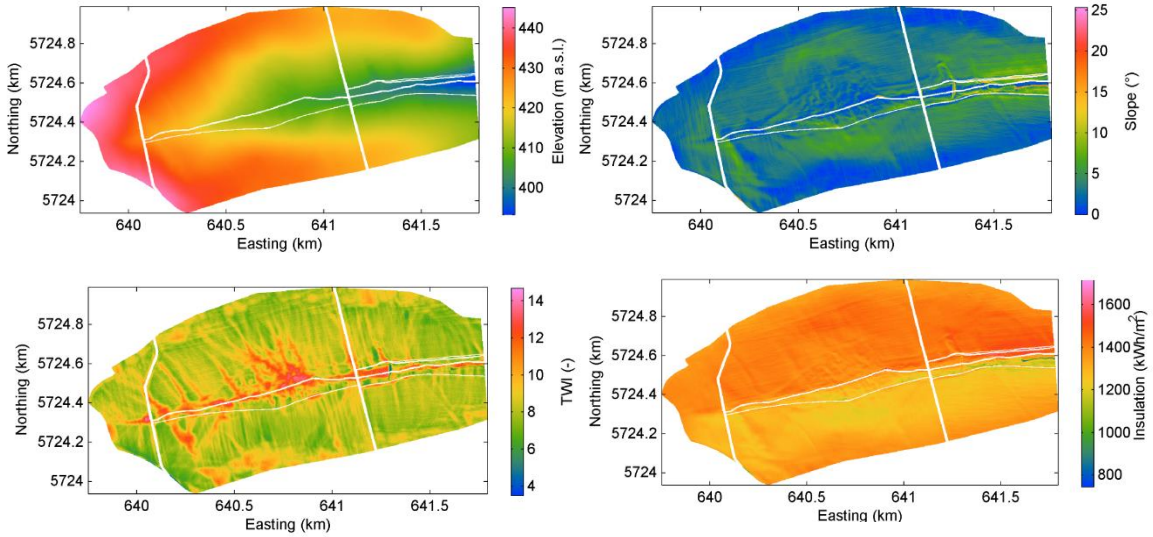

Figure A-2: Four topographic attributes as objective or technical maps; (a) elevation, (b) slope, (c) SAGA wetness index, and (d) annual potential incoming solar radiation derived form a 2-m digital elevation model for the Schäfertal catchment (Schröter et al., 2015).

Figure A-2 presents elevation, slope, SAGA wetness index (SWI), and total annual incoming solar radiation (TIR), respectively. Each map independently contributes information about contextual and local landscape conditions commonly used in the literature (Western et al., 1999; Wilson et al., 2005; Takagi and Lin, 2012, Schröter et al., 2015).

Schröter et al., (2015) used this information in a fuzzy c-means clustering for segmentation of the area to support efficient sampling for soil moisture measurements and to predict the soil moisture distribution in the area based on the clustering results. The location of the soil moisture sample points is shown in Figure A-3. At each location three soil moisture samples had been measured by Schröter et al., (2015). In this work I use 71 samples in the crop and 23 samples in the grass area. At each point 3 samples for soil moisture are considered. Based on this dataset I calculate prediction models for 1000 random selections of samples resulting in probabilistic prediction of soil moisture in the crop and grass area, independently. I follow here the strategy outlined in chapter 2 using ANNs to realize the predictions. Prediction uncertainty results here solely from uncertainty in soil moisture measurements.

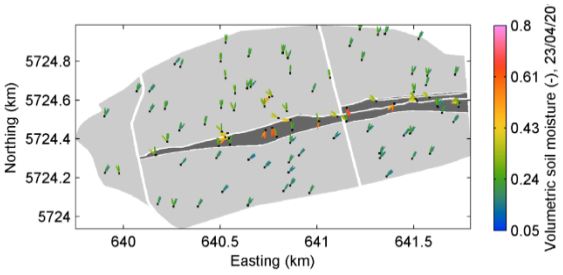

Figure A-3: Soil moisture measurements in the Schäfertal. Locations for sampling the target parameter volumetric soil moisture are indicated by black dots.

Figure A-4 presents an exemplary result of regression in the training phase (Figures 4-4a and 4-4b) and the test phase (Figures 4-4c and 4-4d) in crop (Figures 4-4a and 4-4c) and grass (Figures 4-4b and 4-4d) area, respectively, for one exemplary model of the 1000 trained models. The regression coefficient R indicates good training results for this example.

Figure A-5 presents the results of probabilistic soil moisture prediction in the Schäfertal catchment using ANN models based on the grass and crop land use. Here I

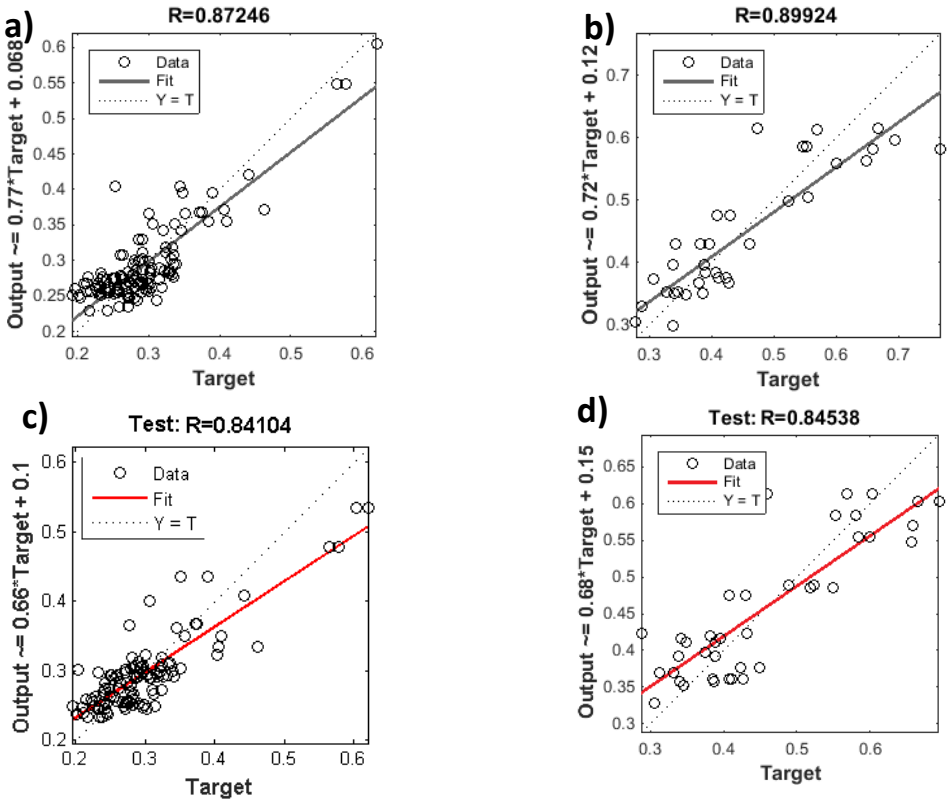

Figure A-4: An exemplary result of regression in the training phase (a and b) and the test phase (c and d) in crop (a and c) and grass (b and d) area.

have trained 1000 ANNs based on different combinations of soil moisture samples by selecting one out of the three soil moisture measurements per sample location. The probabilistic results are shown using glyphs representing the posterior probability density function scaled on a 50*50 m grid. Each glyph depicts predicted soil moisture (color) and relative frequency (length) resultant from 1000 ANN models. Figure A-5b shows the most likely soil moisture extracted from Figure A-5a.

A.4 Conclusion

In this work I tried to offer a probabilistic prediction of soil moisture in a small-scale catchment. For doing so I used elevation, slope, TWI, and insulation as input data to estimate their relation to the measured soil moisture values. Based on the land use I separated the area in crop and grass land. I created 1000 trained ANNs with random

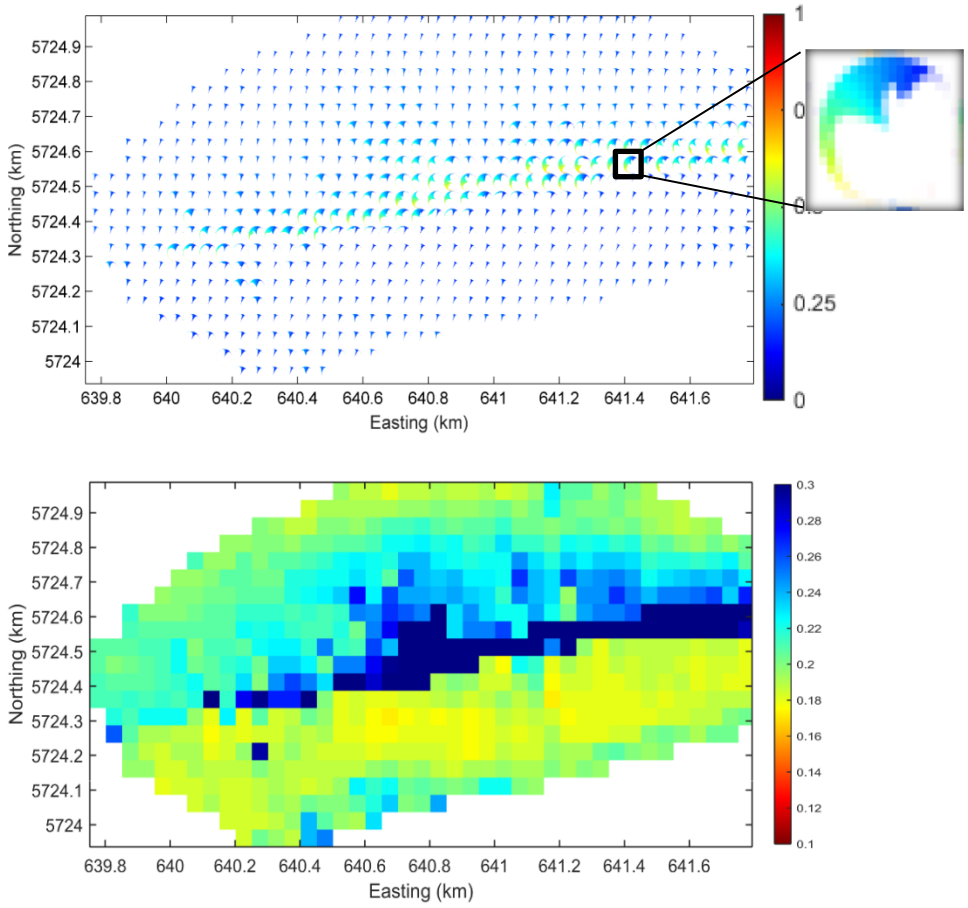

Figure A-5: (a) Probabilistic map of volumetric soil moisture content using ANN models based on the grass and crop land use. The scale for each point is 50*50 m. Each glyph depicts soil moisture (color) and relative frequency (length). One exemplary glyph is zoomed in Figure A-5a to present the results of the 1000 ANN models for the related position. (b) shows the most likely soil moisture extracted from (a).

selection of target soil moisture data. Accordingly, with this method I was able to generate probabilistic information and could quantify a prediction interval at each location. I showed the results of this probabilistic prediction method for the Schäfertal catchment. I believe that more research is necessary to prove the target parameter prediction in the Schäfertal catchment based on the introduced methods in this thesis, since no independent validation set of soil moisture data remains. Such method can be developed in any field catchment for probabilistic prediction of target parameters also if non-topographic attribute maps shall be included. If uncertainty information for each pixel in the considered maps would be available, it could be considered when training the ANNs to avoid overfitting (see chapter 3).

