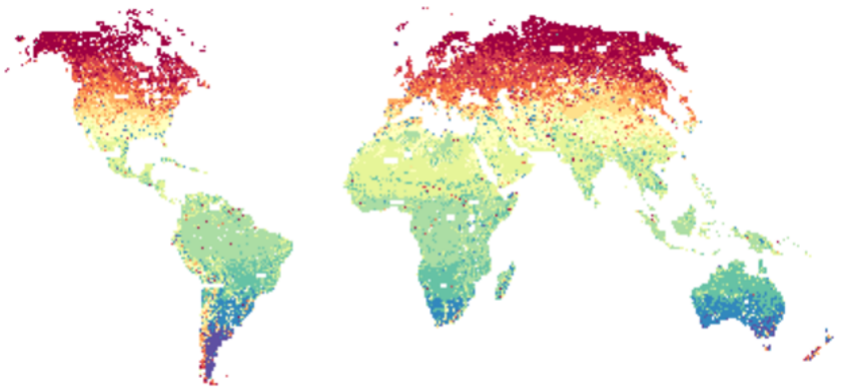
Geo-information Science and Remote Sensing

Thesis Report GIRS-2020-22

ADDED VALUE OF SYNTHETIC PROFILES

With Particular Reference to Soil Organic Carbon Predictions in SoilGrids



Anna Christien Schoneveld
May 12, 2020

WAGENINGEN
UNIVERSITY & RESEARCH

Added value of synthetic profiles

With Particular Reference to Soil Organic Carbon Predictions in SoilGrids

Anna Christien Schoneveld

Registration number 94 12 27 744 080

Supervisors:

Gerard Heuvelink (SGL)
Sytze de Bruin (GRS)
Luis de Sousa (ISRIC)

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

May 12, 2020
Wageningen, The Netherlands

Thesis code number: GRS-80436 Thesis Report: GIRS-2020-22
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Abstract

The purpose of this study is to investigate the effects of synthetic profiles on predictions in Digital Soil Maps (DSM). Synthetic profiles are artificial observations based on expert knowledge. This study was conducted on SoilGrids a data-driven product from ISRIC. SoilGrids is a DSM predicting soil properties at standard depths within a global soil mask. The process of making predictions in DSM relies heavily on soil sample data being representative for an area. The soil samples should, therefore, be allocated in geographical space and environmental covariate space (feature space) in such a way that soil-environmental relationships are adequately represented. Statistical inference beyond this sub-population implies extrapolation and can result in incorrect predictions. In this thesis, a solution to avoid extrapolation in feature space was found in the incorporation of synthetic profiles.

In the first part of this thesis the location of geographical areas in SoilGrids which are underrepresented in feature space were determined. The shortest distance towards training profiles in feature space for aggregated prediction raster cells in SoilGrids, the so-called dissimilarity indication, was calculated. The patterns of dissimilarity indication coincided with mountainous areas. An interview was then held with soil expert from ISRIC to acquire quantitative estimates for the target soil property SOC (Soil Organic Carbon) for two selected study areas (Tarim and Qaidam Basin in West-China) to generate synthetic profiles. In the execution phase of the thesis the SoilGrids models were run with synthetic profiles to make predictions for SOC in these areas. The new SOC predictions support the expectations that synthetic profiles influence the prediction. The SOC predictions decreased and went towards the direction of the expert estimates. The results suggested also that synthetic profiles have a smoothing effect and should therefore be used with care.

Acknowledgements

I would like to thank my supervisors for their supervision and especially Luis for his helpful advice regarding scripting and his positive thinking throughout this research. I would like to thank the experts for their time and providing me with their knowledge on Soil Organic Carbon in Western China. Furthermore, I would like to thank all MGI students in Thesis ring 1 who provided me with tips and tricks for writing a thesis and for sharing their thoughts on my writing style. Last but not least I would like to thank my parents and friend Jeroen for their moral support which assisted me in the pursuit of my academic goals and writing this master thesis.

Acronyms

CSM Conventional Soil Maps 1

CLORPT CL-imate, O-rganisms, R-elief, P-arent material and T-ime

DSM Digital Soil Maps 1

GIS Geographic Information System 51

GRASS Geographic Resources Analysis Support System 51

HPC High Performance Computing cluster 13

ISRIC International Soil Reference and Information Centre 2

LHS Latin Hypercube Sampling 13

NDVI Normalized Difference Vegetation Index 2

QR Quartile Reduction approach 13

SOC Soil Organic Carbon: fine earth fraction in g per kg 1

SOM Soil organic matter 1

stddev Standard Deviation 51

TP Training Profiles 51

WoSIS The World Soil Information Service 2

Contents

Chapter 1

Introduction

1.1 Context and Background

Worldwide, soil is recognized as a major contributor to ecosystem services, such as food production and climate regulation (Sanchez et al. 2009). Correct and comprehensive maps of soil properties can help policy makers to make decisions that reduce risks associated with climate change, natural and man-made hazards, and food security (Panagos et al. 2012). One of these important soil properties is Soil Organic Carbon: fine earth fraction in g per kg (SOC), which plays an important role in controlling the function and quality of soil and offsetting the emissions of greenhouse gases (Guo et al. 2019). SOC is closely related to Soil organic matter (SOM). SOM represents the remains of roots, plant material, and soil organisms in various stages of decomposition and is variable in composition. While soil organic matter occurs in small amounts in the soil it has a major influence on soil aggregation, nutrient reserve and its availability, moisture retention, and biological activity (Amrita 2013). Soil organic carbon (SOC) refers only to the carbon component of SOM. The spatial patterns and total amounts of SOC are important for studies of soil productivity, soil hydraulic properties, and the cycling of C-based greenhouse gases (Kern 1994).

1.1.1 Conventional soil mapping

The mapping of SOC is done traditionally with Conventional Soil Maps (CSM). In CSM, soil experts first construct a soil– and landscape model of the area of interest through intensive and costly fieldwork (Yang et al. 2011). CSM consists of soil bodies represented as discrete, homogeneous entities in so-called soil classes (Kempen et al. 2012). The Data in CSM are therefor highly summarized to fit in the polygon-based framework (soil classes), which reduces the level of detail (Hartemink et al. 2010). CSM is especially limited for global quantitative environmental studies since these studies ask for soil properties rather than soil classes.

1.1.2 Digital soil mapping

Since CSM is limited in terms of both the level of spatial detail and the accuracy of the soil attributes (Zhu 1997), Digital Soil Maps (DSM) have emerged and developed over the past decades to fulfill the increasing global demand for quantitative information on soil properties like SOC (Kempen et al. 2012; Lagacherie 2008).A digital soil map typically is a raster composed of two-dimensional cells (pixels) organized into a grid in which each cell has a specific geographic location and contains soil

data (*SSM - Ch. 5. Digital Soil Mapping*). The soil data or property predictions are based on the relationship between the the soil and environmental covariates. Environmental covariates are measurable properties such as elevation, Normalized Difference Vegetation Index (NDVI), annual rainfall, and temperature. Environmental covariates must be available at all prediction locations (raster cells), in other words they must cover the global soil mask. In a statistical model the relationships are fitted and the learned relationships are then applied to locations where soil data are not available using statistical (machine) learning techniques (Heung et al. 2016). Digital soil maps illustrate thus the spatial distribution of soil classes or properties with the use of statistical models that predict these soil properties from covariates based on training data. DSM is an alternative to CSM and is rapidly taking over due to the following advantages:

- DSM is easy to integrate with most other forms of natural resource data that are grid-based e.g. satellite imagery, digital elevation models and climate data (Hartemink et al. 2010).
- DSM enables updating maps and is reproducible since the prediction models are stored and can be rerun when new data become available (Heuvelink et al. 2010).
- Different models of spatial variation can be chosen, and the proper use of (geo)statistical methods results in predictions with quantified uncertainty (Kempen et al. 2012).

1.1.3 SoilGrids

SoilGrids is a data-driven product developed and maintained by International Soil Reference and Information Centre (ISRIC) – World Soil Information. SoilGrids is a DSM predicting soil properties (among others SOC) on a global scale at 250 m resolution for six standard depths (ISRIC 2019). The 2.0 version of SoilGrids predicts at the midpoints of the depth intervals, that is at the centre of the 0-5, 5-15, 15-30, 30-60, 60-100, and 100-200 cm intervals. In SoilGrids, soil training profiles from The World Soil Information Service (WoSIS), a database for soil samples from national and regional soil profile databases with their corresponding coordinates are used as input data (Batjes et al. 2017). This input data is overlay-ed with more than 150 remote sensing-based environmental covariates (Hengl et al. 2017). These covariates come from remote sensing data repositories and are related to the soil-forming factors from the CLORPT model. The concept of this model is that soils are a function of five key environmental factors, namely climate (cl), organisms (o), relief (r), parent material (p), and time (t) (Jenny 1994). Machine learning methods like random forest in SoilGrids need to be trained on the feature space with a training dataset to detect meaningful patterns in the data. is used in SoilGrids to fit a model based on the training data, i.e. the WoSIS soil training profiles and make spatial predictions (see Figure 1.1). (Shalev-Shwartz and Ben-David 2014)

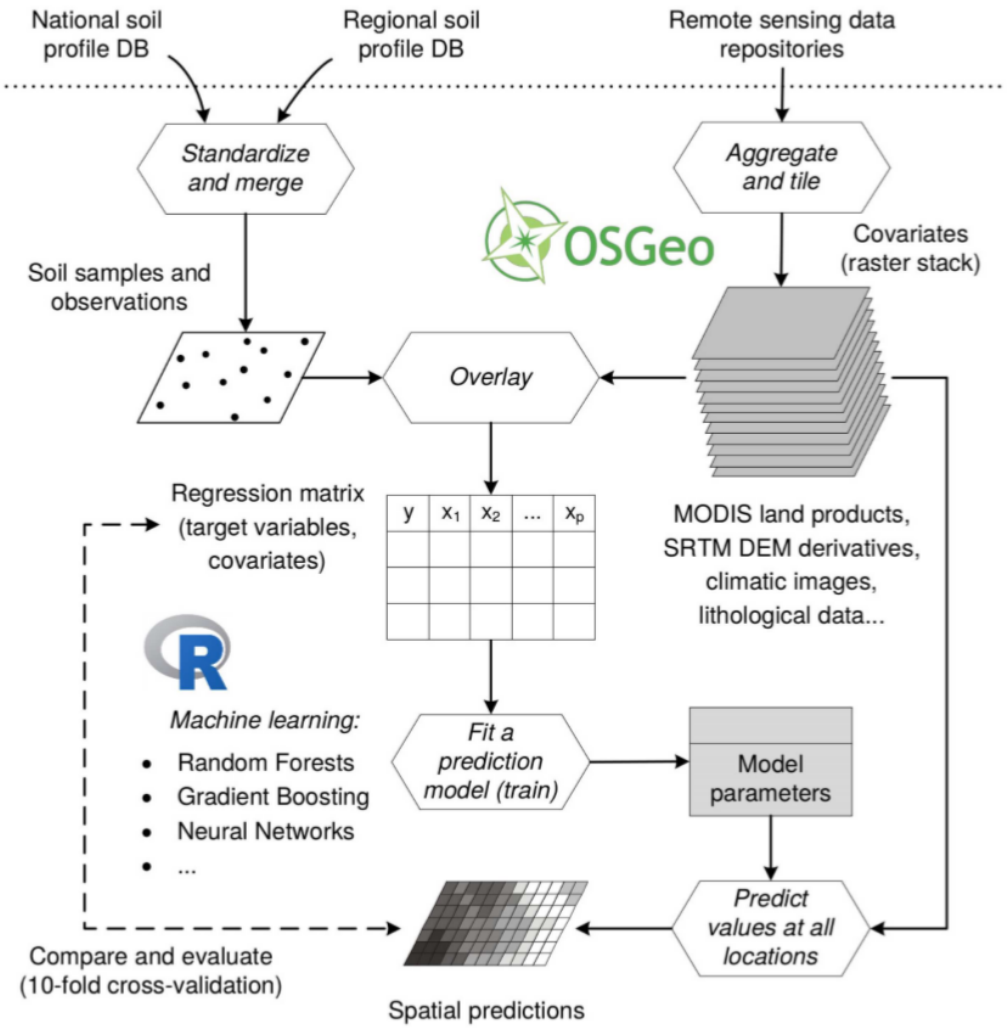


Figure 1.1: The (data-driven) statistical framework used for generating SoilGrids (Hengl et al. 2017)

1.2 Problem definition

Digital soil mapping relies on soil sample data for model fitting. Unfortunately, it is not possible to obtain large volumes of soil data due to the high costs of field sampling and laboratory analysis and the limited financial resources of soil projects. Each soil project used a sampling design to make a selection from the total population from which measurements were made (Pennock et al. 2007). To acquire sampling data that are 'representative' for an area (or the Earth) it is required to allocate soil samples in geographical space and environmental covariate space (feature space) in such a way that soil-environmental relationships are adequately represented (Zhu et al. 2015).

1.2.1 Soil sample distribution in geographic space

The number of soil samples in WoSIS is 150,000, however some large areas that have extreme climatic conditions and/or have very restricted access may be systematically excluded from soil sampling projects (Hengl et al. 2017). These areas result in a significant under-representation in DSM that

predict on a continental to global scale, like SoilGrids (Hengl and MacMillan 2019). Also, often agricultural areas and developed countries are more densely sampled than non-agricultural zones and the developing world (Hengl et al. 2017). See Figure 1.2 for the distribution of soil training profiles in the world).

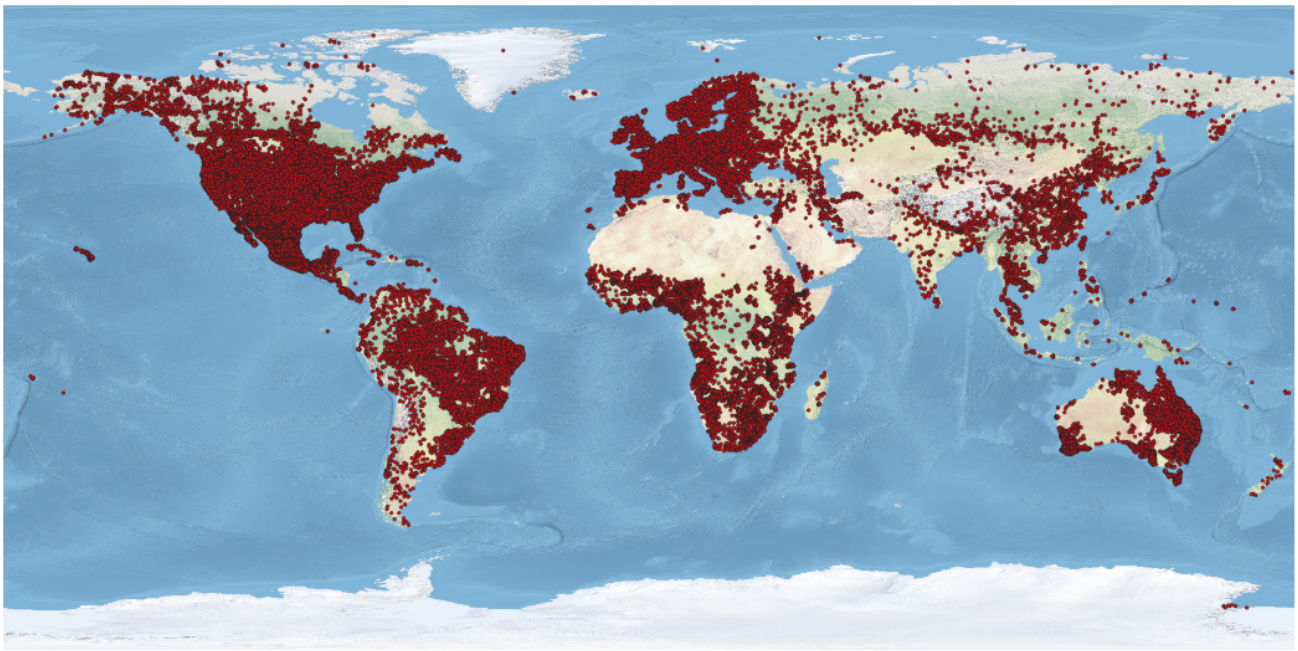


Figure 1.2: World distribution of soil profiles from WoSIS via WFS (September 2019). (ISRIC 2019)

Statistical inference concerns the sub-population which has been sampled. Predictions beyond this sub-population implies extrapolation, i.e., drawing conclusions about something beyond the range of the data (QA et al. 2002). Extrapolation in feature space is risky therefor in the 2017 version of SoilGrids pseudo-observations were introduced. These pseudo observation tried to avoid incorrect predictions for underrepresented areas (Hengl et al. 2017). The 100 – 400 synthetic profiles in SoilGrids were randomly positioned within sand dune, glacier, and mountain areas with assigned soil properties and soil classes based on the knowledge of experts(Hengl et al. 2017). The way pseudo observations were constructed in the SoilGrids version of 2017 was in a approximate manner, but not exact, fully formed, or scientific reliable.