Appendix B

A New Methodology for Prediction of 2D Distributions of Sparsely Measured Logging Data under Full Consideration of Tomographic Model Generation Ambiguity

Abduljabbar Asadi, Peter Dietrich, Hendrik Paasche
Extended Abstract of the Near Surface Geoscience,
Turin, Italy, 6-10 September 2015

B.1 Abstract

We present a novel methodology to probabilistically predict spatial distributions of sparsely measured borehole logging data constrained by multiple geophysical crosshole tomograms. In doing so, we fully account for the ambiguity of the tomographic model reconstruction procedure by taking advantage of a recently developed fully non-linear inversion approach. We use Artificial Neural Networks to link the results of the non-linear inversion with sparse information of tip resistance logging data. Additionally, we achieve information during the training phase of the ANN about the compliancy of tomographic models found by the inversion with the available logging data, which may help to identify those tomographic models that may reconstruct the subsurface more realistically.

B.2 Introduction

Geophysical tomographic datasets have proven valuable in supporting many near-surface hydrological, and engineering exploration tasks. They uniquely offer the ability to image physical parameter variations, e.g., radar or seismic wave velocities, in a spatially continuous manner. However, for many geophysical tomographic imaging problems, the geophysical model generation suffers ambiguity due to limited number of observations and limited observational accuracy. Traditionally, deterministic approaches are employed

to generate geophysical tomographic models, which do not allow for realistic and quantitative ambiguity appraisal of the model generation ambiguity inherent to a tomographic dataset. For answering geotechnical or hydrological issues, physical parameter variations imaged in geophysical tomograms have to be converted or linked to other target parameters of higher relevance for engineers or hydrologists, such as porosity, tip resistance or sleeve friction. These additional target parameters can usually only be recorded sparsely or along one dimension.

Numerous approaches are available to link geophysical tomograms and sparsely measured hydrological or engineering target parameters. Particularly popular are deterministic transfer functions linking one physical parameter imaged in a tomogram with a sparsely measured target parameter, e.g., recorded in a borehole. An example for such approaches is the equation suggested by Raymer et al. (1980) linking P-wave velocity and porosity explicitly.

Here, we follow a different approach striving to link multiple geophysical tomograms to the same sparsely measured target parameter in order to achieve spatially continuous predictions of the 2D distribution of the target parameter. In doing so, we employ an Artificial Neural Network (e.g., Alpaydin, 2014) to link the tomograms and the target parameter. Thus, we avoid the utilization of an explicitly formulated deterministic transfer function. Additionally, we assess the tomographic ambiguity inherent to the considered tomographic datasets by fully non-linear self-organizing inversion (SOI; Paasche 2015) and propagate them into the prediction of several thousand scenarios of 2D distributions of the sparsely measured target parameter.

B.3 Artificial Neural Networks (ANN)

Nowadays prediction is one of the most important tasks in machine learning that has great advantages and applications in different scientific fields like Geosciences, Computer Science, Bioinformatics, Marketing and so on. Among the most powerful algorithms for the creation of prediction models are Artificial Neural Networks (ANN). The functionality of ANN is similar to the behavior of networks of neurons in the human brain. ANNs are created from different layers of interconnected nodes; each node producing a non-linear

function according to its input values. The input values of a node may come from the results of other nodes or directly from the input dataset. Also there are some nodes that prepare the output of the ANN. The complete network represents a complex system, which can incorporate any degree of nonlinearity, that allows general functions, such as linear or exponential, to be modelled according to the training dataset as a predictor. The general structure of neural networks is shown in Figure B-1.

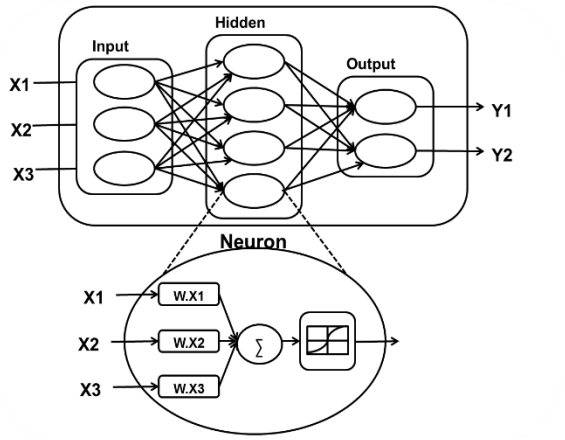

Figure B-1: Structure of an Artificial Neural Network. We have three layers. The input layer prepares data for feeding the ANN. The operation of the hidden layer is determined by inputs and weights of inputs (W). The operation of the output layer is guided by the hidden layer and connected to the results of ANN training.

B.4 Methodology

Fundamental assumption of our work is that a dataset can be considered as an image of reality, i.e., a certain 3D volume or 2D plane of the ground. When imaging reality into a dataset, informational loss occurs due to experimental limitations, e.g., limited resolution capabilities and limited number of observations. Nevertheless, any dataset recorded over the same subsurface area must comprise compliant information about the same reality, even when being expressed in a different way, for example in the form of tip resistance logging data or radar wave traveltime information emanating from a tomographic experiment. Furthermore, a unique noise component may affect every dataset. In our methodology we strive to take advantage of the fact that tomographic datasets and sparse logging data are images of the same reality, albeit with different and unknown imaging functions.

Figure B-2 outlines the processing steps when working towards the probabilistic prediction of spatially continuous distributions of sparsely measured tip resistance data constrained by radar and seismic tomograms. In our example, radar, S-wave and P-wave crosshole tomographic datasets comprise complimentary information about the ground and allow for spatially continuous 2D imaging of radar, S-wave and P-wave velocity variations. We use fully non-linear SOI to achieve ensembles of 30 equivalent radar, S-wave and P-wave tomographic models illustrating the ambiguity of the tomographic model reconstruction for each underlying dataset. These ensembles allow for 27 000 possible combinations of radar, S-wave and P-wave tomograms (30^3).

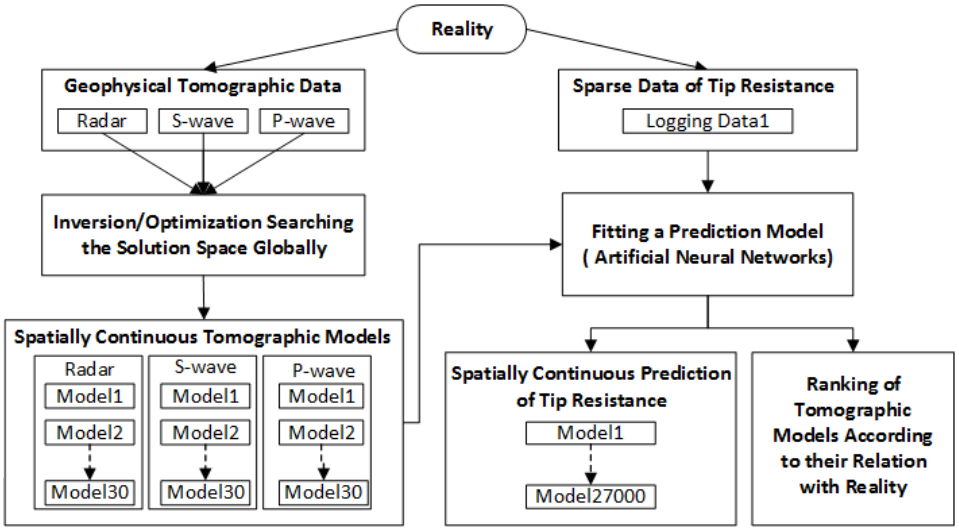

Figure B-2: Flowchart of processing steps for the prediction of 2D tip resistance models and tomographic model ranking

One combination of radar, S-wave and P-wave velocity information provided by the tomograms at locations where 1D logging data about tip resistance are available forms the input layer of an ANN. The corresponding tip resistance information forms the output layer. Then, the network is iteratively trained to learn the relationship between input and output layer information. Here, we strive to train the neurons in the hidden layer in an optimal manner to suite a linear relationship between input and output layer information. Practically, this training procedure strives to learn an optimal transformation of input and output layer information described by the information in the hidden layer. We repeatedly train the ANN for 27 000 different input layer information according to all possible combinations of geophysical tomograms.

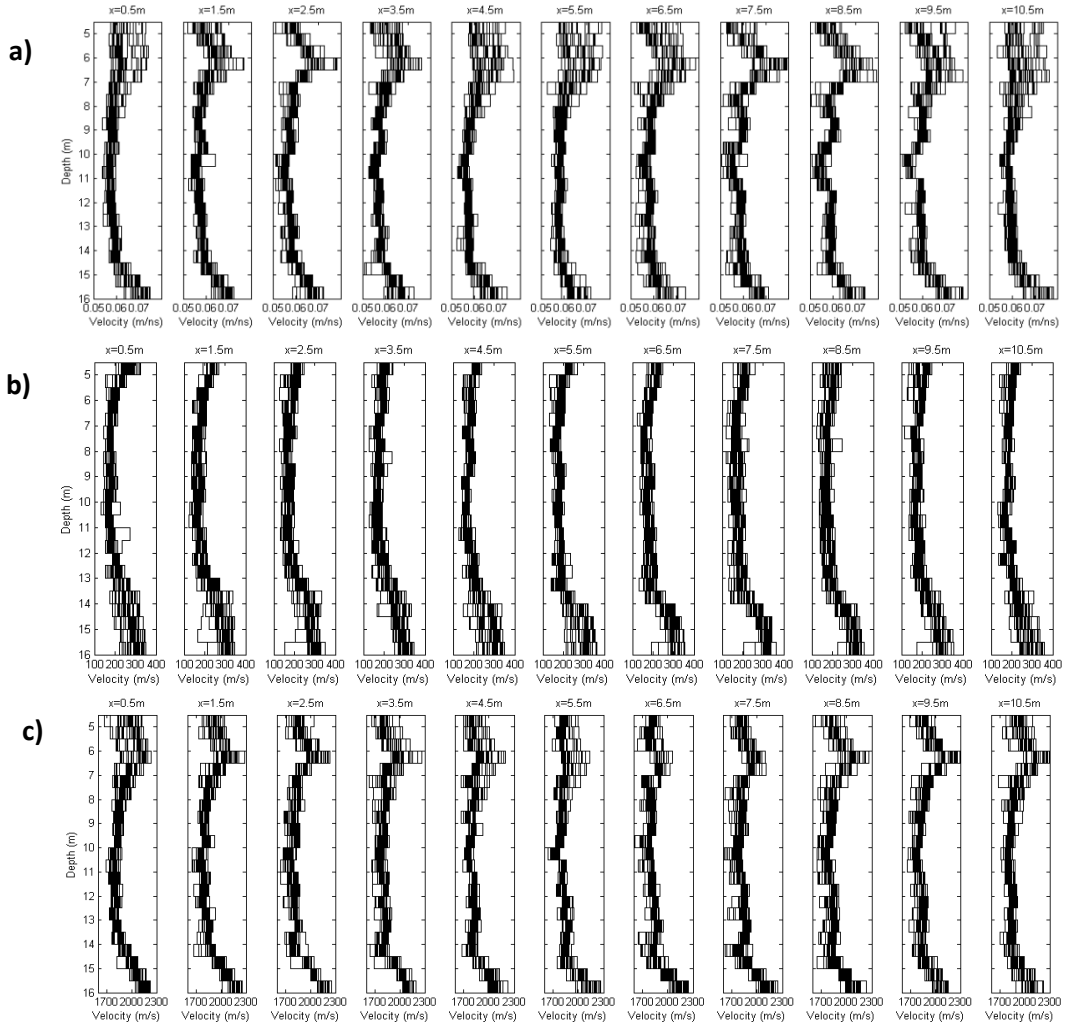

Figure B-3: Geophysical velocity tomograms achieved by fully non-linear SOI. Rectangular grid cells of 1 m lateral and 0.5 m vertical side lengths have been used for model parameterization. The black lines illustrate (a) 30 radar, (b) 30 S-wave, (c) 30 P-wave velocity models.

Next, we apply the 27 000 learned prediction models to the remaining parts of the tomograms to achieve predictions for 2D distributions of tip resistance. Finally, we achieve 27 000 2D scenarios of tip resistance constrained by all available geophysical tomograms and the sparse calibration data.

Additionally, we measure the mean square error (MSE) between predicted and measured tip resistance data, which allows us to judge the quality of the learned prediction model. High MSE values indicate a generally lower compliancy between tomographic models and sparse logging data. This information may be helpful to identify

tomographic models that are rather mathematical solutions with model features that cannot easily be linked to additional data describing the same reality.

B.5 Application to a Field Dataset

We apply our methodology to a field dataset previously measured by Linder et al. (2010). The recorded crosshole tomographic radar, S-wave and P-wave datasets have been re-inverted using the SOI to achieve ensembles of equivalent radar, S-wave and P-wave tomograms reaching from 4.5 m to 16 m depth and covering a 2D plane with 11 m lateral extension (Figure B-3). At two different lateral locations, 1D tip resistance data have been recorded (Figure B-4). The information from the tomograms and logging data is delivered to the ANN. When fitting a prediction model we set up our ANN with 50 neurons in the hidden layer.

Figure B-4 shows the spatially continuous prediction of tip resistance for all possible combinations of tomographic models. At the locations of the logging data, the training has been performed, which results in highly accurate predictions at these two locations. In all other areas, the variability of the tomographic models results in increased prediction uncertainty.

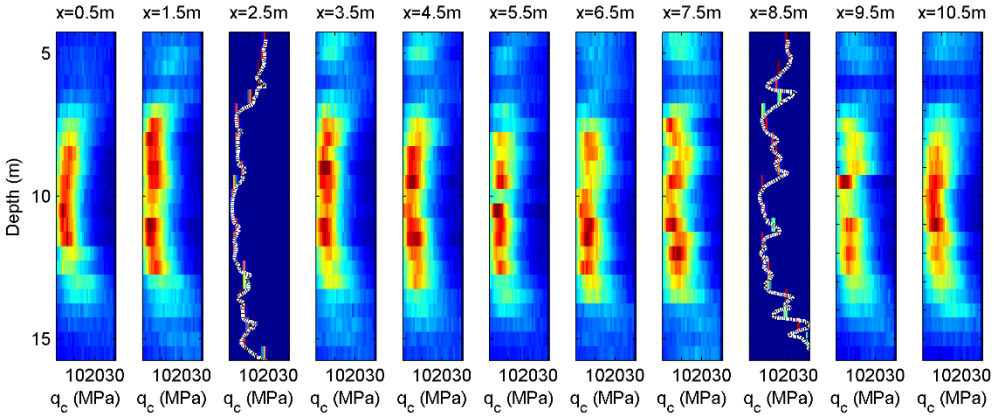

Figure B-4: Results of our 2D tip resistance prediction shown as histographic plot. The dotted white lines show the measured logging data of tip resistance that are used for training the ANN. Red colors correspond to high relative frequencies. Blue colors correspond to low relative frequencies. Note the reduced sharpness of prediction at depths where the logging data are different and cannot be brought in full compliance with the velocity variations in the tomograms.

In Figure B-5, we show the corresponding MSE information for every radar, S-wave and P-wave velocity model. These 90 models fit their underlying datasets equally, and are thus purely equivalent solutions of the geophysical inverse problem. However, differences exist how well they can be linked by the ANN to the tip resistance data. Following our assumption that tomograms and logging data carry compliant information of the same reality, tomographic models with higher MSE can less easily be connected with the available logging data indicating that these models are rather mathematical solutions to the inversion which may not be close to reality in their physical parameter variations.

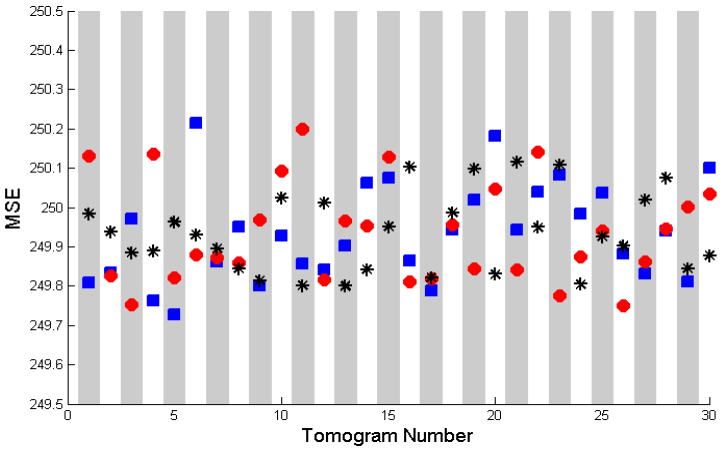

Figure B-5: MSE from ANN training for 30 models of radar, S-wave and P-wave velocity. Blue color for radar models, black color for S-wave and red color for P-wave seismic models. Models with low MSE can be brought more easily in compliance with the tip resistance logging data.

B.6 Conclusions

We suggest a new workflow allowing to link multiple tomographic geophysical models with sparse information about engineering or hydrological target parameters, e.g., logging data. By taking advantage of a fully non-linear inversion procedure we are able to predict many scenarios of 2D models of only sparsely measured logging data. The training procedure of the employed Artificial Neural Network allows for ranking the available equivalent geophysical tomograms achieved by fully non-linear inversion according to their closeness to reality as defined by complimentary logging data.

Appendix C

2D probabilistic prediction of sparsely measured geotechnical parameters constrained by tomographic ambiguity and measurements errors

Abduljabbar Asadi, Peter Dietrich, Hendrik Paasche
Extended Abstract of the 78th EAGE Conference & Exhibition 2016, Vienna, Austria, May 30 - June 2, 2016.

C.1 Abstract

We present a new approach for 2 D probabilistic prediction of sparsely measured target parameters, e.g., measured by direct push technology or borehole logging. Geophysical tomography is used to constrain the prediction. The presented approach fully accounts for tomographic ambiguity and transduces it into prediction uncertainty. Furthermore, errors of the logging data can be considered to avoid overfitting when learning the optimal link between tomograms and logging data by means of Artificial Neural Networks. Consideration of errors results in improved predictions, which we exemplary illustrate here by 2D sleeve friction prediction.

C.2 Introduction

For solving many near-surface engineering exploration tasks, geophysical tomographic datasets are increasingly used to support more traditional geotechnical exploration techniques, such as borehole and direct push logging. Geophysical tomographic datasets are unique in their potential to image ground variability in a spatially continuous manner. The imaged parameters are physical ground properties, for example propagation velocities of seismic waves. Geophysical tomographic imaging suffers ambiguity due to measurement errors and limited number of observations. Traditionally,

deterministic tomographic reconstruction approaches relying on regularized local-search optimization are employed to generate geophysical tomographic models. Such approaches do not allow for realistic and quantitative ambiguity appraisal inherent to the model generation.