In this master thesis we want to redo this for the large gaps which exist in terms of representing the entire feature space of the prediction points. The pseudo observations focused on geographically unsampled areas. The term synthetic profiles is therefor used to make a clear distinction between what is done in SoilGrids 2017 and what will be investigated in this master thesis. Synthetic profiles are artificial observations based on expert knowledge to enlarge the training dataset and therefore make better predictions. Since expert elicitation provides an estimate of the possible outcome without the need for large expensive soil sampling fieldwork, synthetic profiles are potentially a cheap tool to fill in the gaps in the training data (Haakma et al. 2011). In the past, most of the studies using expert knowledge used it to identify soil-forming factors and processes. These are examples of epistemic uncertainty. where there is a lack of knowledge and expert elicitation is used to fill the gap (Knol et al. 2010). The direct placement of field observation with expert informed guesses will be further

explored.

1.2.2 Environmental covariate space

The environmental covariate space (also called feature space) is the collection of environmental covariates that are used to characterize the data. The feature space is a d-dimensional space where d is the number of covariates. Recall that SoilGrids has over 150 predictor variables. so d can be quite large. The set of feature (covariate) values for a particular soil profile location form a covariate vector. All raster cells in SoilGrids (also called prediction points) have a similar covariate vector in the environmental covariate space. The covariate vector from the exact geographical location of the soil training profiles are linked with the training profiles soil observations and thereby define a training dataset.

In high dimensions, human geometric imagination is limited. So in Figure **??** a simplistic feature space is represented with only two covariates. Imagine that all circles are covariate vectors, blue circles represent soil training profiles, while red circles are prediction points, i.e. SoilGrids raster-cells. Covariate vectors that are close to each other in feature space have similar covariate values.

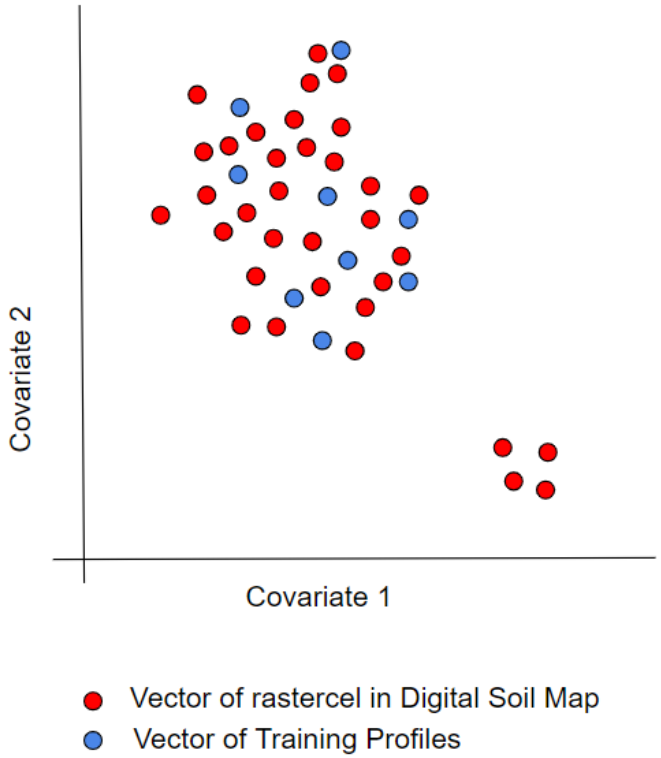


Figure 1.3: Simplistic representation of the feature space with only two dimensions.

In the feature space shown in Figure 1.3, two clearly distinct groups of covariate vectors for prediction points (red circles) in feature space can be observed. The ones close to covariate vectors with soil information (blue circles) form the first group, these are covariate vectors in feature space that

have covariate vectors with soil observations nearby in feature space. The homosoil method assumes that similar environmental conditions have similar soil properties (like SOC) (Mallavan et al. 2010). Training profiles that are close to a covariate vector in feature space of a prediction point can serve as a surrogate for those locations since similar environmental conditions occur (Zhu et al. 2015). Since statistical inference is more capable of doing interpolation than extrapolation, whenever there are training data close in feature space it will be easier to predict for that prediction point, and the prediction will be more accurate. For example, many deserts may not be sampled but there may be data available from another desert which has similar conditions (covariate values). In such case, these data contribute largely to predicting accurately these unsampled deserts.

The second group consists of red circles that have no blue circles nearby (see Figure 1.3 bottom right). For these covariate vectors, machine learning techniques become unreliable since it has to extrapolate in feature space rather than interpolate. Our main concern is this second group because extrapolation in feature space can result in incorrect predictions. In order to improve SoilGrids, more focus will need to be put on improving the feature space representation in the training data, by adding synthetic profiles (Hengl and MacMillan 2019).

1.3 Research Objective and Questions

From the above is is clear that there are two ways in which extrapolation for predicting a prediction point for soil organic carbon may occur.

- **Extrapolation in Geographic Space** - Locations on Earth (250m SoilGrids raster cells) without training profiles with soil information for soil organic carbon close by.
- **Extrapolation in Feature Space** - Locations on Earth (250m SoilGrids raster cells) where their vector in feature space is far away from covariate vectors of training profiles for soil organic carbon.

This master thesis addresses the latter. The following research objective is defined:

Assess the added value of incorporating synthetic profiles in SoilGrids for areas which are under-represented in feature space.

The objective will be realized by answering the following research questions:

1. Which geographical areas in the world are under-represented in the WoSIS database in feature space for soil organic carbon?
2. Which methods can be used to generate synthetic profiles and how can the level of confidence of the synthetic data be quantified?
3. What is the effect of using synthetic profiles and their spatial density on soil organic carbon predictions made in geographical areas which are under-represented in feature space?

Chapter 2

Methodology

The process of determining the added value of incorporating synthetic profiles in geographical areas where predictions otherwise would be extrapolated is divided into three steps. The first step is to determine the geographical areas where their covariate vector in feature space is far apart from the covariate vectors of soil training profiles. These areas are underrepresented in feature space. The second step is to generate synthetic profiles for these areas with expert knowledge. The third and last step is to incorporate these synthetic profiles and recalculate the target soil property, Soil organic carbon (SOC), for these areas, and compare predictions with a case in which synthetic profiles are not used. The first step is addressed in Sections 2.1 and 2.2, the second in Section 2.3 and the third in Section 2.4.

2.1 Calculating distance in Feature Space

For determining the geographical areas which in feature space are far apart from the nearest training profile we need to calculate distance in feature space. In order to provide a meaningful notion of dissimilarity between two covariate vectors in high dimensional feature space, a distance function is needed. In this thesis the Manhattan distance has been chosen (see section2.2). The distance function must compute the distance between the covariate vector of all raster cells (prediction points) in a DSM and the covariate vector of each soil training profile to find the nearest soil training profiles. The shortest possible distance from a covariate vector of a raster-cell (prediction point) towards a covariate of a soil training profile (observation point) for SOC is the dissimilarity indication. When the dissimilarity indication is high for all raster cells in a geographical area, this indicates a geographical area that is underrepresented in feature space.

2.1.1 Standardization

Before the Manhattan distance can be calculated the covariates need to be standardized, because they were measured at different scales. This is necessary to compare the distances in every dimension (covariate). When a covariate is measured on a scale ranging from 1000 to 5000 a difference of 5 does not have much impact for that covariate, however this same difference is on a scale from 1 to 10 enormous. The covariates of the soil training profiles and raster cells of the DSM both need to be standardized. The standardizing process produces standard scores (z-scores) that represent the number of standard deviations above or below the mean that a specific observation falls. The mean

of each covariate was substracted from its value, and the result divided by the standard deviation, see Equation 2.1

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

where:
$\mu = Mean$
$\sigma = Standard\ deviation$

2.1.2 Manhattan Distance

The Manhattan distance measures the distance in feature space. Given two random points A and B in a d-dimensional space, the Manhattan distance is the sum of the absolute differences in the d dimensions:

$$|d_{AB}| = \sum_{i=1}^{d} |a_i - b_i| \tag{2.2}$$

Note that each dimension (covariate) is deemed equally important. Since the aim is to avoid extrapolation in feature space when making predictions for SOC, only the covariates that were important in predicting the SOC-values in SoilGrids were used. The covariates list that consists of the 77 covariates can be found in the Appendix (Appendix .1).

2.1.3 Qualitative variables

Calculating the dissimilarity indication is only applicable when all covariates are quantitative variables. However, some of the covariates in SoilGrids are qualitative variables. Fortunately these covariates do not consist of several categories with corresponding integers representing the categories but are rather binary. A raster cell in SoilGrids falls in a certain category or it does not. Consequently the distance in feature space between a training profile and a prediction point can be easily computed for these qualitative covariates. Whenever both fall or do not fall in a certain category their distance for that dimension is 0, otherwise, there is a distance between the training profile and the prediction point.

2.2 Data reduction

Given the many $250\ m \times 250\ m$ raster cells and training profiles, the computation of all Manhattan distances between training and prediction points will be a formidable task. The computational burden of this task is reduced by data-based parallelism based on tiles, the aggregation of the 250 m raster-cells to 10 km raster-cells, and the reduction of soil training profiles, (see Figure 2.1 for an overview of the data reduction methods). Each of these three steps is explained in the next subsections.

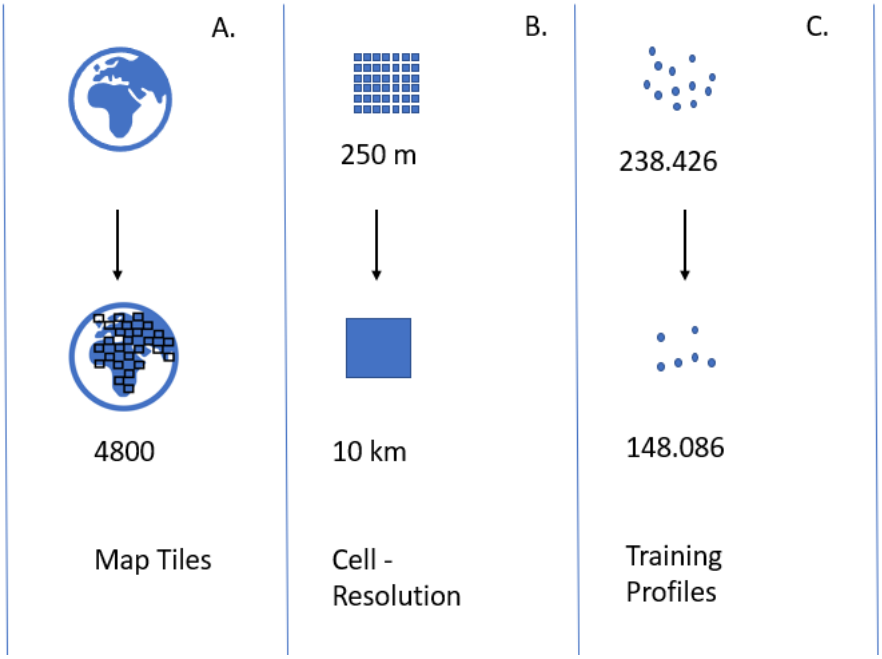


Figure 2.1: Data Reduction: A. Data-based parallelism based on tiles, B. Aggregation of raster-cells to 10 km, C. Reduction of training profiles.

2.2.1 Tile-based Parallelism

Part of increasing the computation speed was achieved by a data-based parallelism. The prediction points are distributed across different computers (called 'nodes') to calculate the dissimilarity indication. The division of the data was done by a tile-based mapping system called tile4800. After a projection the two-dimensional surface of the Earth was divided into a series of regularly spaced grid cells (Sample and Ioup 2010). This resulted in a total of 4800 different tiles. Each of the 4800 tiles overlap entirely or partly with a part of the soil mask of the Earth, (see Figure 2.2). Each tile could then in parallel with other tiles calculate the dissimilarity indication for all prediction points in the tile.

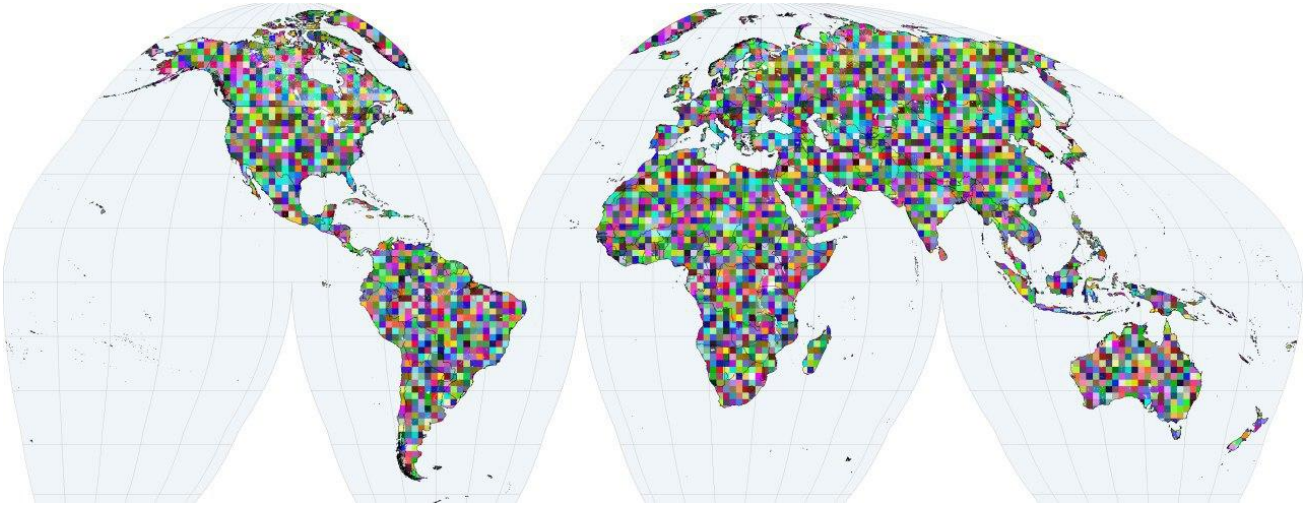


Figure 2.2: World divided in 4800 tiles.

The 4800 tiles were submitted as jobs to Anunna. Anunna is the High Performance Computing cluster (HPC) cluster which has a large number of nodes to utilize parallel computing and is hosted by Wageningen University. The workload manager of Anunna, SLURM, distributes the different tiles across different nodes and therefore access more computing power.

2.2.2 Aggregation of cell resolution

Since a tile covers an area of 200 km by 200 km and the raster cells have a cell resolution of 250 m by 250 m, a total of 640 000 SoilGrids prediction points (raster-cells) could potentially be on one tile. To increase the computational speed 40 by 40 raster cells of each 250m by 250m were combined to one cell. Hence the cell size was enlarged from 250 meter to 10 km. The $10\ km \times 10\ km$ cells received the covariate value of the most centered 250 m cell in the 10 km extent. After aggregation, a tile had a maximum of 400 cells instead of 640 000.

2.2.3 Reduction of the Training Profiles

Only the training profiles with soil information regarding SOC were used. This is an initial total of 238.426 soil training profiles. With data cleansing, the training profiles which missed column values (covariate) were omitted (3,086 in total). However, the most reduction in training profiles took place with the Quartile Reduction approach (QR) were the number of training profiles went down from 235.340 to 148.086.

The Quartile Reduction approach aims to get an even distribution of training profiles in feature space close to the frequency distributions of the different covariates. The Quartile Reduction approach is based on Latin Hypercube Sampling (LHS). LHS is a sampling method that stratified the input probability distribution into equal intervals whereby each sample is randomly taken from one axis-aligned hypercube (Iman 2016).

In feature space, covariate vectors are merely a list of values, where every covariate/dimension has its position in the list. For the Quartile Reduction approach, the value of the covariate was altered according to the quartile in which the covariate value is in. The global quartile boundaries of covariates were used. The values of the covariates of the training profiles were transformed to quartiles represented by integers 1 to 4, with 1 meaning the value falls between the minimum value and Quartile 1, 2 meaning the value is between Quartile 1 and the median (Quartile 2), and so on, (see Figure 2.3).

The covariate vectors are transformed to a list of 77 items and each item consists of one of the integers 1, 2, 3 and 4. These items were pasted together in a string of length 77, the so-called Q-Label. the length is 77 since the total number of covariates is 77. Every training profile has its own Q-Label. Covariate vectors with identical Q-labels are in the same hypercube and are near to each other in feature space. The assumption was made that only one observation was needed in a hypercube to get an even distribution of observations in feature space. Duplicate Q-labels were therefore deleted, the first occurring training profiles in the list of the two duplicates was kept. The final list of training profiles (TP_QR) consists of 148.086 training profiles.

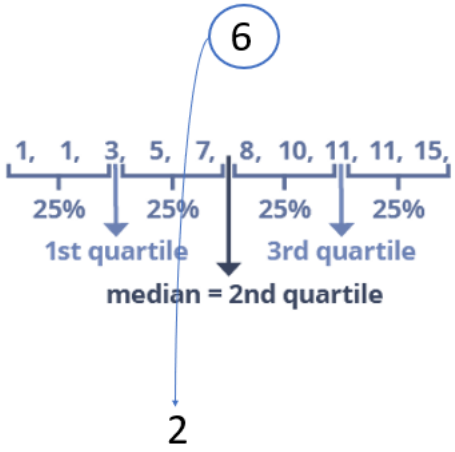


Figure 2.3: The value of a covariate is changed to an integer between 1 and 4 according to the quartile in which the covariate value belongs to. In the example, the value 6 becomes a 2 since it is between the boundary of Quartile 1 and Quartile 2

2.2.4 Calculating Manhattan distance

The last step is calculating the dissimilarity indication which is the shortest distance in feature space from each prediction point (10 km) towards one of the training profiles of TP_QR. An R script was written (T01_Run_Tile) to calculate the dissimilarity indication. On the HPC Anunna, every tile (4800 in total) was run once with this script. The result for every tile was saved and later compressed together, resulting in a world map showing the dissimilarity indication. For an overview of all software implementations, see Appendix .2.

2.3 Expert elicitation

The second research question of this research addresses the generation of synthetic profiles. Synthetic profiles are artificial observations based on expert knowledge, obtained via expert elicitation. Expert elicitation is a decision analysis technique used to gather the professional judgments of an individual with expertise in a required field (Cruickshank 2018). The structure of this section will follow the seven-step protocol of Knol and colleagues (see Figure 2.4 for this seven-step protocol). This flexible seven-step procedure towards organizing expert elicitation is based on existing protocols (Knol et al. 2010).

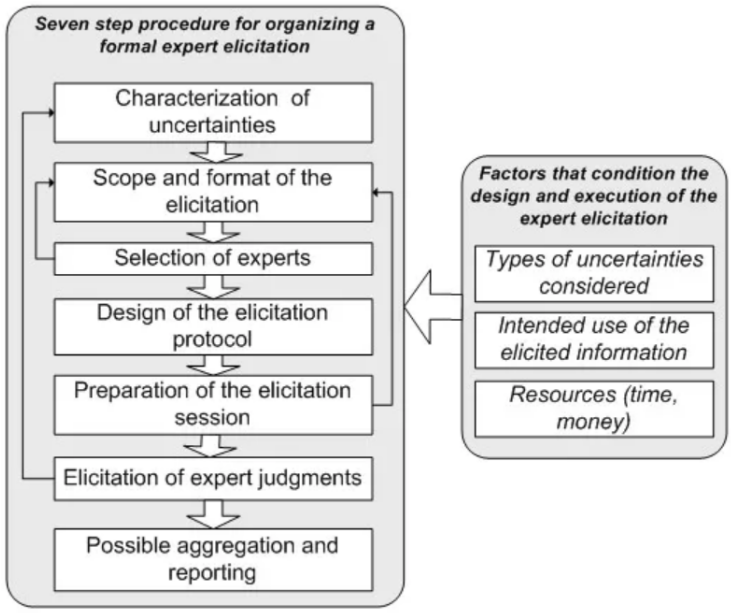


Figure 2.4: Seven step procedure for a formal expert elicitation (Knol et al. 2010)

Step 1:Characterization of Uncertainties. The uncertainty to addresses is the value of the target property: SOC in geographical areas which are unrepresented in feature space. This uncertainty (SOC-Values) can be expressed in statistical terms such as a (subjective) credibility interval (Garthwaite et al. 2005) (Knol et al. 2010).

The geographical areas which are unrepresented in feature space are brought back to actual geographical areas on Earth after research question 1. As will be shown in Section 3.1 and shown in Figure 3.2, WoSIS has insufficient data about SOC for the Tarim and Qaidam Basin in Western China, because the distance in feature space to training profiles is relatively large in these areas. The Tarim and Qaidam Basin in Western China were therefore chosen as study areas for the rest of this thesis.

Step 2: Scope and format of the elicitation. At the start of expert elicitation the format of face-to-face interviews was preferred by all involved parties (supervisors and student). There was no absolute threshold set on the number of experts to be invited.

The face to face interviews were held individual in an attempt to reduce as much as possible external influence on the experts' judgment. A potential downside of a group interview is the implicit suggestion of achieving consensus, where dominant opinions influence other opinions (Knol et al.

2010). The individual interviews also allowed for more targeted questions and explanations as well as increasing the feeling of responsibility of the experts to provide informed judgments(Knol et al. 2010). The experts were given the location of the study areas in Google Earth as well as a fact-sheet of the key soil forming factors in the study areas (see Appendix .3 and .4).

Step 3: Selection of experts. Which experts take part in an elicitation can greatly affect its outcomes and therefore the acceptance in the scientific community (Knol et al. 2010). The selection of experts, therefore, requires careful consideration. Two soil expert from ISRIC, both respected as authorities in their field of expertise, were chosen.

- Expert A, Sustainable land management and soil classification expert
- Expert B, Senior researcher, soil and land degradation assessment and restoration

Step 4: Design of the elicitation protocol The elicitation protocol contains the questions to be asked during the elicitation and the desired format for the answers (Knol et al. 2010). The questions for the soil experts about the quantity of soil organic carbon is shown in Table 2.1. The experts were asked to provide SOC estimations directly. These estimations were expressed in probabilistic terms (min, max, and most likely values). To quantify their uncertainty they were asked to give a 90% credibility interval. The width of this interval should indicate their uncertainty. Qualitative uncertainty words such as "likely" and "unlikely" were on purpose not used to avoid linguistic ambiguity which leads to confusion (Tversky and Kahneman 1974). The soil layer depth intervals where on purpose left open for the expert to fill in. Their SOC estimate could be the same for a very long soil column. Although SoilGrids makes map layers for different soil depth intervals (0-5cm, 5-15cm, 15-30cm, 30-60cm,60-100cm, and 100-200cm) the model is suited to use training data ranging over two or more depth intervals.

Table 2.1: Questions for the soil experts.

1	Is the value of Soil Organic Carbon equal for all soil depths from 0 to 200 cm below the surface in the Tarim Basin? If no which soil layers/depths can be roughly distinguished? i = a,b,c. . . . .
2	Could you give me an estimate of Soil Organic Carbon in the Tarim Basin for depth i?
3	Can you give a range for Soil Organic Carbon for depth i (upper boundary and lower boundary) where the probability that Soil Organic Carbon is in that range is 90%, So there is a 5% chance the true value of Soil Organic Carbon is above the upper boundary of the range and there is a % chance the true value of Soil Organic Carbon is below the lower boundary of the range
4	Is the value of Soil Organic Carbon equal for all soil depths from 0 to 200 cm below the surface in the Qaidam Basin? If no which soil layers/depths can be roughly distinguished? i = a,b,c. . . . .
5	Could you give me an estimate of Soil Organic Carbon in the Qaidam Basin for depth i?
6	Can you give a range for Soil Organic Carbon for depth i (upper boundary and lower boundary) where the probability that Soil Organic Carbon is in that range is 90%, So there is a 5% chance the true value of Soil Organic Carbon is above the upper boundary of the range and there is a % chance the true value of Soil Organic Carbon is below the lower boundary of the range

Step 5:Preparation of the elicitation session. Before the interview experts were via an email provided with the target soil property (SOC) and the study areas. It appeared more reliable to explain everything in further detail during the interview and give them the fact-sheet of key soil forming factors in the study area during the (first) meeting.The experts requested some time to read the full document therefore, a second meeting was scheduled. So the experts could read this document in the time between the two meetings.

Step 6:Elicitation of expert judgements. The goal of the expert elicitation was told to the experts. Their quantitative estimates of SOC were used in the execution phase of the thesis when the SoilGrid models were run with their estimates. By enlarging the training dataset with synthetic profiles for the study areas with SOC information, the expectation is that better SOC predictions are made for these areas.

During face-to-face interviews, both soil experts were eager to help and expressed their difficulties with formulating their estimation. They both felt that they did not have sufficient data about SOC in the study areas, both had therefore their approach and tools beside the factsheet provided to them to help them answer the questions.

Expert A felt somewhat uncomfortable making a general estimation for such a large area. This expert stated that there is always more variation in soils than you beforehand see, this also applies for

desert-like areas. For both basins the expert kept in mind a typical desert floor in the middle of the basin. The estimations of this expert can be found in Table ?? and 3.2 in the Results chapter. The estimations were essential for the creation of synthetic profiles.

Expert B could not provide estimates without actual soil data. This expert explained that soil data in China was not in abundance especially with open access. The expert gave solely advice on how to find soil research done in the Tarim and Qaidam basin. Unfortunately we could not use the information which was provided by this expert.

Step 7:Possible aggregation and reporting. The quantitative estimates are to be combined into one final estimate. Initially the weight of the individual estimates of the experts would be determined by their 90% credibility interval. The weight of their estimation would be based on their subjective judgment of their "level of certainty". It is very doubtful whether such estimates provide a good indication of the actual value of the elicited information, or merely introduce more bias (Knol et al. 2010). However with only one expert providing quantitative estimation there was no need for aggregation. The creation of synthetic profiles was only proceeded with the estimates of expert A.

2.4 Incorporating Synthetic Profiles in SoilGrids

The third and last research question is to incorporate synthetic profiles and recalculate the soil property SOC for the study areas by calibrating SoilGrids using both real SOC observations as well as synthetic SOC data and predict SOC in the selected areas. Since the prediction models are stored and can be rerun when new data become available, DSM is able to update maps. The model will be re-train and the feature selection and cross-validation will be repeated to get in the end a new model. The process is the same but only the model and predictions changes.

The number of synthetic profiles within the area can be changed by using a courser or finer grid of synthetic profiles, in this way distributing the synthetic profiles evenly over the area (Figure 2.5). Since SoilGrids is a computationally expensive model to run, only two different grid resolutions for synthetic profiles, grid resolutions 7 and 8 from the R package dggridR, were used (Barnes et al. 2017). The distance between the synthetic profiles was respectively 140 km and 80 km which resulted in 23 and 78 synthetic profiles for both study areas (see table 2.2. The coarser resolution (140 km) is the standard resolution that was used for incorporating synthetic profiles in the 2017 version of SoilGrids250m.

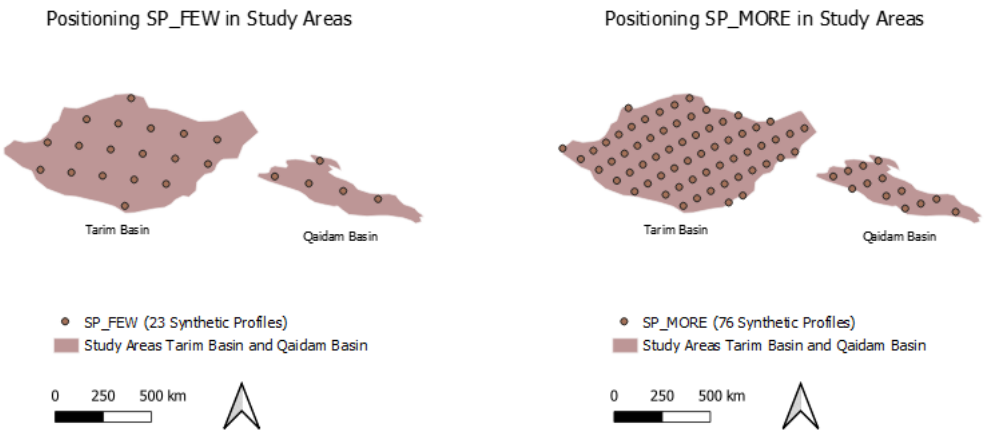


Figure 2.5: positioning of the Synthetic profiles (SP_FEW(23) and SP_MORE(76)) in study area: Tarim and Qaidam basin

Table 2.2: The different reruns of SoilGrids with and without Synthetic Profiles

Run Name	grid Resolution	Number of Synthetic Profiles
SP_NO	-	0
SP_FEW	140 km	23
SP_MORE	80 km	78

Chapter 3

Results

3.1 Areas Unrepresentative in Feature Space

In the methodology section it was explained how the geographical areas that are under-represented in feature space are determined. The different steps to reduce the computational burden and standardization of the data were all written in the programming language R and in further detail discussed in Appendix .2 where a general flowchart is shown. The computation time of one tile is between 7 and 14 minutes depending on the amount of ocean raster cells within a tile. Anunna, the HPC hosted by Wageningen University, made it possible to run the script for multiple tiles at the same time. The total computation took less than 1 day. Figure 3.1 shows the distances for Europe along with an elevation map to show the similarity in the geographical patterns.

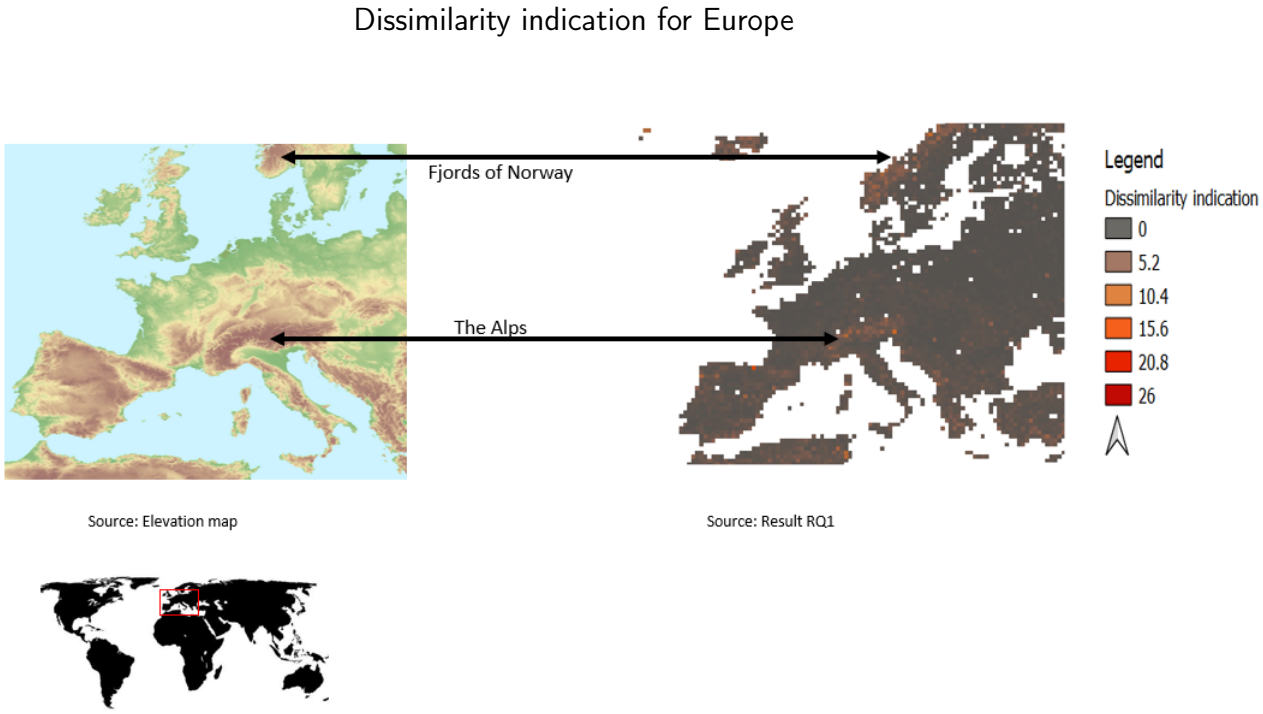


Figure 3.1: Elevation map of Europe and a map of the dissimilarity indication for the closest by training profile of SOC in feature space for Europe

Especially the mountainous areas like the Alps and Pyrenees light up and have a high distance in feature space towards the nearest soil training profile with information about the property Soil Organic Carbon. It makes sense that these areas have a high dissimilarity indication since they are less likely to be sampled since the are less suitable for agriculture and have a very particular vector in feature space.

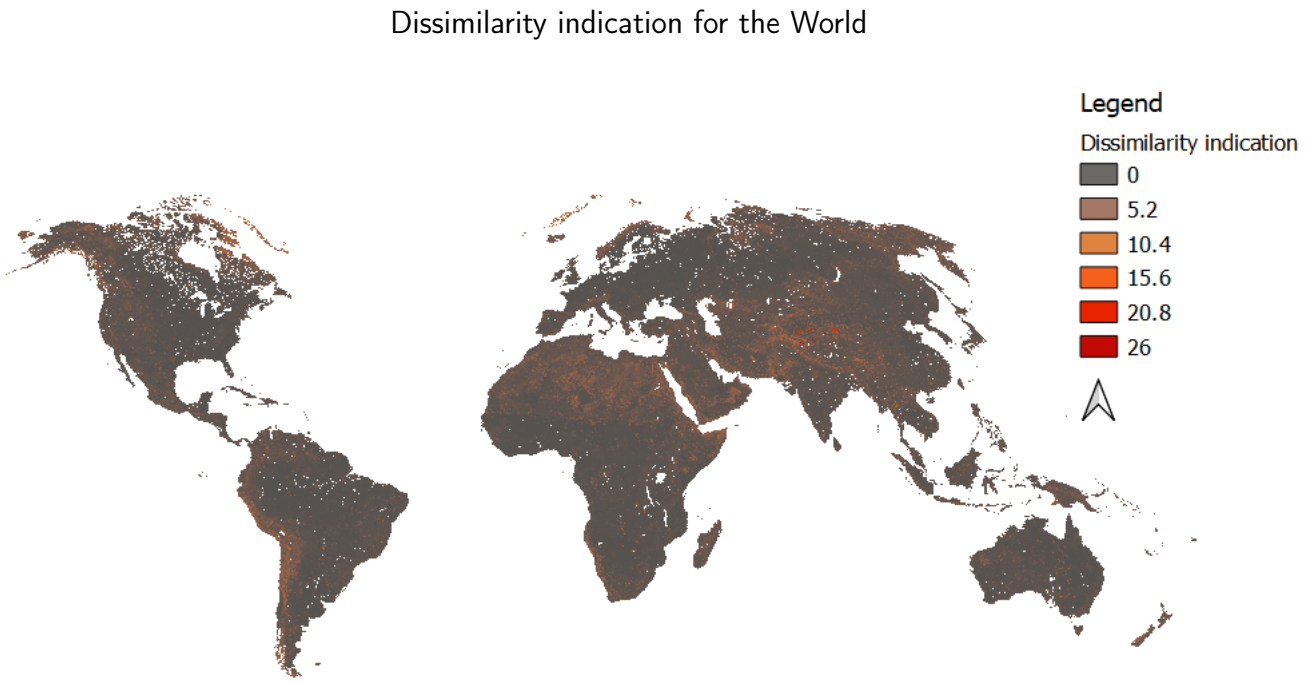


Figure 3.2: Global dissimilarity indication for the closest by training profile of SOC in feature space.