Recently, fully non-linear inversion techniques gain increasing popularity. They search the solution space of the tomographic reconstruction problem globally and result in ensembles of tomographic models equally fitting the underlying dataset. Thus, the tomographic ambiguity is represented by a number of equally plausible models.

When employing geophysical tomograms for spatially continuous geotechnical site characterization physical parameter variations imaged by geophysical tomograms have to be converted or linked by prediction models to other geotechnical target parameters of higher relevance for engineers, such as tip resistance or sleeve friction emanating from cone penetration tests. Such target parameters are measured at a few locations and offer detailed 1D logging information. Like all experimental datasets, also logging data carry some measurement errors. When striving to integrate geophysical tomograms and sparse geotechnical logging data for inferring a spatially continuous model of the geotechnical target parameter tomography ambiguity, logging data errors as well as spatially variable and usually unknown inter-parameter relations between the datasets to be integrated must be taken into account.

Traditionally, deterministic transfer functions have been used to link geophysical parameters, e.g. P- and S-wave velocities, with geotechnical target parameters, e.g., sleeve friction. Such approaches have severe limitations handling unknown and non-unique inter-parameter relations. Recently, statistical or geo-statistical frameworks have been proposed which allow for improved incorporation of uncertainty and non-unique inter-parameter relations. Usually, they are used to link deterministically derived geophysical tomograms with the target parameter and thus do not incorporate tomographic reconstruction ambiguity. Only recently, statistical approaches have been developed and applied to link ensembles of equivalently plausible tomographic models

with target parameters measured in boreholes (e.g. Rumpf and Tronicke, 2014; Asadi et al., 2015).

Based on Asadi et al., (2015), we strive to link multiple ensembles of equivalent seismic P-wave, S-wave and radar velocity tomograms with sleeve friction logs. In doing so, we transduce the tomographic ambiguity described by 30 equivalent P-wave, S-wave and radar velocity models into probabilistic statements about 2D sleeve friction distribution. We extent the approach of Asadi et al. (2015) in a way that also estimated logging errors can be considered in the 2D probabilistic sleeve friction prediction. We employ an Artificial Neural Network (e.g., Asadi et al. 2015) to link the tomograms and the target parameter. Thus, we ensure sufficient flexibility to cope with unknown and even non-unique relations between sleeve friction, P-wave, S-wave and radar velocity.

C.3 Artificial Neural Networks (ANN)

Feed-forward artificial neural networks are well-studied and widely applied machine learning algorithms for earth science applications to prediction nonlinear functions between inputs and target parameters of the ANN. This machine learning method does not require the assumption of a prior solution structure or data linkage model. ANNs consist of different layers, referred to as input, hidden and output layer. In the hidden layer interconnected elements known as neurons are present acting as linking elements between input and output information. During training, test, and validation phase ANNs try to find the best fit between input and target parameters. The structure of feed-forward neural networks is shown in Figure C-1. During the learning phase the ANN tries to

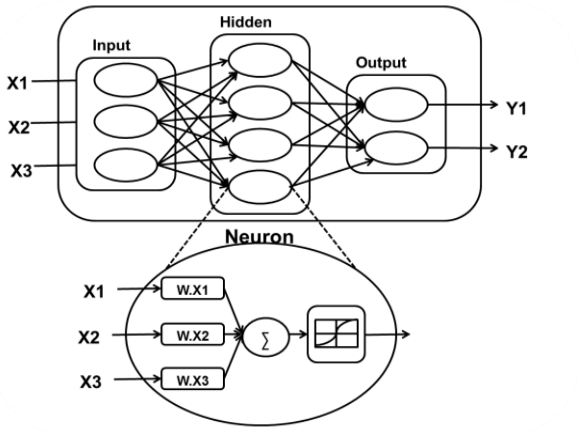

Figure C-1: Structure of a three-layer Artificial Neural Network. The input layer prepares data for feeding the ANN. The operation of the hidden layer is determined by inputs and weights of inputs (W). The operation of the output layer is guided by the hidden layer and connected to the results of ANN training.

minimize the mean squared error (MSE) or weighted mean squared error (WMSE), i.e., the error between the predicted output of the ANN and the measured target values. MSE and WMSE are parameters evaluating the accuracy of the trained ANN. If $\{(x_1,t_1),(x_2,t_2),\ldots,(x_N,t_N)\}$ be a set of training tuples, where $x_i \subset X$ be a vector of input attributes, and $t_i \subset T$ a vector of target attributes, the MSE is defined as $MSE=\frac{1}{2}\sum_{i=1}^{N}(Y_i - t_i)^2$. *N* is the number of observations in the training dataset. If measurement errors shall be considered during the training, the MSE must be replaced by the WMSE. The WMSE is defined as $WMSE=\frac{1}{2}\sum_{i=1}^{N} w_i(o_i - t_i)^2$, *w* determines the weight of the related results of ANN, which can be the accumulated relative errors of the logging data and the tomographic reconstruction ambiguity.

C.4 The Processing Flow

Basic assumption of this study is that logging data and tomographic datasets image the same reality and are therefore compliant. Figure C-2 shows the processing flow followed to create a probabilistic prediction of 2D sleeve friction constrained by ensembles of equivalent radar, P-wave, and S-wave tomograms. Prediction is repeatedly done using the MSE and the WMSE as performance measure when training the ANN. When using the MSE, tomographic ambiguity and logging errors are ignored during the training.

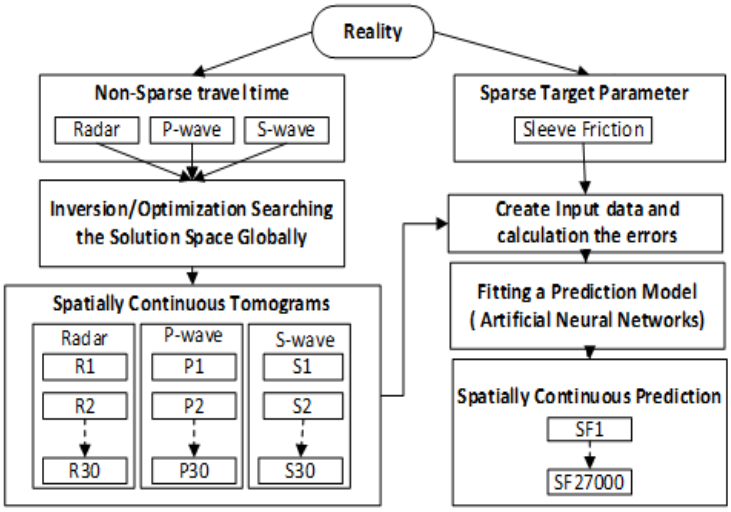

Figure C-2: Processing steps for the prediction of 2D sleeve friction models constrained by ill-posed geophysical tomography.

Instead, the ANN considers the provided training information as true and strives to link tomograms and logging data as good as possible. This results in prediction models that can be excellent for the given training data, but poor at different locations in the model area. When using the WMSE training residuals are error normalized. This reduces the importance of tomographically determined velocity information suffering from poor determination accuracy and particularly noisy sleeve friction readings during the training procedure. The learned linkage model may be less accurate but not over-fit the data beyond their uncertainty limit.

In our example, radar, S-wave and P-wave crosshole tomographic datasets comprise complimentary information about the ground and allow for spatially continuous 2D imaging of radar, S-wave and P-wave velocity variations. We use fully non-linear self-organizing inversion (Paasche, 2015) to achieve ensembles of 30 equivalent radar, S-wave and P-wave tomographic models illustrating the ambiguity of the tomographic model reconstruction for each underlying dataset. This results in 27000 possible combinations of radar, S-wave and P-wave tomograms (30^3) for training. Finally, we apply the 27000 trained prediction models to the model areas, where no sleeve friction information has been measured to achieve 2D probabilistic sleeve friction information.

C.5 Results

We use the database recorded by Linder et al. (2010) comprising two cone penetration tests within the 2D tomographic plane of cross-borehole P-wave, S-wave and radar traveltime datasets. Figure C-3 shows 30 equivalent radar, P-wave, and S-wave tomograms achieved by self-organizing inversion (Paasche, 2015) of the tomographic datasets. The solutions diverge increasingly towards the upper and lower model edge, whereas the velocities in the central parts of the tomograms are well defined. The velocity variability in the tomograms illustrates the ambiguity of the tomographic reconstruction problems.

For each combination of tomograms a relative error weight $w_i=1/(ER_i+ES_i+EP_i+ESF_{i,})$ is calculated for every tomographic grid cell *i* used for training. ER_i, ES_i, and EP_i are the relative errors of Radar, S-wave and P-wave velocity expressed

as ratio of the actual velocity of the chosen model at the *i*-th cell over the velocity range of all 30 models at the *i*-th cell. ESF_i is the relative error of measured sleeve friction integrated over the *i*-th tomographic grid cell. W_i is considered as error weight in the WMSE when training the ANN.

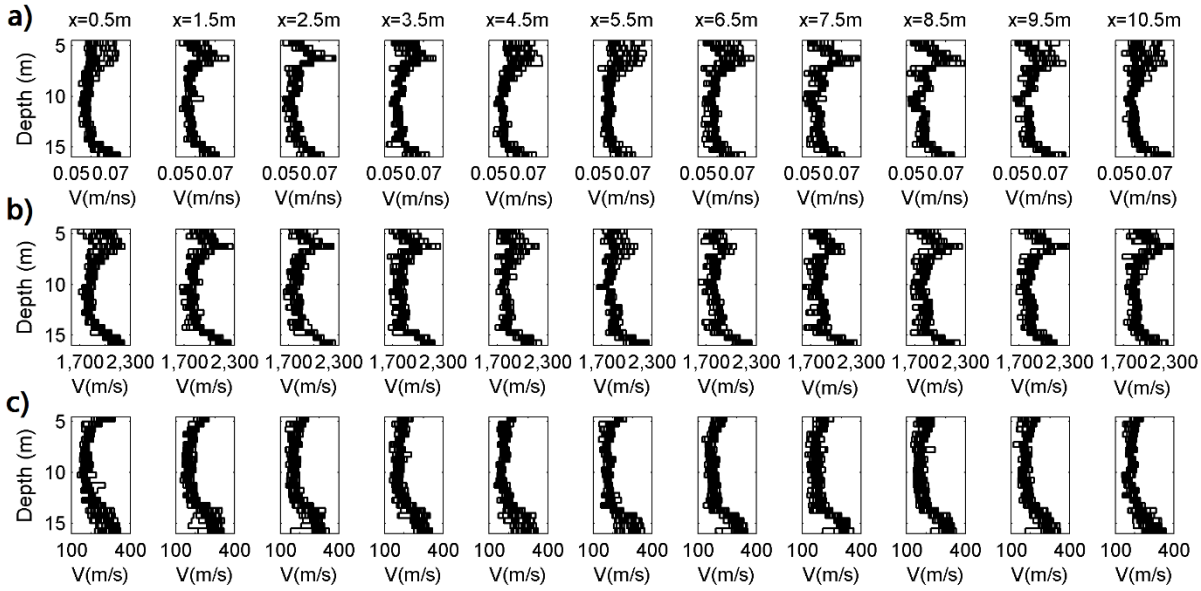

Figure C-3: 2D geophysical velocity tomograms illustrated as laterally neighboured 1D velocity panels. The black lines illustrate 30 equivalent (a) radar, (b) P-wave, (c) S-wave velocity models. Tomographic grid cells have 0.5 m and 1m vertical and lateral side lengths, respectively.