Figure 3.2 gives an overview of the results for research question 1 for the entire world. Note that the same patterns with regard to mountainous area can be seen. The Arctic areas, especially the coast line, has also some high dissimilarity indication.

Several areas with multiple adjacent raster cells that have a very peculiar vector and therefor have a high dissimilarity indications shows up in the global map. Three of these areas: Salar de Uyuni, the Sahara, and the Tarim and Qaidam basin will be highlighted in the following subsections.

3.1.1 Salar de Uyuni

In the Southern Altiplano Plateau, a high plateau in Bolivia, The Salar de Uyuni is located near the crest of the Andes. The Salar de Uyuni is characterized by a large salt lake in the basin center, it is the world's largest salt flat. The basin has experienced successive periods of lake expansion-contraction. Since the fluvial systems directed from mountains are feeding the basins and the same mountain ranges generate rain shadows which allow the lake to grow (Li 2014). While the basin has no drainage outlets the water has evaporated, high salinity levels caused a thick salt crust to form, leaving behind the impressive salt flat. This unique landscape is sometimes called the mirror of the earth due to the reflection of light when a thin sheet of water from the floods lay on top of it. The Salar de Uyuni is located 450 km to the south of Titicaca lake which is on the border of Bolivia and Peru, both are visible from space see Figure 3.3.

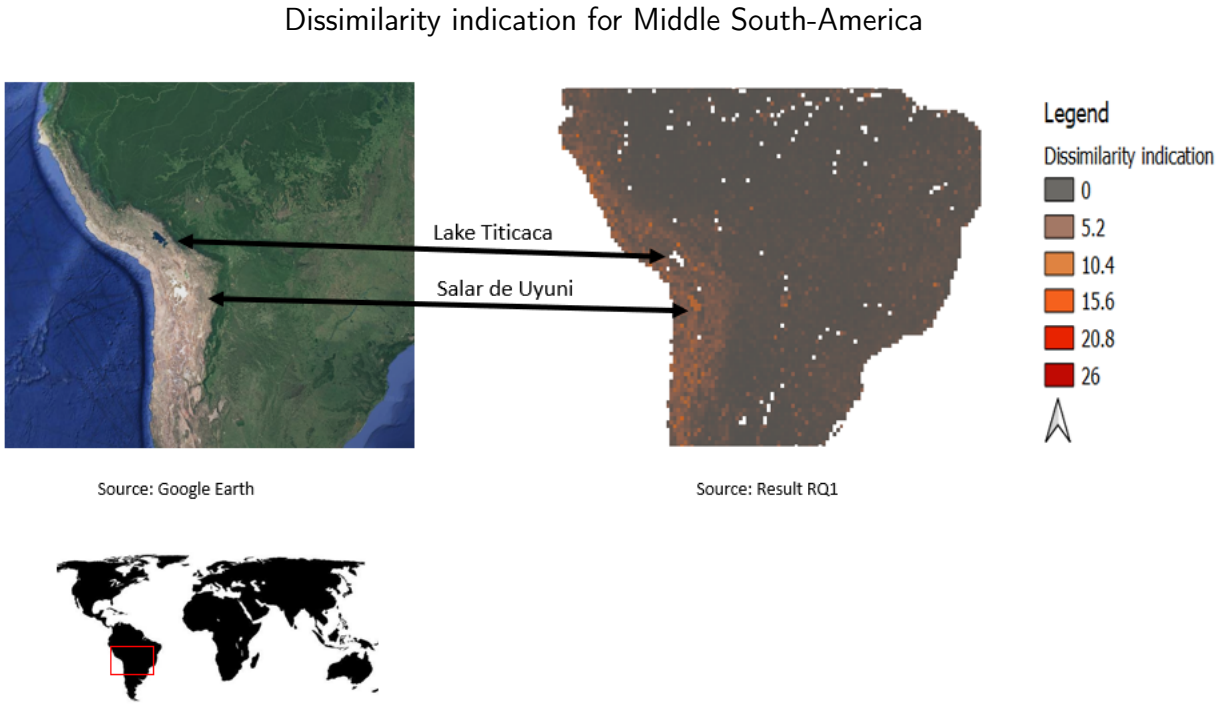


Figure 3.3: Satellite image and the dissimilarity indication to the nearest SOC Soil Trainign Profile for middle-South America. The Salar de Uyuni is the largest salt flat in the world, occupying 10 582 km^2. because of its size and distinct colour it is visible from space.

The Salar de Uyuni is within the World Soil mask but is eventually masked out in the final product of SoilGrids as a water body. This is clearly an area with a very particular vector with no soil training data profiles and shows therefore the result of calculating the dissimilarity indicator.

3.1.2 The Sahara

The Sahara is the largest hot desert in the world, and the third-largest desert overall after Antarctica and the Arctic. The Sahara is one of the harshest environments on Earth, covering nearly a third of the African continent. In this area very little rainfall and high temperature occur since the trade winds blow over land. The trade winds, blowing from higher latitudes, are very drying, and clouds are therefore almost absent in these desert regions (Walker 2000). The temperature can rise above 50 degrees in the shade. The Sahara is bordered by the Atlantic Ocean on the west, the Red Sea on the east, the Mediterranean Sea on the north and the Sahel Savannah on the south. It has a variety of land features: mountains, plateaus, sand- and gravel-covered plains, salt flats, basins, and depressions.

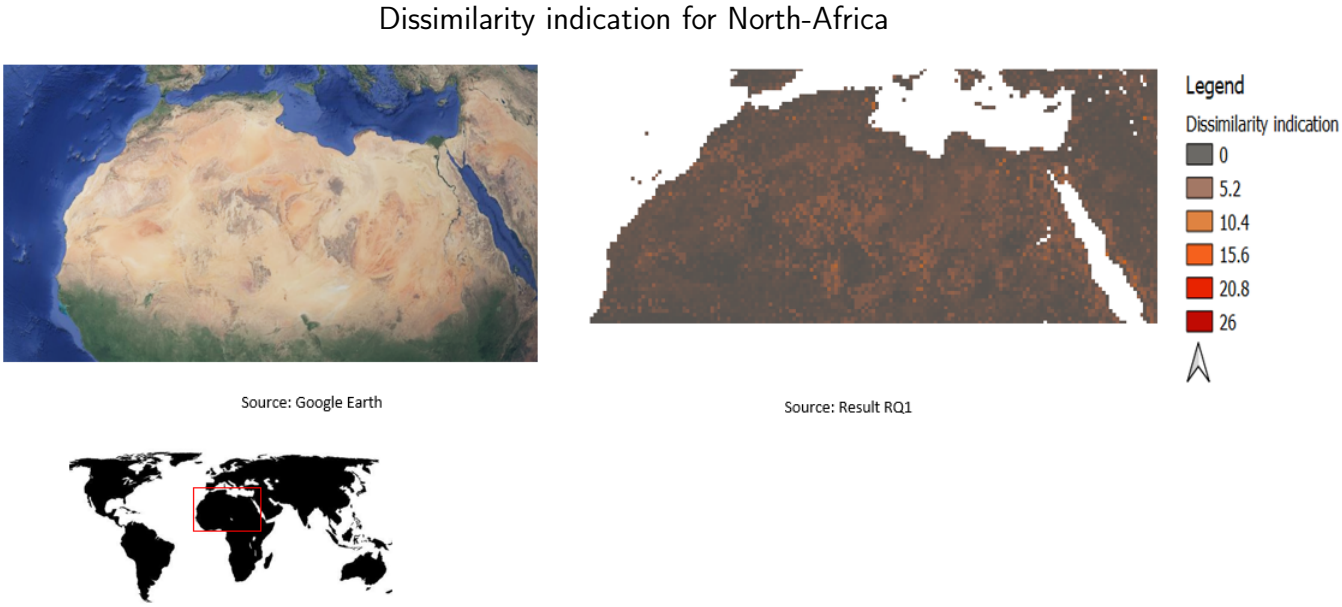


Figure 3.4: Satellite image and the dissimilarity indication for Northern Africa showing the Sahara which has an area of 9 200 000 km^2

The Sahara is less underrepresented as beforehand though. This is because it benefits from observations in similar environments in other parts of the world. However there are still some geographical patterns in the dissimilarity indication in the Sahara to be observed. These patterns coincide with the vast plateau like the Tassili N'Ajjer in Algeria and Aïr massif in northern Niger.

3.1.3 Tarim and Qaidam Basins

To the north of the Himalaya Mountains two areas have a very high dissimilarity indication (value 18). These areas are part of the Tarim and Qaidam basin which are surrounded by mountain ranges that already have quite some distance in feature space towards soil training profiles with SOC information see Figure 3.5.

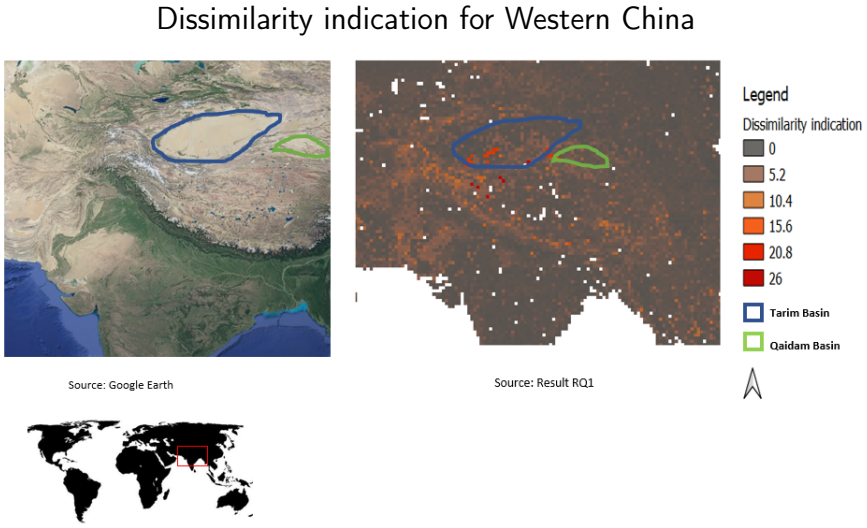


Figure 3.5: Satellite image and the dissimilarity indication for Western China, the Tarim (blue) and Qaidam (green) basin are delineated

What makes these areas so distinctly different in their environmental characteristics that their vector in feature space is so far away from SOC soil training profiles is a question still not completely answered. It could simply be an unusual combination of environmental characteristics that causes the high dissimilarity indication.

3.1.4 Study Areas

All the above-mentioned areas show up with their high dissimilarity indication suggesting they are being unrepresented in the feature space as to SOC information. The reason for this is because their vector is unique and can not be found somewhere else in the world where soil has been sampled. The areas have in common that the soil is quite bare and with variable elevation.
The decision to select the study area(s) is based on the result of research question 1. The Tarim and Qaidam basins were deliberately chosen to be the study areas. Because their dissimilarity indication was intriguingly large and these areas have a clear geographical cut with their surroundings see Figure 3.6.

1. **Tarim Basin**[a] is the larger from the two and is located in the province of Xinjiang in China. Xinjiang is divided into the Dzungarian Basin in the north and the Tarim Basin in the south separated by the Tian Shan mountain range. Much of the Tarim Basin is dominated by the Taklamakan Desert. Special interest in the south-western part of the Taklamakan Desert/Tarim basin. Where the Kunlun mountain range starts which is the northern edge of the Tibetan plateau.

2. **Qaidam Basin**[b] is located in the province of Qinghai. It seems the highest dissimilarity indication occurs in the Qaidam Basin which is a comparatively low area in the northeastern part of the Tibetan Plateau. The Altyn-Tagh is a mountain range between Kunlun mountains in the west and the Qilian Mountains in the east. The Altyn-tagh is the northern edge of the Qaidam Basin.

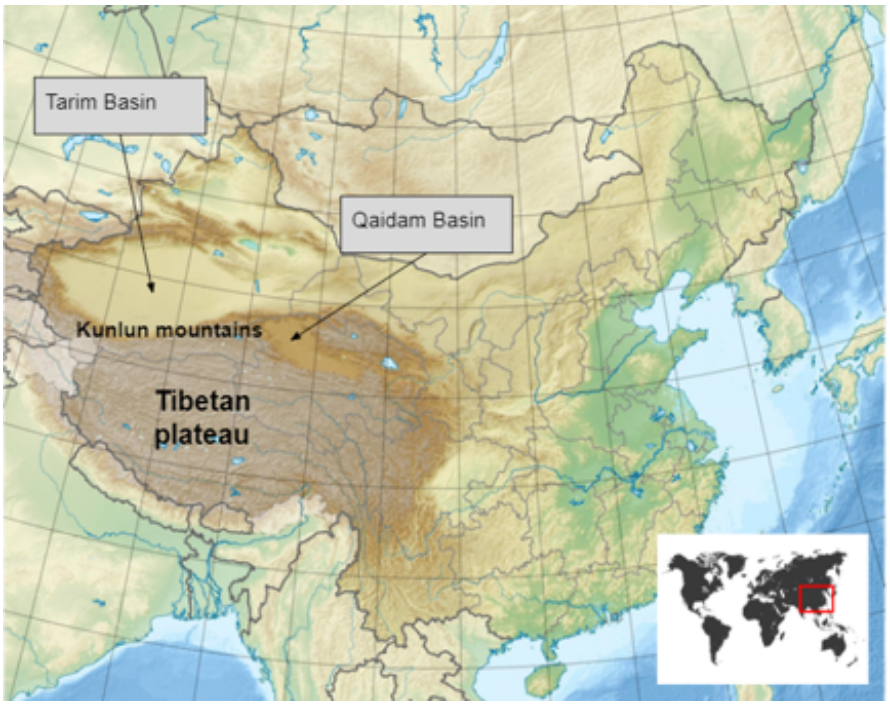


Figure 3.6: Location of the study areas in Western China surrounded with areas on a higher elevation.

[a] More information about the Tarim Basin can be found in Appendix .4
[b] More information about the Qaidam Basin can be found in Appendix .3

3.2 Generating Synthetic Profiles

As was stated in the methodology section, soil expert A gave his estimations for the Soil Organic Carbon in the Tarim and Qaidam basin regions in western China. Tables 3.1 and 3.2 show these educated estimates for the different depth intervals and the upper and lower limit of their 90% credibility intervals.

Table 3.1: Soil Organic Carbon in Tarim Basin

Tarim basin SOC			
Depth in cm	estimate(dg/kg)	Lower(dg/kg)	Upper(dg/kg)
0 - 20	30	10	130
20 - 40	15	10	80
40 - 60	10	4	20
60 - 100	5	1	10
100 - 200	0	0	8

Table 3.2: Soil Organic Carbon in Qaidam Basin

Qaidam basin SOC			
Depth in cm	estimate(dg/kg)	Lower(dg/kg)	Upper(dg/kg)
0 - 20	80	60	300
20 - 40	70	50	250
40 - 60	40	10	200
60 - 100	10	5	150
100 - 200	0	0	80

The estimated SOC value is for every depth interval never approximately in the middle of the 90% credibility interval. The probability distribution is therefore not a normal distribution or symmetric distribution. All the estimates of the expert are to the left of the middle of the 90% credibility interval, all estimates have therefore a right-skewed distribution. Synthetic profiles had to be generated from these expert estimates for SOC in the Tarim and Qaidam basin.

3.3 Soil Organic Carbon prediction values with Synthetic profiles

In Figure 3.7 the predictions of Soil Organic Carbon for the topsoil (0 to 5 cm) obtained with a SoilGrids model without synthetic Profiles is shown. These data are equivalent to those in the SoilGrids version 2019.

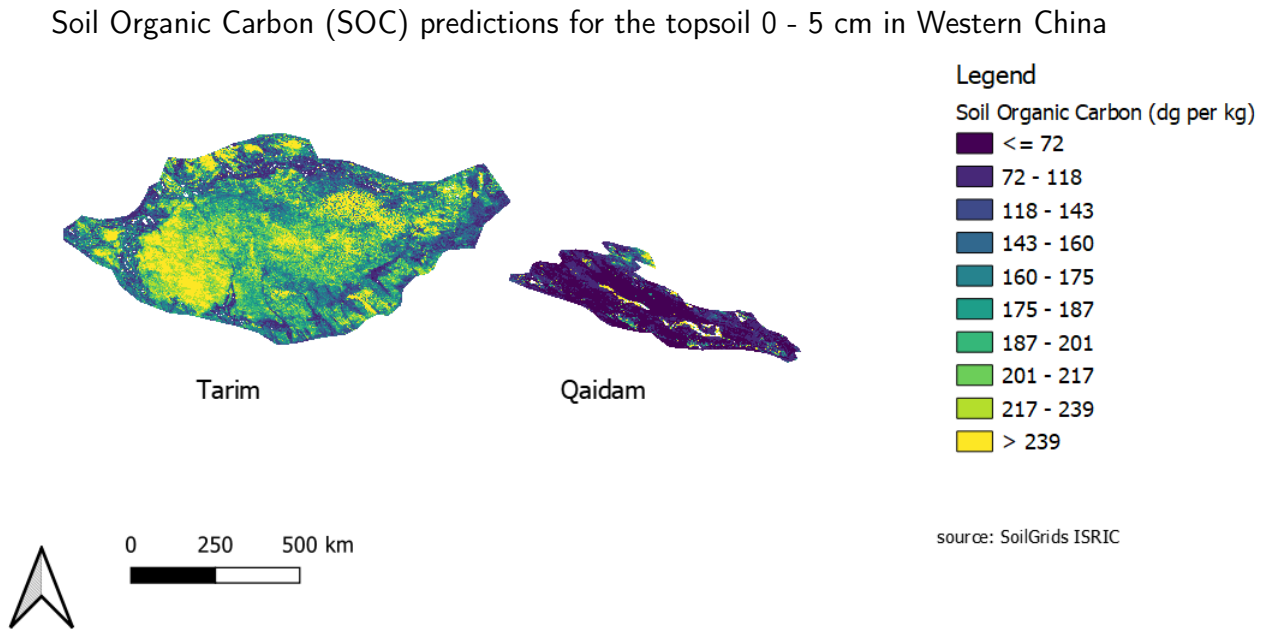


Figure 3.7: Soil Organic Carbon predictions for the two study areas Qaidam and Tarim Basin for the top soil 0-5 cm with SP_NO (0 Synthetic profiles in total)

The mode of the SOC predictions for the topsoil 0-5 cm in the two study areas Qaidam and Tarim Basin is around 180 dg per kg, see Figure 3.8. This is much higher than the expert estimates for Qaidam (30 dg/kg) or Tarim (80 dg/kg). It was therefore expected that the SOC predictions would become lower when incorporating synthetic profiles in the training data. In histogram 3.9 and 3.10 the histogram becomes less skewed to the right and more uniform with high peaks in certain SOC values.

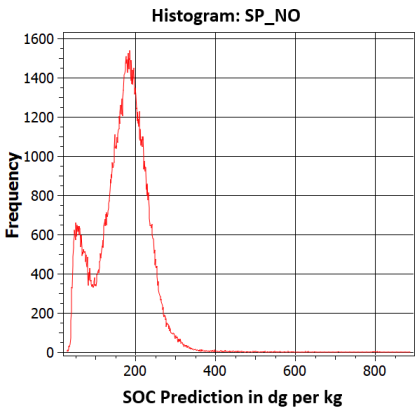


Figure 3.8: Histogram of Soil Organic Carbon predictions with SP_NO

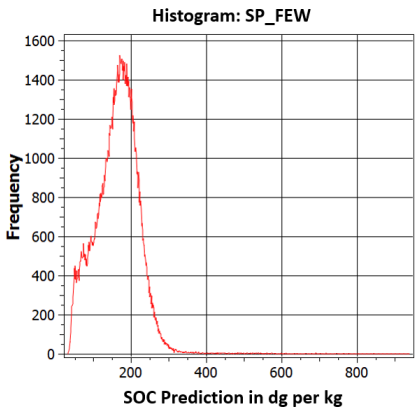


Figure 3.9: Histogram of Soil Organic Carbon predictions with SP_FEW

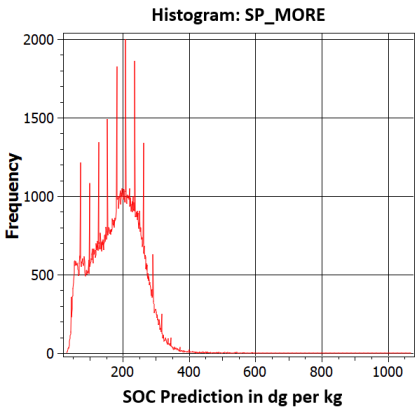


Figure 3.10: Histogram of Soil Organic Carbon predictions with SP_MORE

In Figure 3.11 the two predictions for SOC for the two study areas with respectively 23 and 76 synthetic profiles are shown next to the prediction with 0 synthetic profiles.

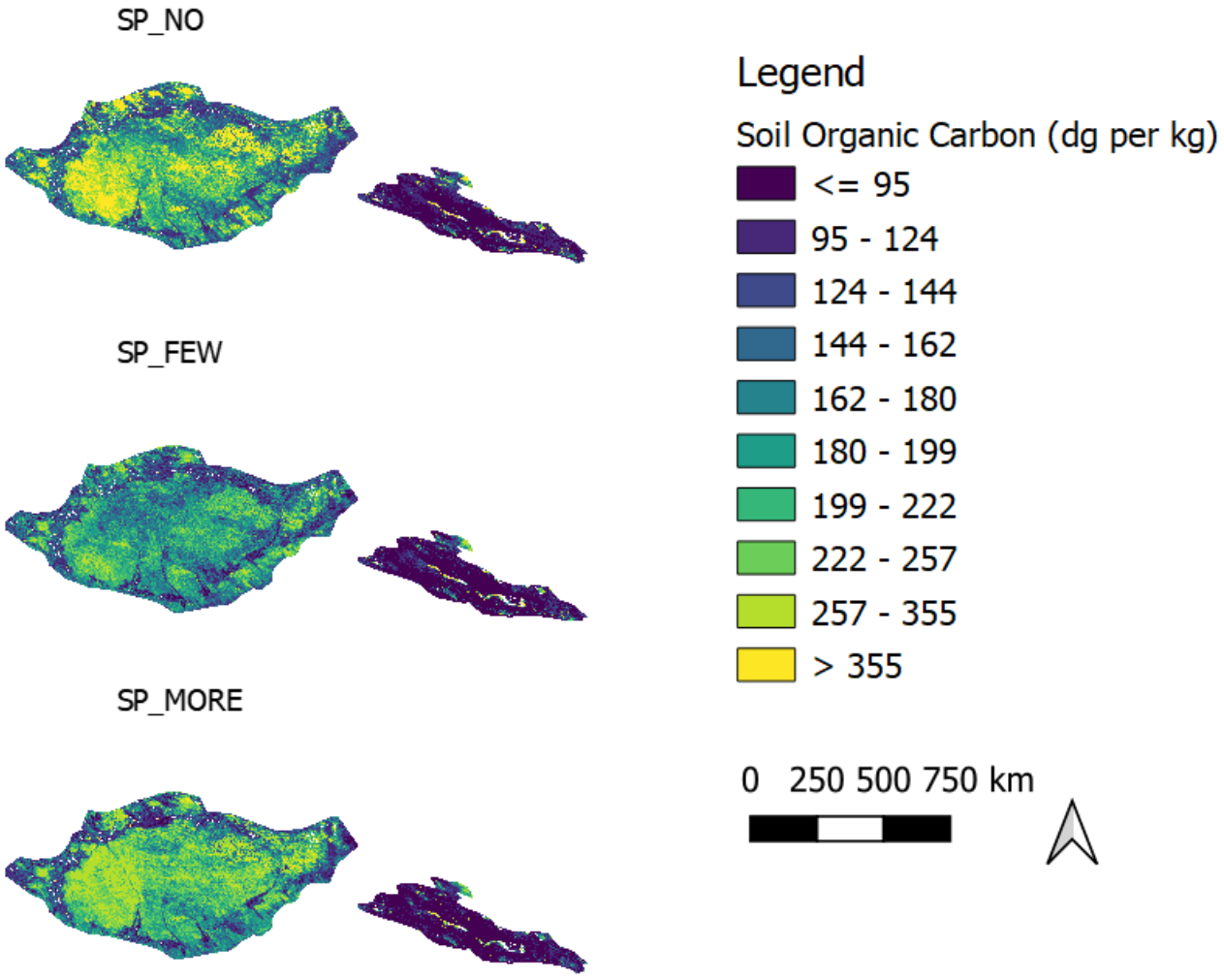


Figure 3.11: Soil Organic Carbon predictions for the two study areas Qaidam and Tarim Basin for the top soil 0-5 cm with SP_NO , SP_FEW and SP_MORE, respectively 0, 23 and 76 Synthetic Profiles

3.3.1 Predictions at different soil depths

The soil organic carbon reduces when going to deeper soil-layers this is also visible in Figure 3.12 and 3.13. This decrease in SOC for lower soil layers is true except for the 100-200 cm soil layer with 23 Synthetic Profiles. The model with no synthetic profiles (Figure 3.14) does not have such an artifact.

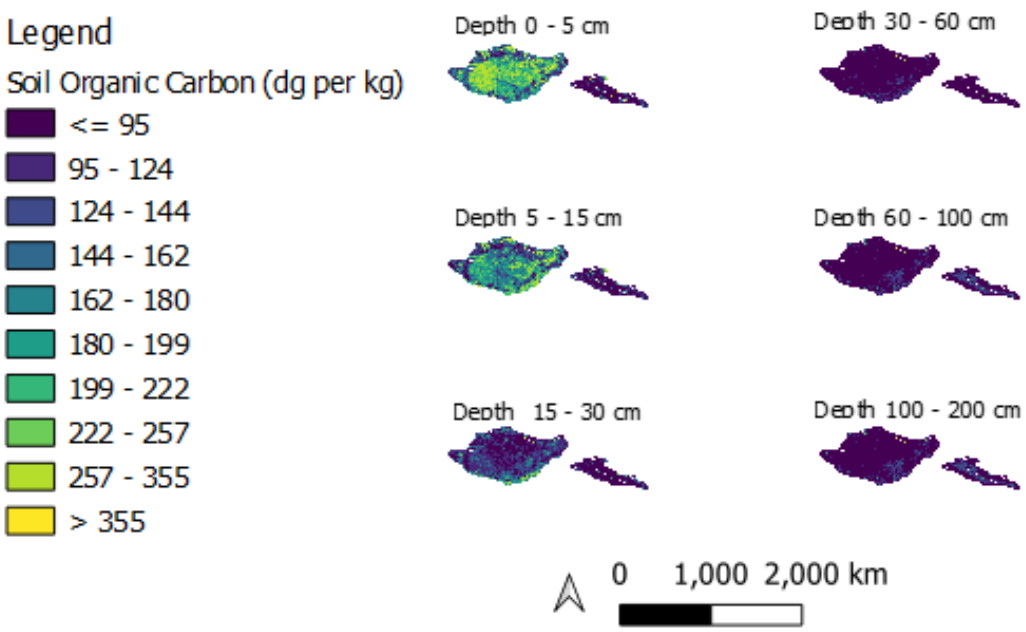


Figure 3.12: Soil Organic Carbon predictions for the two study areas Qaidam and Tarim Basin for the different soil depth intervals with SP_MORE (76 Synthetic profiles in total)

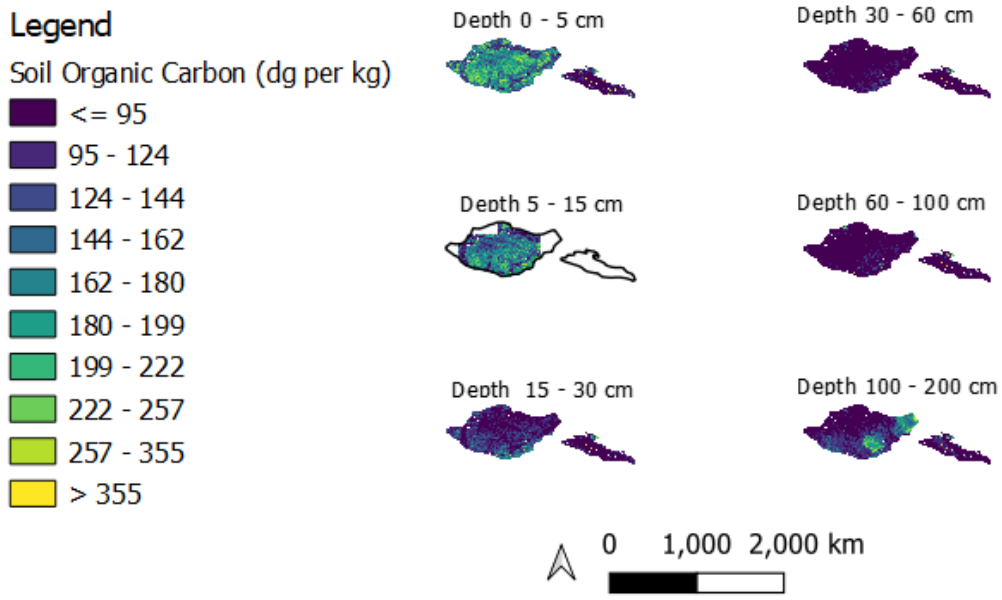


Figure 3.13: Soil Organic Carbon predictions for the two study areas Qaidam and Tarim Basin for the different soil depth intervals with SP_FEW (23 Synthetic profiles in total)

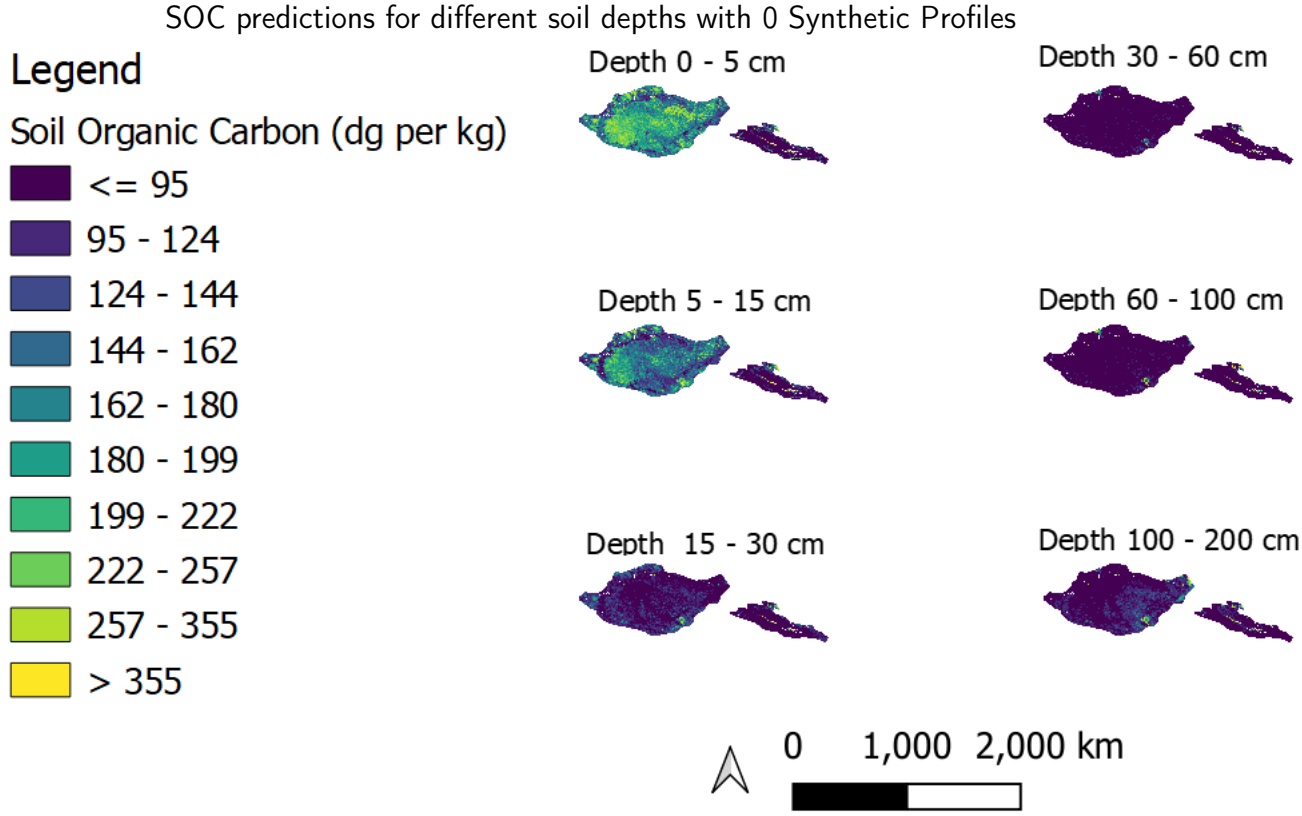


Figure 3.14: Soil Organic Carbon predictions for the two study areas Qaidam and Tarim Basin for the different soil depth intervals with SP_NO (without synthetic profiles)

3.3.2 Difference in Predictions with different synthetic profiles densities

An interesting question is whether the incorporation of synthetic profiles lead to lower predicted SOC values in the two study areas. The leftmost map of Figure 3.15 shows the difference between the prediction with 23 synthetic profiles and the prediction without synthetic profiles. A similar difference map is shown for prediction with 76 synthetic profiles (centre) and the difference between 76 and 23 synthetic profiles (right).