Figure C-4 shows the probabilistic prediction results of 2D sleeve friction based on the ANNs trained with all combination of radar, P-wave, and S-wave tomograms. The result achieved when using the MSE performance measure for training is shown in Figure C-4a. At the location of the logging data the training has been performed and the learned prediction model offers highly accurate prediction at these positions. In the remaining part of the 2D area the ranges of predicted sleeve friction are large, due to simple error propagation when applying the over-fitted prediction model to this model regions.

When trying to account for measurement errors in the training step, the predicted range is clearly reduced. Furthermore, the regions of high relative frequencies are now more sharply contoured. At the position of the boreholes the prediction accuracy is decreased

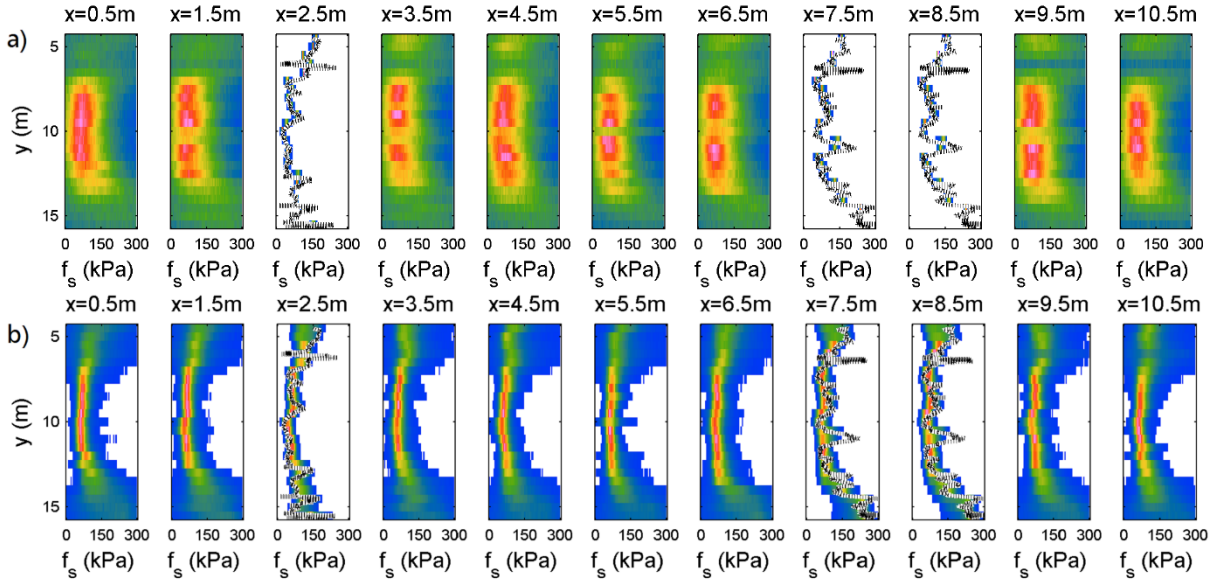

Figure C-4: Results of our 2D probabilistic prediction of sleeve friction. The dotted black lines show the measured logging data of sleeve friction that are used for training the ANN and calculation of the logging error. Red color corresponds to high relative frequencies. Blue color corresponds to low relative frequencies. Note that the ANNs trained with (a) with the MSE performance measure offer increased ranges of sleeve friction compared to the results (b) achieved when using the WMSE.

but at the rest of the area this type of training strategy can offer better constrained predictions.

C.6 Conclusions

In this work we offer a strategy for linking multiple geophysical tomograms to sparse geotechnical information, e.g., logging data. We fully account for tomographic ambiguity expressed by ensembles of equivalent tomograms and transduce this ambiguity into prediction uncertainty. Additionally, logging errors can be incorporated to avoid an overfitting of the ANN used to learn the linkage model relation tomographic and logging information. The results in this work are exemplary illustrated based on the sleeve friction as target parameter. However, this method can be applied to any combination of tomograms and logging data to achieve 2D or 3D probabilistic predictions of the target parameter variability.

Appendix D

Predicting Porosity According to Ensembles of Collocated Radar and Seismic Tomographic Models with Artificial Neural Networks

A. Asadi, P. Dietrich, H. Paasche
Abstract of the 75th annual conference of the DGG, 23. - 26. March 2015, Hannover

D.1 Abstract

Predicting the porosity of the ground according to radar and seismic tomographic information is an important task in geophysics but there are some problems in traditional ways for doing this prediction. Frequently, people select deterministic models of radar and seismic velocities and use them for prediction of porosity distributions using deterministic petrophysical transfer functions. This traditional method cannot offer realistic confidence and prediction intervals. A confidence interval determines a range of values that are considered as acceptable for prediction result with a specified probability that the value of a parameter lies within it. A prediction interval is an estimate according to the observed information offering a range in which future observations will fall with a certain probability. In our work we use Artificial Neural Networks (ANN) that are one of the most powerful tools for prediction of target parameters. ANNs are able to learn from multi-layer input information, such as ensembles of equivalent tomographic geophysical models achieved by fully non-linear inversion, during the train step and can reveal hidden and strongly non-linear dependencies, even when there is a significant noise in the training set. Calibration with sparse information about the target parameter, e.g., porosity information achieved by borehole logging, is required for training ANN. Results show that our approach is

powerful for predicting the porosity distribution of the ground while providing quantitative information about confidence and prediction intervals.

Appendix E

Conceptual Developments for Clustering Mapped Data Emanating from Technical Sensors and Subjective Insights of Human Experts

A. Asadi, P. Dietrich, H. Paasche
Abstract of the 75th annual conference of the DGG, 23. - 26. March 2015, Hannover

E.1 Abstract

When exploring the ground, geophysicists and other Earth scientists frequently map the available observations in individual but collocated thematic images, e.g. geological, hydrological, magnetic, electrical conductivity or radiometric maps. As humans we are trying to analysis these images according to three features: Color or absolute value of every point, edge information of structures and textures in the maps. Integrating and segmenting multiple thematic images, e.g., by cluster analysis, such as fuzzy c-means, k-means or expectation maximization (EM) is one of the important tasks in geoscientific map analysis. Traditional algorithms have some problems in this subject. For example, partitioning or model-based cluster analyses, e.g., fuzzy c-means or EM, analyze just the color or absolute values and they do not consider other features like edges and texture in the maps. Furthermore, there is no way for intelligent combination and utilization of (partly) subjective and technical information in such cluster analyses, e.g. pre-classified geological maps or geophysical maps, respectively. We are going to discuss conceptual ideas inspired by data mining for integrating and clustering maps, while paying attention to their subjective and technical acquisition procedure. The new concepts may potentially allow for multi-map integration and cluster analysis according to the color or absolute value, edge and texture information in the mapped technical

information and additional consideration of knowledge provided by human experts. We illustrate critical aspects of our conceptual ideas using small synthetic datasets illustrating problems and potential when clustering data emanating from technical (geophysical) sensors and subjective insights or expectations of human experts.

Appendix F

Probabilistic Integration of Tomograms and Logging Data Accounting for Tomographic Ambiguity and Logging Data Errors

A. Asadi, P. Dietrich, H. Paasche
Abstract of the 77th annual conference of the DGG, 2017, Potsdam, Germany

F.1 Abstract

Probabilistic prediction of 2D or 3D images of hydrologic or geotechnical parameters solely measured along one dimension in boreholes can contribute to solve hydrological, petroleum, or engineering exploration tasks. We build on a recently developed fully data-driven workflow interpolating logs of the same hydrological or geotechnical target parameter acquired in different boreholes by considering ill-posed geophysical tomography. Tomographic images between the boreholes are reconstructed using a particle swarm optimization algorithms searching the solution space of the underlying inverse problem globally. We compute multiple tomograms for each available tomographic dataset, which all fit the underlying dataset equally well. We use Artificial Neural Networks (ANNs) to find the optimal prediction models between the computed ensembles of equivalent geophysical tomograms and the sparse measured logging data. During the training phase of ANNs we take the uncertainty of logging data into account as well as the ambiguity of geophysical tomographic image reconstruction to avoid data overfitting when learning the prediction model. Additionally, we account for differences in the spatial resolution of logging data and tomographic models. This approach can be applied to any combination of geophysical tomograms and hydrologic, petroleum or engineering target parameters solely measured in boreholes. To illustrate our workflow,

we reprocess an available field dataset collected at a field site South of Berlin, Germany, to characterize near-subsurface sedimentary deposits. In this example we employ 2D cross-borehole tomographic radar, P-wave, and S-wave velocity models to constrain the prediction of tip resistance, sleeve friction, and dielectric permittivity as target parameters.