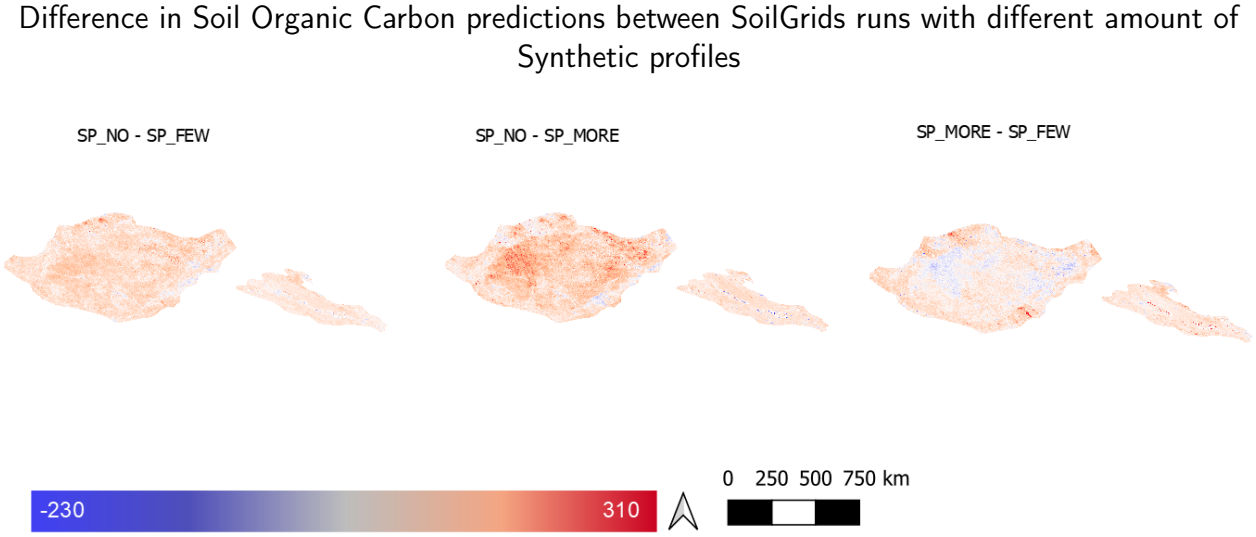


Figure 3.15: The Difference in Soil Organic Carbon predictions for the two study areas Qaidam and Tarim Basin with the different runs (SP_NO, SP_FEW and SP_MORE)

The SP_FEW decreases the prediction values for some small areas speckled on the edge of the Tarim basin and in the center of the Qaidam Basin. While the SP_MORE in general decrease the prediction values for the whole area. In the SP_MORE - SP_FEW it is shown that SP_MORE has higher prediction values for the center of the Tarim basin than the SP_FEW although there is a trend: more SP result in lower prediction values.

3.3.3 Smoothing effect

The synthetic profiles had the same SOC values for an area that was assumed to be homogeneous. Therefore the topological features like riverbeds and oasis get in a run with randomly positioned synthetic profiles a prediction that reflects more the surrounding area. This smoothing effect is visible in Figure 3.16. The geographical features like elevation get higher SOC predictions. The mechanism behind this is that by placing synthetic profiles randomly in an area, the synthetic profiles can be located on specific topographic or geological features. Therefore, they receive the expert estimates for a more general representation of the area. Since a denser grid of synthetic profiles will better represent the spatial variability in the study area, the smoothing effect will increase. A dense grid will cause a more homogeneous prediction of the study area and therefore a removal of topological and geological features in the prediction values.

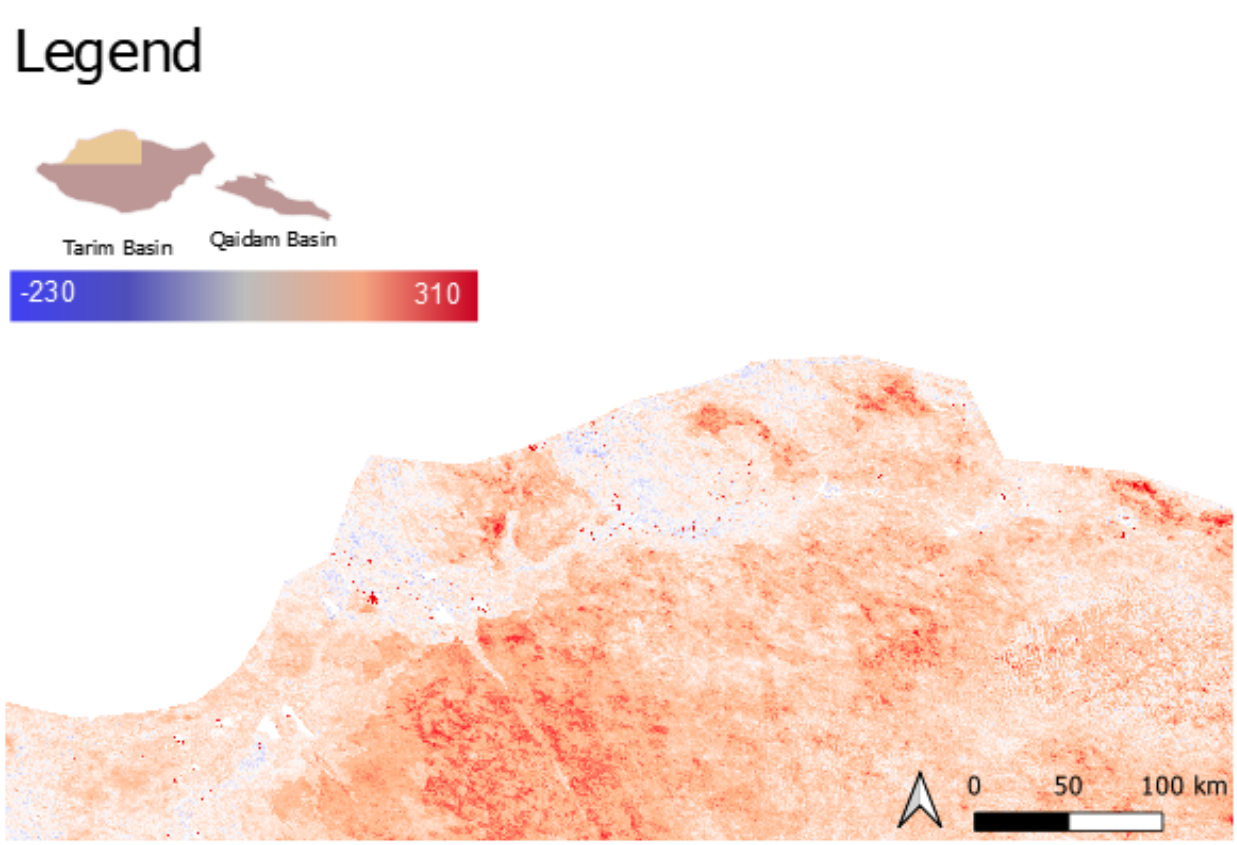


Figure 3.16: The Difference in Soil Organic Carbon predictions between 76 Synthetic profiles and 0 for an area in the North West part of the Tarim Basin.

Chapter 4

Discussion

This research aimed to find geographical areas in the world that are less represented in feature space than other areas and describe the effect of incorporating synthetic profiles on SOC predictions made in these areas. Although this research had some interesting findings,in this chapter the results are discussed and limitations regarding the analysis and setup are brought forward.

4.1 Distance Function

To answer research question 1 (see Figure 3.2) a distance function for distance in feature space was needed to find the locations on earth that are very dissimilar to all soil training profiles. This so-called distance function provides the dissimilarity indication for the prediction points (raster-cells).

The Euclidean distance has been widely used in scientific research since it provides a simple and mathematically convenient metric. In Figure 1.3, in the introduction chapter, the dissimilarity between two vector is simply the length of a straight line between two points in a Euclidean space. A small distance indicates a low degree of dissimilarity and a large distance indicates a high degree of dissimilarity. Euclidean space is easy to conceptualize but there are three reasons why it is inappropriate for use with the SoilGrids feature space:

1. Euclidean distance is very sensitive to the scales of the variables involved. It only makes sense to use the Euclidean distance when all the dimensions have the same units (Dwinnell 2006).
2. Euclidean distance is blind to the correlation of the different predictor features (covariates) (Dwinnell 2006). In the feature space of SoilGrids the environmental covariates or dimensions are not independent (or uncorrelated), but Euclidean metrics assume (at-least) non-correlation.
3. Calculating the Euclidean distance between vectors within a high dimensional feature space becomes nearly impossible since a curious phenomenon arises. The ratio between the nearest and farthest points approaches 1, i.e. the points essentially become uniformly distant from each other. In other words, the contrast in distance to different vectors becomes non-existent (Aggarwal et al. 2001).

The covariates used in SoilGrids are on different scales. Therefore, the values of the covariates were standardized before calculating the distance in feature space. The 77 most uncorrelated covariates from the more than 150 covariates were selected to reduce the correlation between predictor features

and at the same time reduce the number of dimensions. The problem concerning the high remaining dimensionality was partly solved by not using the straight-line (Eucli dean distance) between two vectors, but rather use the easy to calculate Manhattan distance (L1 distance metric)(Salzberg 1991). Which takes the sum of the absolute values of the differences in all dimensions (covariates).

The field of machine learning is continuously developing new distance metrics expressing the similarity between two elements (vectors). Therefore, there could potentially be a better distance function than Manhattan distance for measuring the distance of a prediction point and a soil training profile in the feature space of SoilGrids. However, to my knowledge Manhattan distance is the most appropriate distance function.

4.2 Data Reduction

Quartile Reduction method aims to increase the computation speed of calculating the dissimilarity indication. This was successfully done by reducing the number of training profiles from 235 340 to 148 086. This resulted in a reduction of calculating the distance in feature space to soil training profile for prediction points, the reduction was approximately 100 000 per prediction point. In the test runs with only one desktop computer this resulted in a computation time reduction of approximately 30-40 minutes to 7-14 minutes per tile. The processing time of calculating the dissimilarity indication for 4800 tiles on one computer went from roughly 35 minutes * 4800 tiles = 116 days to 10,5 minutes * 4800 = 35 days. Anunna, the HPC would divide the working load over multiple nodes and therefore decrease the computation time even more. Afterward with only a computation time of less than 24 hours it is debatable if the data reduction of the soil training profiles with the QR was necessary.

4.3 Synthetic profiles based on expert knowledge

In this thesis, expert knowledge was used as surrogate data to fill in gaps in the training dataset. The experts were asked for estimates for the whole Tarim and Qaidam basin while only a fraction of these areas are actually underrepresented in feature space (see Section 3.1.3). The area was enlarged to the whole Tarim basin to have a better understanding of the study area, information of the south west border of the Tarim basin was almost absent. The estimates of the experts have not been questioned much. This is a major pitfall in this research since the estimates of the expert might be contested for a number of reasons.

Overconfidence is a common bias seen in expert elicitation (Cruickshank 2018). The expert might, for instance, believe very strongly in a certain scientific hypothesis. Then he will give his estimation a narrow credibility distribution, since he is convinced about his estimation which reflects this certain scientific hypothesis (Garthwaite et al. 2005). Even if the rest of the scientific community is much more skeptical towards this scientific hypothesis (Knol et al. 2010). The expert hypothesis might not even be fully true resulting in incorrect estimations and model predictions.

Another concern in expert elicitation is satisficing. Satisficing may occur during a long elicitation and is caused by cognitive tiredness. Research has shown that satisficing is more likely to occur when there is an increase in the difficulty of the task or a reduction in the participant's ability and motivation to complete the task well (Visser et al. 2000). It is debatable whether the expert in our

research had the knowledge to answer all questions. Some questions may have been too specific and the expert may not have been able to give accurate estimations. Even-though the expert provides his estimations with subjective probability distributions this will not compensate for the lack of expertise from the expert side (Knol et al. 2010).The expert estimation which are might be far from the true value will propagate to the models predictions.

There is no absolute guideline on which to base the number of experts to be invited. In Step 3: Selection of experts two soil experts were selected, however only one of them provided SOC estimates which could be used to generate syntethic profiles. According to a panel of expert elicitation practitioners during a Resources for the Future Workshop in 2006, at least six experts should be included; otherwise, there may be questions about the robustness of the results (Cooke and Probst 2006) Although using the estimates from a single expert can be risky this was for the time being the only short time solution.

Not only is the usefulness of expert knowledge determined by the quality and quantity of the expert judgments. It is also bounded by the delicate positioning of the synthetic profiles (Truong 2014). Due to the smoothing effect of randomly positioning Synthetic profiles in a large area, incorporating more synthetic profiles is not wise. Hence it is questionable whether a regular grid is a smart choice, Since spatial position is as important as soil variable values. A possibility would be to let the expert position the synthetic profiles themselves.

4.4 SoilGrids with Synthetic profiles

The model of SoilGrids that predicts soil properties like SOC is approached as a black box. The synthetic profiles were incorporated in the training dataset of SoilGrids and the model was run to get new SOC predictions. The model select covariates upon which it based the predictions for a certain soil property. The covariates that were important in predicting the SOC-values in SoilGrids without Synethic profiles are listed in the Appendix (Appendix .1). With only 23 synthetic profiles the covariates selected by the model are the same as without synthetic profiles. However, with 76 synthetic profiles the number and mix of covariates selected changes markedly, just 76 profiles influence a global model trained with more than 200 000 profiles.

Chapter 5

Conclusions and recommendations

This master thesis studied the use of synthetic profiles for creating a digital soil map (SoilGrids) to improve the predictions for areas that are underrepresented in feature space. The current results are another step towards understanding the role of synthetic profiles in correct and comprehensive digital soil maps. In this chapter, first the research questions are answered and next ideas for further research are presented.

5.1 Under-represented regions

Geographical locations that are underrepresented in feature space for SOC in SoilGrids coincide with mountainous areas and the coastline of Artic areas. It is advisable to invest more in the acquisition of soil training profiles from these areas or incorporate synthetic profiles. Since statistical inference only applies to the sub-population which has been sampled and these areas have not been sampled or do not have a soil sample that represents the area via the homo-soil method.

5.2 Generating synthetic profiles

The creation of synthetic profiles with expert elicitation was more difficult than initially thought. Consulted scientists prefer to have actual data to base their statements on, which we can not provide them. The expectation on what a general soil expert knows about specific study area had to be drastically lowered. In the end one expert could provides me with SOC estimates upon which the synthetic profiles were based.

5.3 Effects of synthetic profiles

The added value of synthetic profiles is to have a higher or lower prediction value for a large (homogeneous) area which without synthetic profiles would have too low or high prediction values according to an expert judgement. The expert must be quite certain of his judgement and the difference between the expert estimate and the prediction value without synthetic profiles is sufficient. Even if these two circumstances are true, incorporating more and more synthetic profiles will invoke a smoothing effect when the area is heterogeneous. Since soil properties almost always experience some degree of

spatial variability within an area, it is not wise to incorporate too many synthetic profiles. Synthetic profiles are therefor not a panacea for underrepresented geographical areas.

5.4 Further research

This thesis provided a good base for future research. First of all further research can be done to identify the areas which are unrepresented in feature space for other soil properties such as pH or bulk-density or Sand content and alike. Secondly a research on whether the impact of Synthetic Profiles in the study area are higher due to the fact that these areas are unrepresented in feature space can be performed, since despite the small number of synthetic profiles added in the Tarim and Qaidam basin, they had a substantial effect on the predictions. The hypothesis is that the model is more robust in areas that are better sampled. Thirdly further research can be done with respect to generating synthetic profiles with expert elicitation, where the expert pinpoint the synthetic profiles them self. The latter only works when the experts have sufficient and up-to-date knowledge of the study area.

Another angle and a fourth option for research in the future is to reduce the geographical extent in which geographical areas that are underrepresented in feature space are search in. Since this thesis focus on finding these geographical areas on a global scale. The computation speed to calculate the dissimilarity indication for every prediction point in the world was increased by reducing the Soil Training profiles and by enlarging the cell resolution from 250 metres to 10 kilometres. The later causes that the result becomes far less visually attractive and useful. Research could focus on a far smaller extent than the whole globe to find geographical locations that are underrepresented in feature space with a cell resolution of 250 metres.

As discussed in section 4.2 the QR may not be necessary in reducing the computation time since a total computation time with Anunna of several days is reasonable in a master thesis. Recalculating the dissimilarity indication with 235 340 instead of 148 086 will result that some raster-cells have a lower dissimilarity indication. I assume the geographical patterns will stay more or less the same, however this is not yet been tested. When comparing the dissimilarity indication with all soil training profiles and with the soil training profiles after the QR method will tell us whether the QR method is an adequate method to get an even distribution of training profiles in feature space.