Appendix G
Incorporating Hyperspectral Datasets in the Integration Strategy Introduced in Chapter 4

G.1 Process

Hyperspectral datasets provide information about the electromagnetic spectrum for each data point in the map domain (or pixel in the image domain) to achieve information about objects, material or physical processes in the domain. I have tested the method presented in chapter 4 to segment a small part of the Schäfertal catchment (100*100 data point) only based on hyperspectral technical information. The goal of this study was showing the ability of the introduced method in chapter 4 to work with different technical datasets and showing the potential of the hyperspectral datasets to identifying and segmenting structures in the Schäfertal catchment. The used hyperspectral dataset comprises 340 spectral bands. I used 72 bands by selecting every 5^{th} band. This has been done to reduce the amount of data based on the knowledge that neighbored spectral bands are usually highly correlated. Alternatively, principal component analyses striving to describe dominant patterns in the hyperspectral cube by a few eigenvectors are popular data reduction tools, but go always along with filtering out some information, which may be critical.

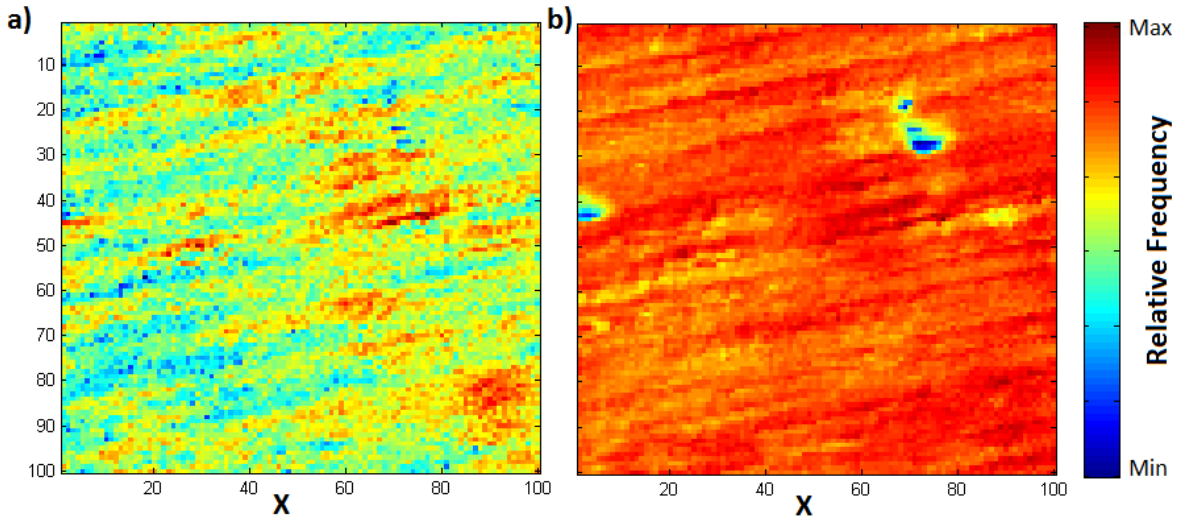

Figure G-1: Two exemplary bands of the hyperspectral data, each band carries random noise and anthropogenic effects.

Figure G-1 presents exemplary three selected bands of the considered hyperspectral information. Like other technical data, the hyperspectral dataset carries noise, e.g., random or anthropogenic effects (e.g., driving lanes of tractors). The data which I had used was $d_{100*100*72}$, a 100 x 100 data point/ pixel region for the considered

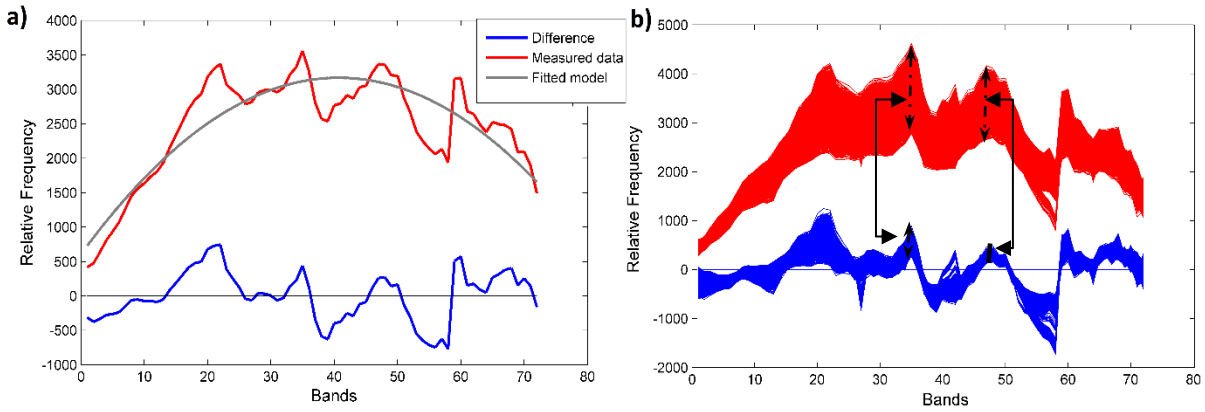

Figure G-2: (a) 72 bands of information for one pixel in 2D area, red, gray, and blue line show measured data, fitted linear model, and difference between measured data and fitted model, respectively. (b) all measured data (red line) and difference between measured data and fitted liner model for 10000 pixel in the 2D area. The selected two position presents the decreasing of the difference between noisy points (lower bands) and normal points (higher bands), in the measured data and fitted data, which in fitted results this difference has been decreased.

72 bands. I considered it as two dimensional data matrix $d_{10000*72}$, with 10000 being equal to the number of data points or pixels in the 2D area. For recognizing the difference between points in the driving lanes (which carry some anthropogenic effect) and normal points (points without any anthropogenic effects), I had compared these points with each other. I recognized that points affected by anthropogenic and environmental noise exhibit systematically decreased values in all bands. For solving this problem I fitted a polygon function in each pixel. Then, I had calculated the difference between fitted model results and measured data. Figure G-2 illustrates this step, Figure G-2a shows measured data, fitted model, and the difference between fitted data and measured data for an exemplary point in the 2D area. Figure G-2b shows the measured data and difference for all 10000 data points in the 2D area. In the further steps for clustering I used the difference data resultant from fitted models instead of the measured data.

Based on the difference data, in each band boundaries have been detected, and I had calculated the total boundary map resultant from all boundary information from 72 bands in the 2D area. Then, based on shortest path calculations the new information vector had been created as input for the clustering method (see workflow of chapter 4). Figure G-3a presents the total boundary map of the 72 filtered hyperspectral bands that resulted from boundary detection and shortest path computation. Figure G-3b shows the

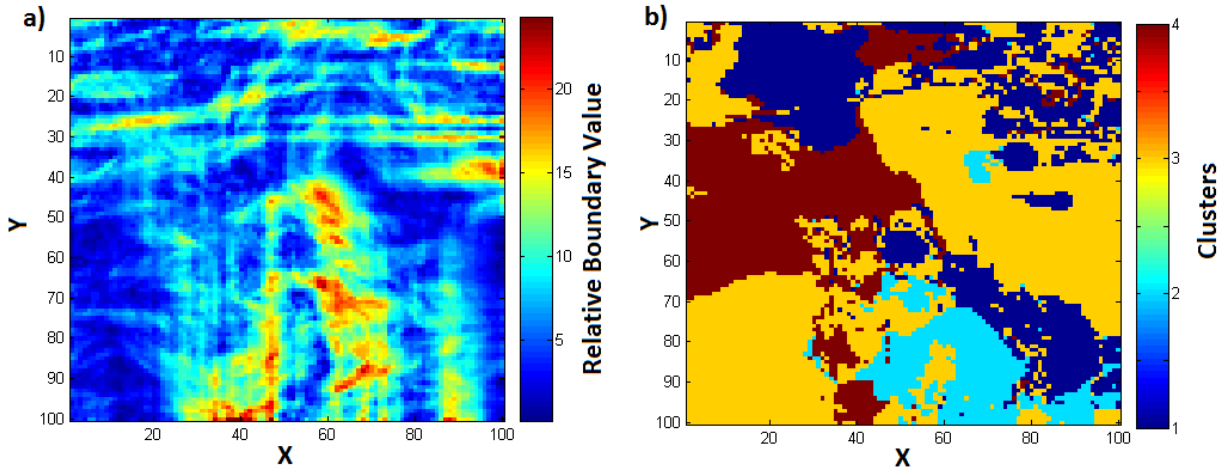

Figure G-3: Results of the clustering method. (a) Total boundary map of selected 72 bands, (b) Clustering results for small part of Schäfertal (100*100 pixel) based on the hyperspectral datasets for desired four clusters.

integration and segmentation results for a 4 cluster solution of the small part (100*100 pixel) of the Schäfertal based on the hyperspectral datasets. For finding the meaning of the clusters, detailed ground sampling would be necessary. In principal, it is possible to expand the results to the entire Schäfertal catchment dataset.

Appendix H
Histogram Normalization

H. 1 Process and Results

In chapter 4 addressing integration and segmentation of subjective and objective maps, I tested different types of data normalization (0 and 1 interval as illustrated in chapter 4, and histogram normalization) in the objective datasets. Here, I present the results of the histogram normalization which is a method in image processing for contrast adjustment based on the image histogram (Hum et al., 2014). Figure H-1 shows the results of the histogram normalization in the objective datasets used in chapter 4 for integration and segmentation of the Schäfertal catchment data. This type of normalization has good ability to highlight the structure in the map or image, but comparing this figure