Glossary

R a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing (Hornik et al. 2002).. 21

References

Sanchez, P., Ahamed, Carre, F, A.E, H., Hempel, J, & Huising, J. (2009). Digital soil map of the world. *Environmental Sciennce*.

Panagos, P., Van Liedekerke, M., Jones, A., & Montanarella, L. (2012). European soil data centre: Response to european policy support and public data requirements. *Land use policy*, *29*(2), 329–338.

Guo, L., Zhang, H., Shi, T., Chen, Y., Jiang, Q., & Linderman, M. (2019). Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. *Geoderma*, *337*, 32–41.

Amrita, V. (2013). Soil analysis-determination of available organic carbon content in the soil.

Kern, J. S. (1994). Spatial patterns of soil organic carbon in the contiguous united states. *Soil Science Society of America Journal*, *58*(2), 439–455.

Yang, L., Jiao, Y., Fahmy, S., Zhu, A, Hann, S., Burt, J. E., Qi, F. Et al. (2011). Updating conventional soil maps through digital soil mapping. *Soil Science Society of America Journal*, *75*(3), 1044–1053.

Kempen, B., Brus, D. J., Stoorvogel, J. J., Heuvelink, G., & de Vries, F. (2012). Efficiency comparison of conventional and digital soil mapping for updating soil maps. *Soil Science Society of America Journal*, *76*(6), 2097–2115.

Hartemink, A. E., Hempel, J., Lagacherie, P., McBratney, A., McKenzie, N., MacMillan, R. A., Minasny, B., Montanarella, L., de Mendonça Santos, M. L., Sanchez, P. Et al. (2010). Globalsoilmap. net–a new digital soil map of the world, In *Digital soil mapping*. Springer.

Zhu, A.-X. (1997). A similarity model for representing soil spatial information. *Geoderma*, *77*(2-4), 217–242.

Lagacherie, P. (2008). Digital soil mapping: A state of the art, In *Digital soil mapping with limited data*. Springer.

Kienast-Brown, S., Libohova, Z, & Boettinger, J. *Ssm - ch. 5. digital soil mapping*. Natural Resources Conservation Service Soils. Retrieved February 11, 2020, from `https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/ref/?cid=nrcs142p2_054255`

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, *265*, 62–77.

Heuvelink, G., Brus, D., De Vries, F, Kempen, B, Knotters, M, Vasat, R, & Walvoort, D. (2010). Implications of digital soil mapping for soil information systems.

ISRIC. (2019). Soilgrids.

Batjes, N. H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., & de Jesus, J. M. (2017). Wosis: Providing standardised soil profile data for the world. *Earth System Science Data*, *9*(1), 1.

Hengl, T., de Jesus, J. M., Heuvelink, G. B., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B. Et al. (2017). Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.

Jenny, H. (1994). *Factors of soil formation: A system of quantitative pedology*. Courier Corporation.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Pennock, D., Yates, T., & Braidek, J. (2007). Soil sampling designs. *Soil sampling and methods of analysis*, 1–14.

Zhu, A., Liu, J, Du, F, Zhang, S., Qin, C., Burt, J, Behrens, T, & Scholten, T. (2015). Predictive soil mapping with limited sample data. *European Journal of Soil Science*, *66*(3), 535–547.

Hengl, T., & MacMillan, R. A. (2019). *Predictive soil mapping with r*. LULU PR. `https://soilmapper.org/`

QA, E. Et al. (2002). Guidance on choosing a sampling design for environmental data collection for use in developing a quality assurance project plan. *Washington, DC, USA, United States Environmental Protection Agency*, *166*.

Haakma, W., Bojke, L., Steuten, L. M. G., & IJzerman, M. J. (2011). Expert elicitation to populate early health economic models of medical diagnostic devices in development: Poster.

Knol, A. B., Slottje, P., van der Sluijs, J. P., & Lebret, E. (2010). The use of expert elicitation in environmental health impact assessment: A seven step procedure. *Environmental Health*, *9*(1), 19.

Mallavan, B., Minasny, B, & McBratney, A. (2010). Homosoil, a methodology for quantitative extrapolation of soil information across the globe, In *Digital soil mapping*. Springer.

Sample, J. T., & Ioup, E. (2010). *Tile-based geospatial information systems: Principles and practices*. Springer Science & Business Media.

Iman. (2016). Random sampling - tutorial 4 - latin hypercube sampling. Retrieved April 28, 2020, from `https://www.youtube.com/watch?v=H0RZ1uezuuw`

Cruickshank, C. (2018). Does the elicitation mode matter? comparing different methods for eliciting expert judgement.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Barnes, R., Sahr, K, Evenden, G, Johnson, A, & Warmerdam, F. (2017). Dggridr: Discrete global grids for r. *R package version*, *1*(1).

Li, J. (2014). Terminal fluvial systems in a semi-arid endorheic basin, salar de uyuni (bolivia).

Walker, A. S. (2000). *Deserts: Geology and resources*. US Department of the Interior, US Geological Survey.

Dwinnell, W. (2006). Mahalanobis distance.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space, In *International conference on database theory*. Springer.

Salzberg, S. (1991). Distance metrics for instance-based learning, In *International symposium on methodologies for intelligent systems*. Springer.

Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research.

Cooke, R. M., & Probst, K. N. (2006). *Highlights of the expert judgment policy symposium and technical workshop*. Resources for the Future Washington, DC.

Truong, N. (2014). *Expert knowledge in geostatistical inference and prediction*. Wageningen University.

Hornik, K. Et al. (2002). The r faq.

Bivand, R., Krug, R., Neteler, M., Jeworutzki, S., & Bivand, M. R. (2019). Package 'rgrass7'.

Hijmans, R. J., Van Etten, J., Cheng, J., Mattiuzzi, M., Sumner, M., Greenberg, J. A., Lamigueiro, O. P., Bevan, A., Racine, E. B., Shortridge, A. Et al. (2015). Package 'raster'. *R package*.

Appendices

.1 List of Covariates

CLM_WCL_W12SPD
ECO_USG_Z10
ECO_USG_Z11
ECO_USG_Z12
ECO_USG_Z13
ECO_USG_Z14
ECO_USG_Z15
ECO_USG_Z16
ECO_USG_Z17
ECO_USG_Z18
ECO_USG_Z19
ECO_USG_Z20
ECO_USG_Z21
ECO_USG_Z22
ECO_USG_Z23
ECO_USG_Z24
ECO_USG_Z25
ECO_USG_Z26
ECO_USG_Z27
ECO_USG_Z28
ECO_USG_Z29
ECO_USG_Z30
ECO_USG_Z31
ECO_USG_Z32
ECO_USG_Z33
ECO_USG_Z34
ECO_USG_Z35
ECO_USG_Z36
ECO_USG_Z37
ECO_USG_Z38
ECO_USG_Z39
LUC_ESA_L100
LUC_ESA_L110
LUC_ESA_L120
LUC_ESA_L121
LUC_ESA_L122
LUC_ESA_L130
LUC_ESA_L140
LUC_ESA_L150
LUC_ESA_L152
LUC_ESA_L153
LUC_ESA_L160
LUC_ESA_L170
LUC_ESA_L180

LUC_ESA_L190
LUC_ESA_L200
LUC_ESA_L201
LUC_ESA_L202
LUC_ESA_L220
LUC_GLC_C01t
LUC_GLC_C03t
LUC_GLC_C04t
LUC_GLC_C05t
LUC_GLC_C07t
LUC_GLC_C08t
LUC_GLC_C09t
MOR_ENV_DEMM
MOR_MRG_TPI
MOR_MRG_VDP
MOR_USG_F01
MOR_USG_F02
MOR_USG_F03
MOR_USG_F04
MOR_USG_F05
MOR_USG_F06
MOR_USG_F07
SAT_L07_B4NIR14
VEG_MOD_EVIMAX
VEG_MOD_EVIRNG
VEG_MOD_NPPY15
WTR_GIE_C00
WTR_GIE_C01
WTR_GIE_C02
WTR_GIE_C03
WTR_GSW_CHA
WTR_GSW_OCC
WTR_HYS_GTD

.2 Implementation of calculating distance in feature space in R

The free and open-source Geographic Information System (GIS) software Geographic Resources Analysis Support System (GRASS) is used by the SoilGrids team to store all geo-information of the covariates used in SoilGrids. In script R01_get_covariates the function initGRASS from the R-package: rgrass7 (Bivand et al. 2019) was used to connect to the GRASS-server from the R-server. The environmental covariates were stored in the mapset COVARIATES which covered the whole world. These were read as raster layers into Server.
In script R02_aggregate_and_stack the cell resolution in the extent of the tiles was set from 250 m to 10 km. This was done in a loop function for every covariate. Afterward each tile (250m and 10km) had to stack the 77 raster layers. This was done with the stack function from the raster R-package (Hijmans et al. 2015). In parallel with preparing the tile(s), the global descriptive statistics were calculated for the standardization, and also the training profiles were reduced.

The standardization process needed the global mean and standard deviation of each covariate. In script R03_get_global_stats the global minimum, maximum, mean, standard deviation, and the quartiles for each covariate was received from GRASS. These descriptive statistics were calculated with the r.quartile function from the rgrass7 package (Bivand et al. 2019) and saved in a data-frame called Global_Stats_Cov.

In script R03B_Mean_Stddev_Stack the global mean and standard deviation of each covariate was received from the Global_stats_Cov. The global mean and standard deviation of each covariate was transformed into a raster layer with the same dimensions as the tile with a cell resolution of 10 km. All the covariate mean and Standard Deviation (stddev) raster layers were then stacked in the same alphabetically order as the 4800 Tiles. The output: Mean_stack and stdev_stack were both used as input for the standardization in script R05_Standardization.
In section 2.2.3 the method to reduce the amount of training profiles is explained. The quantiles of each covariates were gained from Global_Stats_Cov. Solely the training profiles in TP_QR were further used to calculate the distance and therefore only these training profiles were standardized.
The script R05_Standardization, consists of two parts, standardizing the tile and standardizing the training profiles of TP_QR. In a loop-function, Mean_Stack and stddev_stack were both set to the extent of the tile in question.The standardization of the tile was done with the overlay function from the raster package. For standardizing the Training Profiles (TP) the sweep function was used twice: to subtract the mean and to divide with the stddev.

The input for the last step (R06_Distance_in_Feature_Space) was the standardized training profiles and the standardized tile(s). The calc function from the raster-package (Hijmans et al. 2015) was used to create a new raster (raster layer) object. The Manhattan equation in 2.2 was applied with the apply function to get the absolute distance in each dimension (covariate). The results were saved as tiff files.
All scripts and outputs on the left of the dotted orange center-line of Figure 1 are compressed to T01_Run_Tile. This script was run 4800 times by Anunna for each tile once. All computations on the right of the dotted orange center-line were computed only once since the global descriptive statistics remain the same for each tile as well as the standardized training profiles.

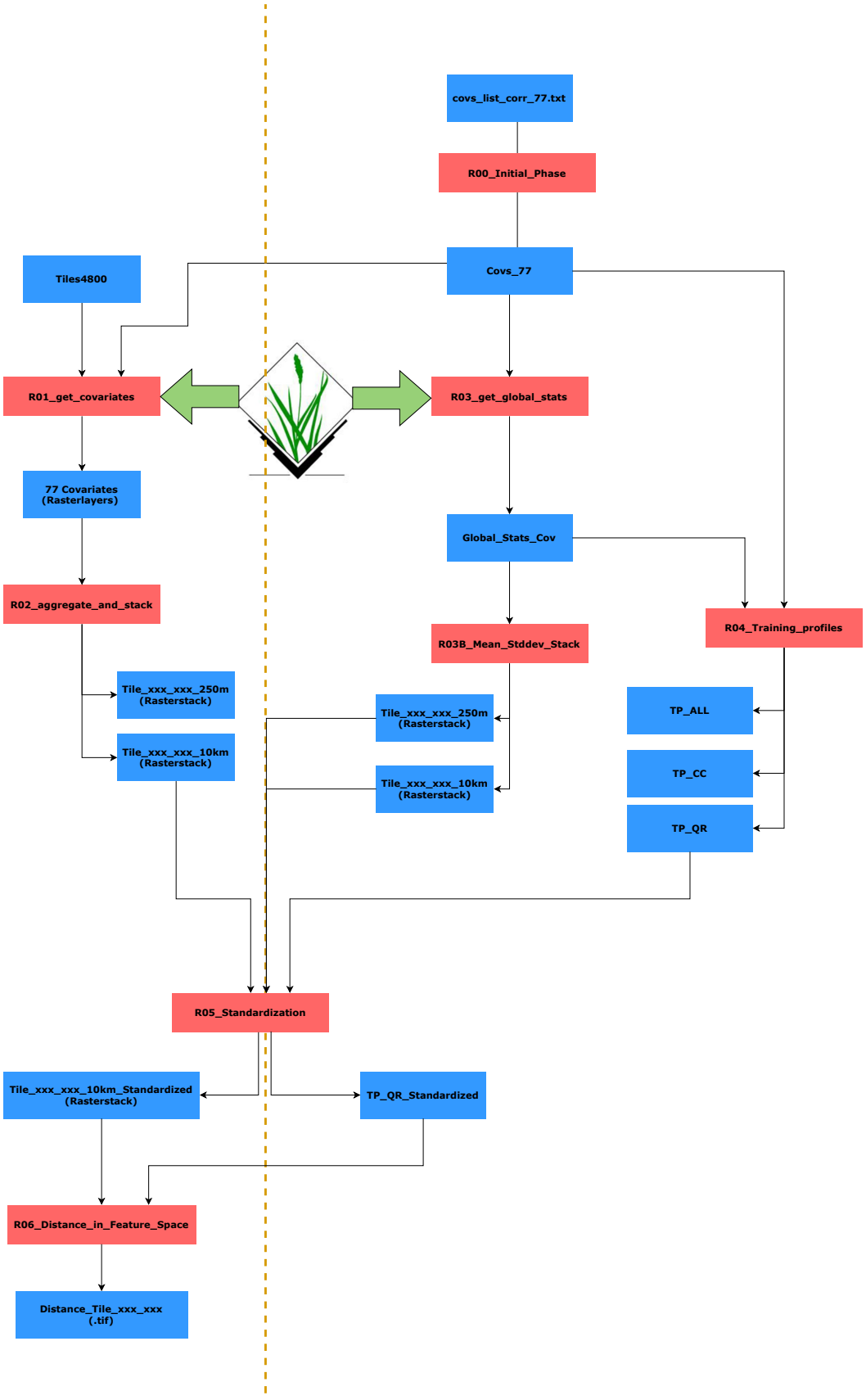


Figure 1: Scripts use for the preprocessing steps for calculating the distance in feature space between training profiles and raster cells in SoilGrids.