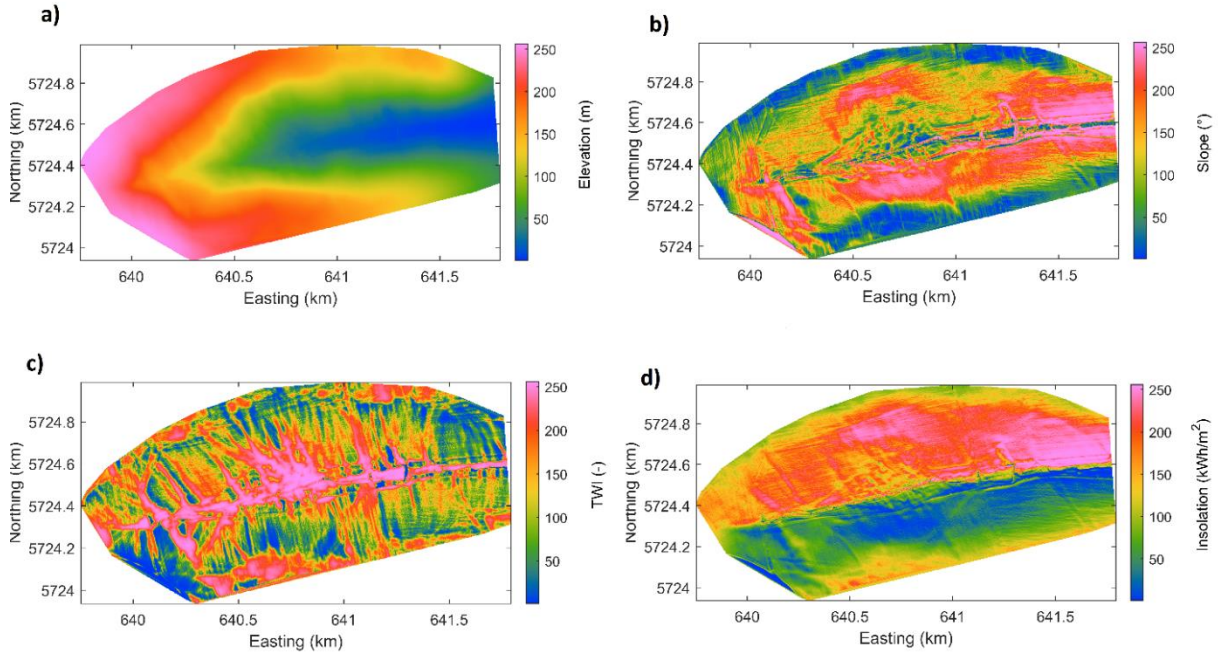

Figure H-1: Histogram normalization for (a) elevation, (b) slope, (c) TWI, and (d) insolation maps of the Schäfertal catchment.

with Figure H-5 in chapter 4, it is obvious that the histogram normalization highlights also the anthropogenic effect (i.e., driving line of the farmer) which is not suitable and realistic for the integration and segmentation method.

I have followed the workflow presented in chapter 4 but now based of the histogram normalized datasets. After normalizing the datasets, the boundary detection had been applied to the subjective (see Figure 4-4 in chapter 4) and normalized maps (Figure H-1). Figure H-2 presents the results of the boundary detection for the histogram normalized data. Figure H-2a and b show the boundary related to the TWI and insolation, respectively. Figure H-2c presents the total boundary of the technical maps which is based on the boundaries of TWI (Figure H-2a), insolation (Figure H-2b) and slope (FigureH-1b). Figure H-2d presents the total boundaries based on the boundary of the subjective maps (Figure 4-11 in chapter 4) and the boundaries of the technical maps (Figure H-2c).

After determining the total boundary map, I had selected 1000 samples based on the sampling strategy described in chapter 4 (presented in Figure H-3a) for calculating

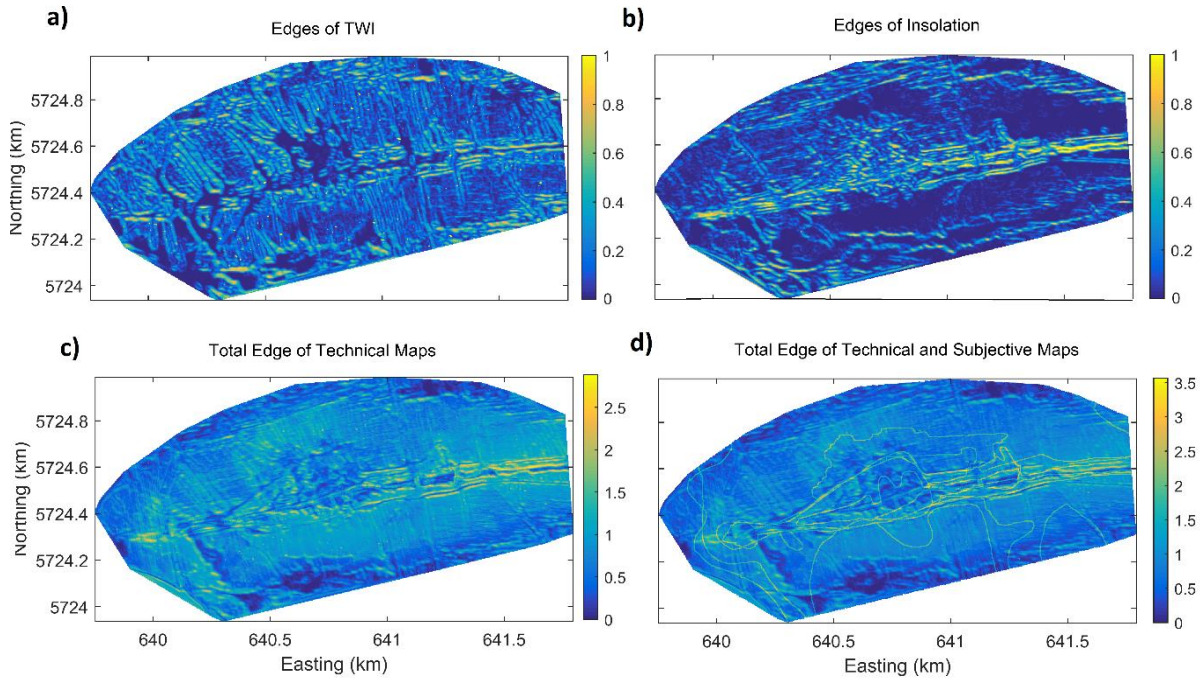

Figure H-2: Boundary detection results. (a) Boundary of TWI, (b) boundary of insolation, (c) total boundary of technical maps, and (d) total boundary of subjective and technical maps.

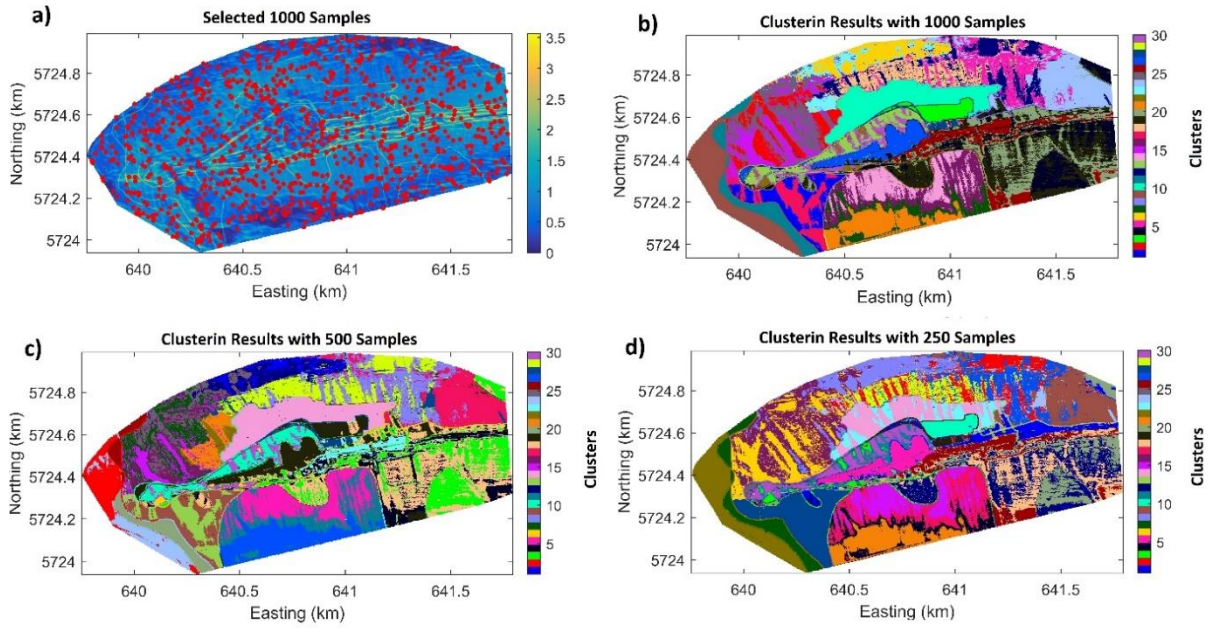

Figure H-3: (a) Distribution of 1000 samples. (b) - (d) clustering results for 30 clusters based on 1000, 500, and 250 samples, respectively.

all shortest paths and path lengths form these samples to all other points in the 2D area of the Schäfertal catchment. After calculating the shortest paths, path lengths, and corresponding similarities the new information vector has been created based on the shortest path results and the similarity matrix. This new information vector had been considered as input for clustering. Figures H-3b and c present the results of the clustering based on 1000, 500, and 250 samples, respectively. In the left part of these figures it is obvious that the driving lanes are presented in the results of clustering. Furthermore, in the up and bottom parts of this area nested clusters are obvious. Comparing these results with clustering results based on the 0 and 1 interval normalization (see Figure 4-14 in chapter 4) I suggest to use 0 and 1 interval normalization in the Schäfertal catchments.

Curriculum Vitae

Abduljabbar Asadi

Date of birth April 04, 1985
Marital status Married, one child
Nationality Iran

Professional Experiences and Leadership

11.2017– now **Lead Data Scientist Africa & Asia Pacific,** Daimler Financial Services AG, Stuttgart, Germany

- Built concepts for different use cases based on the advanced analytics like Customer lifetime value, Next best offer, and Fleet Management.
- Coordinated and successfully implemented the customer risk prediction for collection department in South Africa (based on the R and Microsoft R Services).
- Cooperate with IT group to convert the use cases to product for different countries.

01.2014 – now **Research Associate,** Data Integration and Parameter Estimation Group, Department of Monitoring and Exploration Technologies, Helmholtz Center for Environmental Research – UFZ, Leipzig, Germany

- Research and development of a probabilistic prediction model of geotechnical target parameters with considering data uncertainty (Matlab)
- Research and modeling of subjective and technical maps clustering based on the different machine learning clustering algorithms (Matlab, Python)
- Research and modeling a probabilistic prediction for soil moisture based on the artificial neural networks (Matlab)
- Focus: Applications of data mining and machine learning algorithms (i.e., Clustering, Regression, Probabilistic Prediction, Feature Selection, etc.) in environmental earth databases, collaborate with different groups, presenting results in the conferences, publishing results in scientific journals

07.2013– now **Member of** Helmholtz Integrated Project IP31 “Water and Matter Flux Dynamics in Catchments” in UFZ - Helmholtz Centre for Environmental Research

- Research and modeling a probabilistic prediction for soil moisture based on the artificial neural networks (Matlab)

07.2013– now **Member of IP Predictions group** in UFZ - Helmholtz Centre for Environmental Research Leipzig, Germany

- Attending different meetings in optimization, data mining, Machine learning, and their application in Environmental earth sciences

07.2013- now **Member** of Helmholtz Interdisciplinary GRADuate School for Environmental Research (HIGRADE), UFZ, Leipzig, Germany

07.2012-07.2013 **Head of Department** of Computer and Information Technology, Applied Science University, Culture and Art Kurdistan Branch, Iran

- Managing different programs, lectures and courses in the computer and IT field of study for students in the bachelor level
- Supervising the research group with focusing in the data mining
- Supervising more than 30 bachelor thesis in the computer and IT field of study

08.2010-05.2012 **Head of Department** of Information Technology, Applied Science University, Divandare Branch, Iran

- Managing different programs, lectures and courses in the computer and IT field of study for students in the bachelor level
- Supervising the research group with focusing in the data mining
- Supervising more than 20 bachelor thesis in IT field of study

Education

01.2014 –now **PhD Student,** Faculty of Sciences, Eberhard-Karls-University of Tübingen, Germany

- Focus: Geoinformatics, Data integration and parameter estimation, Data mining and Machine learning algorithms, Environmental earth sciences data sets, Remote sensing data sets, Geophysics
- Advisor: Prof. Dr. Peter Dietrich
- Thesis theme: Advanced data mining and machine learning algorithms for integrated computer-based analyses of big geoscientific databases

09.2009 - 09.2012 **M.Sc.**, Software Engineering, Islamic Azad University Branch of Zanjan, Zanjan, Iran

- Focus: Data mining and machine learning algorithms, Mathematics, Databases, Software engineering, Optimization algorithms, Networks
- Advisor: Dr. Mehdi Afzali
- Thesis theme: A new method for detecting positive and negative association rules using PSO algorithm
- Overall GPA: 3.572 of 4 (very good)

09.2006 -10.2008 B.A., Software Engineering, Shahid Bahonar Technical and Vocational University, Shiraz, Iran

- Focus: Programming, Data structure, Databases, Operation system, Computer modelling, Mathematics, Software engineering, Networks
- Thesis theme: Development a GUI Based Application for Cinema based on the C# Programing Language
- Overall GPA: 3.14 of 4 (good)

2003-2006 **College (Fachschule),** Computer Sciences **–** Software, Technical College I Tehran, Iran

- Focus: Programming, Data structure, Databases, Operation system, Computer modelling, Mathematics, Software engineering, Networks, Information, Internet
- Thesis theme: Development of a website for shoes marketing based on the C# ASP.Net Programing Language
- Overall GPA: 2.83 of 4 (average)

1999- 2003 **High school,** Sanandaj, Iran

- Ranked as best student

Teaching Experience

2012-2013 Lecturer, Department of Computer and Information Technology, Applied Science University, Culture and Art Kurdistan Branch

2010-2012 Lecturer, Applied Science University, Divandare Branch

2010-2012 Lecturer, Applied Science University, Bijar Branch

2009-2011 Lecturer, Department of Computer, Marivan Branch, Islamic Azad University

2009-2012 Lecturer, Applied Science University, Jahad Danshghahi Branch

2009-2010 Lecturer, Department of Computer, Sannandaj Branch, Islamic Azad University

Scientific Journal Papers

2016 **Asadi, Abduljabbar,** Peter Dietrich, and Hendrik Paasche. "2D probabilistic prediction of sparsely measured earth properties constrained by geophysical imaging fully accounting for tomographic reconstruction ambiguity." Environmental Earth Sciences 75.23 (2016): 1487.

2017 **Abduljabbar Asadi**, Peter Dietrich, and Hendrik Paasche (2017). ”Spatially continuous probabilistic prediction of sparsely measured ground properties constrained by ill-posed tomographic imaging considering data uncertainty and resolution.” GEOPHYSICS, 82(3), V149-V162.

2012 **Abdoljabbar Asadi**, Mehdi Afzali, Azad Shojaei and Sadegh Sulaimani, ”New Binary PSO based Method for finding best thresholds in association rule mining”, *Life Science Journal*, , 9(4), pp: 260-264.

2012 **Abdul-Jabbar Asadi**, Azad Shojaei, Salar Saeidi, Salah Karimi and Ebad Karimi, ”A new method for the discovery of the best threshold value for finding positive or negative association rules using Binary Particle Swarm Optimization”, *IJCSI Journal*, Volume 9, Issue 6, No 3, pp:315-320.

2012 Azad Shojaei, **Abdoljabar Asadi**, Rahim Rashidi,”Distributed Routing Algorithms in Dynamic Wireless Networks“, *Journal of American Science*;8(7),pp:335-337.

2012 Azad Shojaei, **Abdoljabar Asadi**,”Presenting a Parallel Algorithm for Constructing Cartesian Trees and its Application in Generating Separate and Free Trees”, *Journal of American Science*, 8(6), pp: 811-813.

Conference Papers and Presentations

30 May - 2 Jun 2016 **A. Asadi** et al.,”2D Probabilistic Prediction of Sparsely Measured Geotechnical Parameters Constrained by Tomographic Ambiguity and Measurements Errors,” *78th EAGE Conference & Exhibition 2016*, Vienna, Austria.

6-10. Sep 2015 **A. Asadi** et al.," Predicting continuous distributions of sparse data under full consideration of tomographic reconstruction ambiguity," *Near Surface Geoscience 2015 - 21st European Meeting on Environmental and Engineering Geophysics*, Turin, Italy.

23- 26. Mar 2015 **A. Asadi** et al., "Predicting porosity according to ensembles of co-located radar and seismic tomographic models with Artificial Neural Networks," *75.Jahrestagung der Deutschen Geophysikalischen Gesellschaft, Hannover*, Poster.

23- 26. Mar 2015 **A. Asadi** et al., "Conceptual developments for clustering mapped data emanating from technical sensors and subjective insights of human experts," *75.Jahrestagung der Deutschen Geophysikalischen Gesellschaft, Hannover*.

12-14. Mar 2013 **A. Asadi** et al., " Provide automatic method for finding the optimal threshold value and discovering efficient association rules using Binary Particle Swarm Optimization," *18th National CSI Computer Conference, Iran, Tehran*, In Persian.

22-28. Feb 2013 **A. Asadi** et al.," A new method for automatic discovery of threshold value and positive or negative association rules," *11th Iranian Conference on Intelligent Systems, Kharazmi University, Tehran,* In Persian.

04-05. Dec 2012 **A. Asadi** et al.," A new method for the discovery of the best threshold value for finding association rules using Binary Particle Swarm Optimization," *THE SIXTH DATA MINING CONFRANCE IDMC'12 . IRAN, TEHRAN*.

04-05. Dec 2012 **A. Asadi** et al.," Provide a new method for detecting positive and negative optimal performance association rules in very large databases using Binary Particle Swarm Optimization, "*THE SIXTH DATA MINING CONFRANCE IDMC'12 . IRAN, TEHRAN*.

2011 B. Maghsodi, B. Nori, and **A. Asadi**," Using data mining to predict the amount of gas," *Third National Conference on Computer Engineering and Information Technology, Hmadan*, Iran, In Persian.

2011 **A. Asadi** et al.," A bidder offering the best service components," *Conference on Computer Engineering and Information Technology, Islamic Azad University, Bukan*, Iran, In Persian.

Attend Courses and Workshops

16-18 Dec 2015	Spatial Point Pattern Analysis, UFZ-Leipzig, Germany.
7-8 Dec 15 / 8 Jan 2016	Time Series Analysis Using Matlab, Leipzig, Germany.
11-13 Mar 2015	Image Processing, UFZ- Halle, Germany.
2014	Optimization Algorithms, UFZ, Leipzig, Germany.
2014	Applied Geophysics and Hydrogeophysics in Geophysical Field Seminar, 10 days, University of Göttingen, Germany.
2014	Geographic Information System (GIS), University of Tubingen, Germany.
2010	Workshop on Data Mining, Tehran, Iran.
2010	Workshop on Data Mining, Kurdistan University, Sanandaj. Iran.

Awards and Honors

2011	Best teacher at the University of Applied Sciences, Divandare Branch
2002	Third stage in the student part of Khwarizmi International Award at the state of Kurdistan, Iran
2001-2003	ranked as best student in Taleghani Vocational School, Sanandaj, Kurdistan, Iran

Interests

Research Areas:
- **Data Mining**
- **Machine Learning**
- **Geoinformatics**
- **Optimization Algorithms**

Application:
- **Business intelligence**
- **Market analysis**
- **Geoinformatics**

Topics:
- Big Data
- Spatial and temporal data
- Incomplete data sets and noisy observations

Skills

- **Languages:** English (Fluent), German (B2), Persian (Native), Arabic (B1), Kurdish (Mother Language).

- **Programming:** Python (Semiskilled), C# (Proficient), ASP.Net (Proficient), C++/C (Semiskilled), MATLAB/Octave (Proficient), Pascal (Proficient), Visual Basic (Proficient), JavaScript(Familiar), Assembly (Proficient).

- **Data Mining Programs:** R (Proficient), Microsoft R Services (Proficient), MATLAB (proficient), Weka (proficient), BI in SQL Server (Semiskilled).

- **DBMS:** SQL Server (Semiskilled), Access (Semiskilled).
- **Visualization:** d3 (Semiskilled).
- **GIS:** ArcGIS (Familiar).
- **Platforms:** Windows (Proficient), Linux (Familiar).
- **Others:** Hadoop (Familiar), DirectX Programing, Photoshop, Multi-Threaded and Parallel Programming, Java (Familiar).

Hobbies and Interests

- **Music & Art:** Visiting Concerts, Cinema, and Theater, Photography.
- **Sport:** Football, Running, and Swimming.
- **Others:** Reading, Internet, Visiting Family and Friends, Computer Programming