.3 Qaidam Basin

Table 1: Qaidam Basin Factsheet

Qaidam Basin	
Description	The Qaidam Basin is a large intermontane depression in Qinghai Province, China. Located in the northeastern part of the Tibetan Plateau, it is surrounded by the Qilian, Kunlun, and Aljun mountains which rise to more than 5000 m. The Qaidam Basin contains some of the richest salt resources in the world. These include deposits of halite and mirabilite, with commercial quantities.
Coordinates	between 90° 160E–99° 160E and 35° 000N–39° 200N in northwestern China
Area	120,000 km2 (700 km long and up to 300 km wide)
Climate	
Climate	Hyper-arid, among the most arid non-polar locations on earth, with some places reporting an aridity index of 0.008–0.04
Mean annual precipitation	20-100 mm Due to the shielding effect of the Tibetan Plateau and surrounding high mountain ranges, little moisture can reach in this basin. Mean annual precipitation ranging from 100 mm in the southeastern to less than 20 mm in the northwestern. Significant rainfall occurs mainly in summer, while precipitation is very low in winter and spring.
Mean annual evaporation	3000-3200 mm The potential mean annual evaporation can be 100 times higher than precipitation.
Wind direction	The Altyn Tagh Mountains are much lower in elevation (4000 m) than the other peripheral mountains (5000 m), i.e. the mountains tend to be taller and have fewer passes through which the wind can travel. The prevailing wind direction in the western basin is NW-SE and becomes nearly W-E in the eastern part see Figure 3. The Qaidam basin is also prone to heavy winds as well as sandstorms from February to April.
Temperature	Due to the high altitude, it has quite cold winters (harsh in the highest elevations), mild summers, and a large temperature difference between day and night. Its mean annual temperature is approximately 5 to 8 °C, with January temperatures ranging from 18 to 7 °C and July temperatures ranging from 15 to 21 °C. With an annual average temperature of 3.5 °C
Organisms	
Vegetation	The open vegetation consists of not too many kinds of plants, most of which are halophytes of highly drought-resisting shrubs, half shrubs, and herbs. Dense sedges form grass dunes along shores of salt marshes, salt lakes, and rivers. Reed and wild rye compose of the main vegetation in salt lakes and swamp periphery
Relief	
Elevation	2800 m / 3000 -3500 m / 2600 -3000 m

Relief	Regularly graduated from the edge to the center, the basin relief appears concentric rings of diluvial gravel fan (Gobi), alluvial-diluvial silty sandplain, lacustrine-alluvial silty clay plain, and lacustrine sludge solonchak plain. In the low-lying area are widely distributed many salt lakes and swamps.
Parent Material	
Soil	The landscape of Qaidam Basin features arid desert with the major soil types of Solonchaks and Gypsisols. Meadow soil and swampy soil are typically salinized, and gypsisols are mainly distributed in the western part of the basin.
Landcover	Qaidam forms an endorheic basin accumulating lakes with no outlet to the sea. Therefore one-fourth of the area which is covered by saline lakes and playas. Around one-third of the basin, about 35,000 km2 (14,000 sq mi), is desert.
Sediment	The basement outcrops emerge along the edge of the Qaidam Basin, consisting of Precambrian-Silurian metamorphic rocks. The overlain strata within the basin are Devonian-Cenozoic sediments that were originated from weathering and denudation of the surrounding mountains. The total thickness of Cenozoic sediments within the basin can reach up to about12,000 m. Research in northwestern Qaidam Basin indicates that very thick gypsum and rock salt sediments were deposited during the period from the Middle Oligocene to the Upper Pliocene. While the Neogene strata consist of mudstone, calcareous mudstones and marls, intercalating siltstone, gypsum, and rock salt beds, many of which are rich in carbonate. And the late Pliocene–Pleistocene sediments are characterized by clay, mudstone, thick halite and gypsum, and saline lacustrine deposits. Overall the sedimentary sequence of the core comprises clay, clay-silt, and siltstone, intercalated with salt layers (mainly halite), marl beds and thin or scattered gypsum crystals
Time	
Geology	More than 1 billion years ago, Qaidam Basin was an integral part of the North China geologic unit. At the beginning of the Eopaleozoic era (about 560 million years ago), it was separated and surrounded by shallow sea as a result of plate disintegration. At the end of the Eopaleozoic era (about 400 million years ago), the basin began to uplift due to the intense tectonic movement caused by plate subduction and collision and later became a land 200 million years ago. Beginning in the Neogene Period, Qaidam Basin was completely separated from the ancient Mediterranean Sea and became a typical inland basin as a result of the fast uplift of the Qinghai-Tibet Plateau. With the high mountains blocking the monsoon from the Indian Ocean, the Pacific Ocean, and the Mediterranean Sea, its ecological environment changed from forest-steppe to desert steppe. After the water evaporated, large amount of salts and rare metals converged and ultimately formed salt lakes see Figure 4

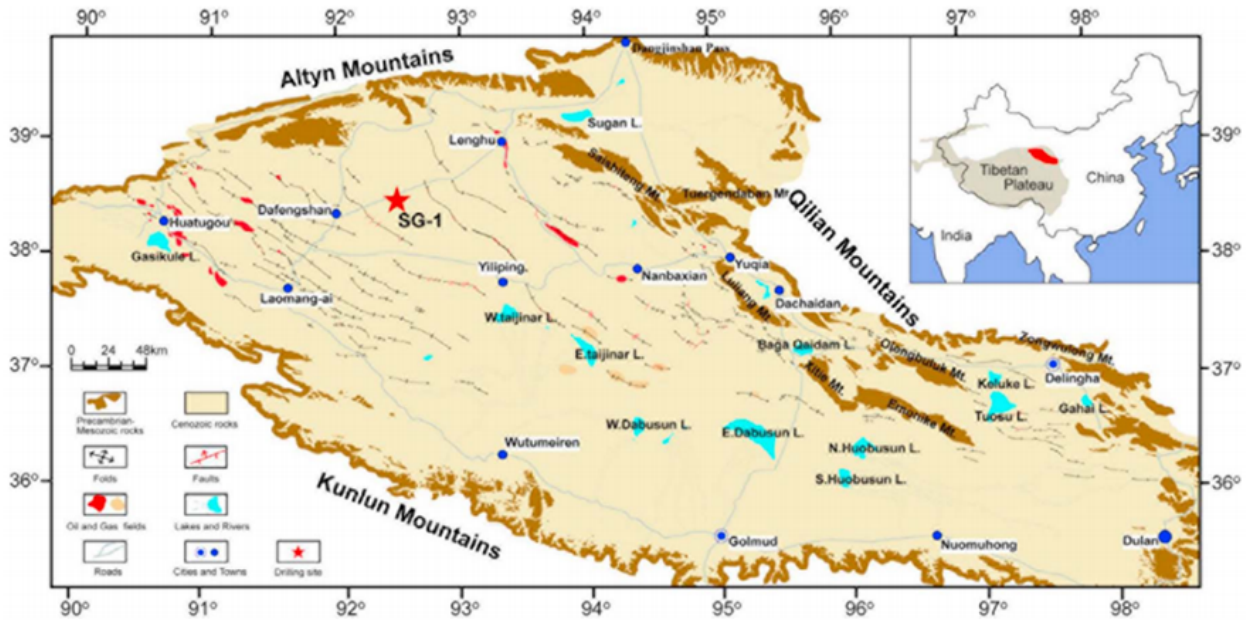


Figure 2: Map of Qaidam Basin and adjacent regions showing surrounding mountains and major structures.

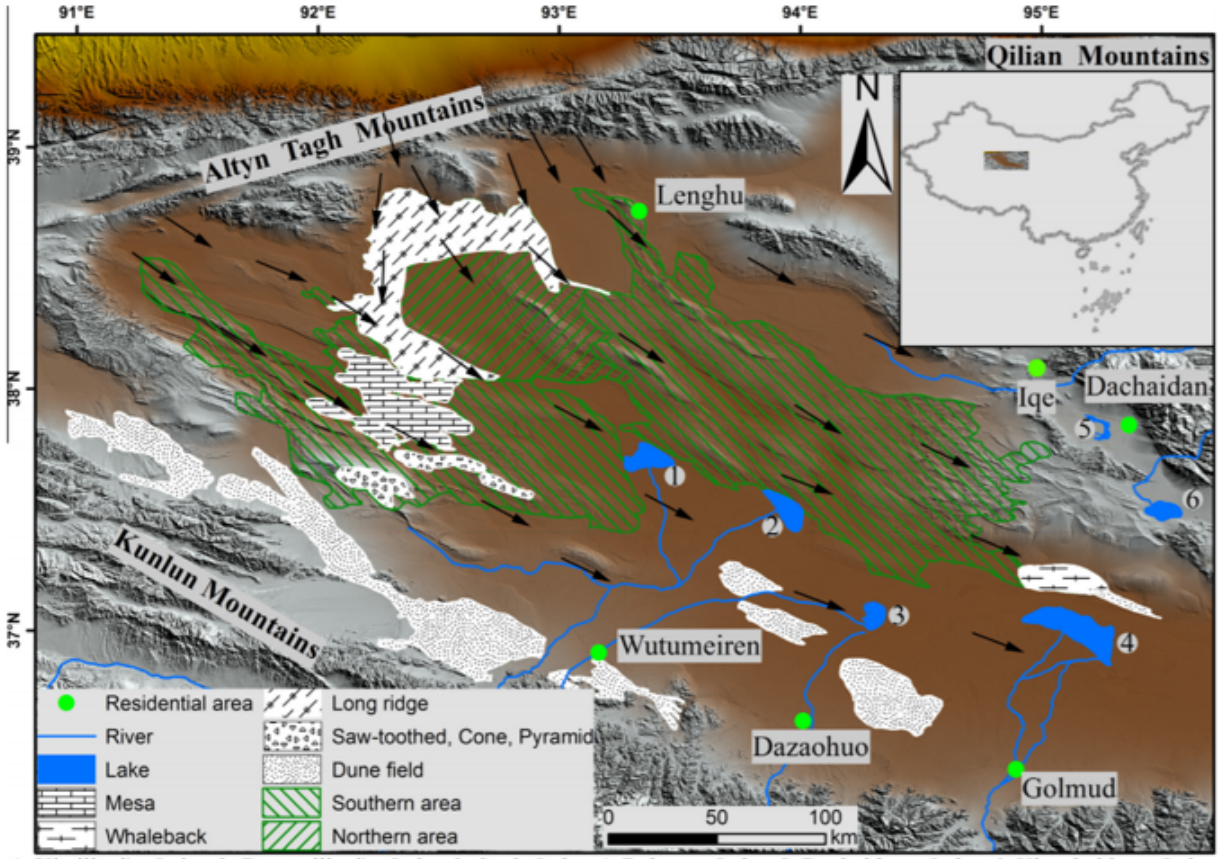


Figure 3: The location of the Qaidam Basin and distribution of the typical yardangs and dune fields within the basin. Black arrows represent the dominant wind direction.

.4 Tarim Basin

Table 2: Tarim Basin Factsheet

Southern margin Tarim Basin	
Description	The Tarim Basin is China's largest inland basin located in the south of Xinjiang Uyghur Autonomous Region between the Tianshan Mountains, Kunlun Mountains, and Arjin Mountain. Its southern boundary is the Kunlun Mountains on the edge of the Tibetan Plateau. a broad escarpment where the altitude drops from around 5,000 metres (16,000 ft) to 1,500 metres (4,900 ft) over a horizontal distance of less than 150 kilometres (93 mi). The Tarim basin is home to the Taklimakan Desert—the biggest, hottest, driest desert in China.
Coordinates	between 75°160E–85°160E and 35°000N–40°200N in northwestern China
Area	The total Tarim basin covers more than 1 million km^2.
Climate	
Climate	The region has a typical arid continental climate, a warm temperate arid desert climate with an extremely dry climate and frequent sandstorm.

Mean annual precipitation	The annual precipitation ranges from 200 to 500 mm in mountainous areas and approximately 50–80 mm in basin plain areas. The annual precipitation is 50–70 mm in the north, 15–30 mm in the south, and only 10 mm in the central desert area. Also, > 80% of the total annual precipitation occurs between May and October, and < 20% occurs between November and April
Mean annual evaporation	The annual evaporation reaches up to 2450.0 to 2902.2 mm.
Wind direction	Wind is common year-round, with prevailing northwesterly winds see Figure 6
Temperature	The annual average temperature is 10–12C with a maximum of 42.2 °C and a minimum value of 30.9 °C.
Hydrology	The Tarim River basin is a relatively closed hydrological system with very limited water resources due to the extremely arid climate. The dryness originating from the long distance to the oceans is greatly enhanced by the surrounding mountains, which block moisture from entering this inland basin. Rivers originating from the surrounding mountains, under the control of elevation, flow toward the basin and then principally eastwards, such as the Tarim River and the Kongque River. The Taklimakan Desert in the central region of the basin is the groundwater discharge region, while the Lop Nur Lake basin in the east is the runoff confluence center.
Organisms	
Vegetation	Runoff from the Tien Shan mountains and the Kunlun Shan mountains feed rivers, which in turn support vegetation. The river valley appears dry in January, but in August, it is lined with vegetation. In the desert-oasis ecotone, the vegetation is dominated by Alhagi sparsifolia Shap., with a coverage of about 38.9%.
Land-use	Farmland is the main land-use type, with cotton, maize, and wheat being the main crops. Farmland, particularly in the Tarim basin where water tables can be high, can become saline in a relatively short time due to the upward movement of salts in the soil solution.
Relief	
Elevation	The elevation is around 750 in the central desert and around 1300 -1400 on the edges see Figure 5
Parent Material	
Soil	The major soil types are brown desert, anthropogenic-alluvial, meadow, and aeolian sandy.
Landcover	At the basin center, there is the Taklamakan Desert, which covers 33.76 104 km2, and is also known as the "Sea of Death". Dunes cover about 85 percent of the Taklimakan desert, often feeding massive dust storms.
Sediment	The Tarim Basin is a large sedimentary basin formed by the superposition of many types of basins at different periods in the long geological evolution.
Time	

Geology	The basin foundation is Proterozoic metamorphic rocks with the sediments of the Sinian system, Paleozoic marine, Mesozoic, and Cenozoic continental deposits with the thickness up to 7,000-10,000 meters. The Tarim Basin is covered by desert, below which a thick upper Proterozoic and Phanerozoic sequence deposits. However, some Precambrian rock outcrops are available along its margins.

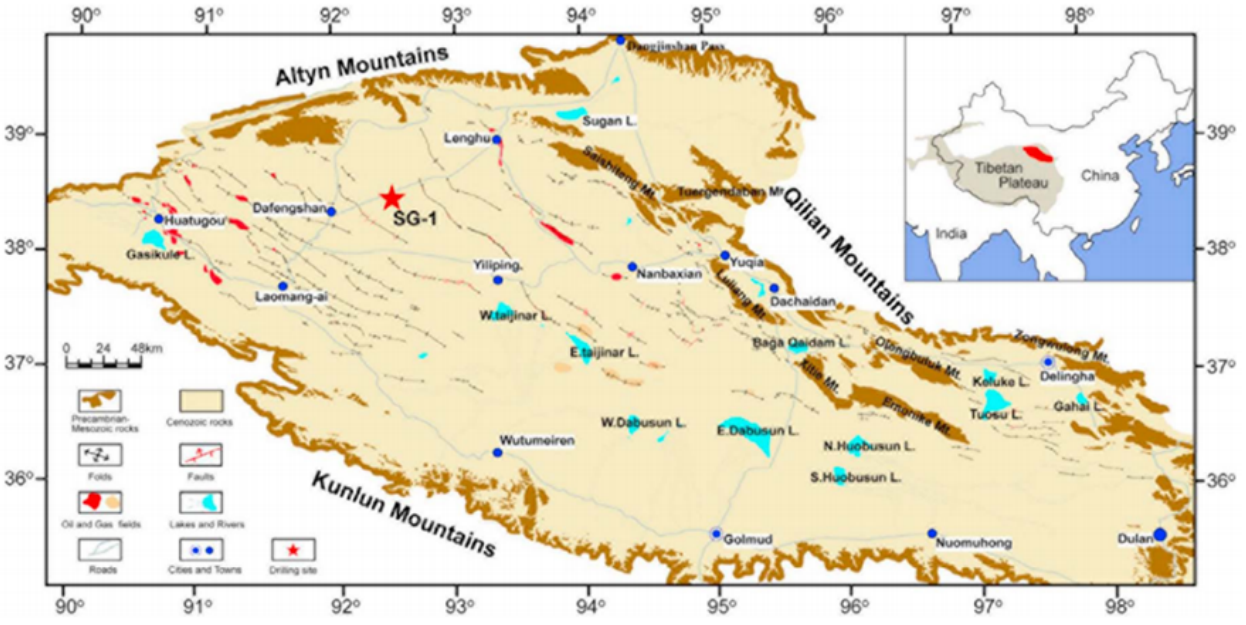


Figure 5: Elevation (in meters) for the Tarim Basin and the locations of meteorological and hydrological stations.

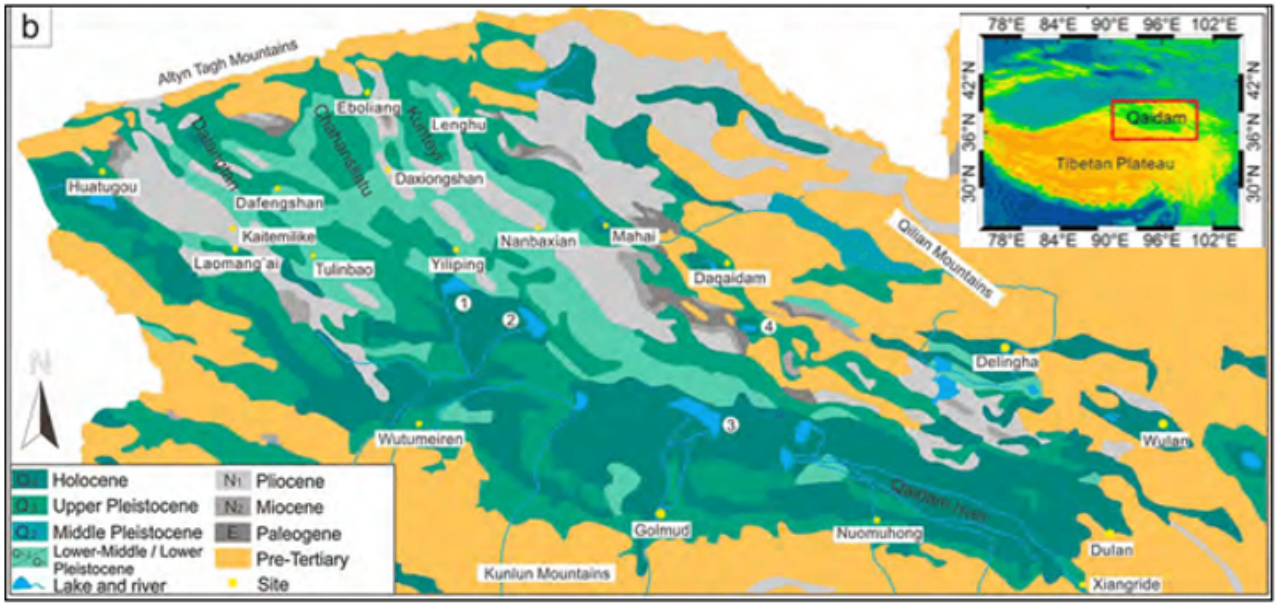


Figure 4: Simplified geological map of the Qaidam Basin (after Li et al., 2016).

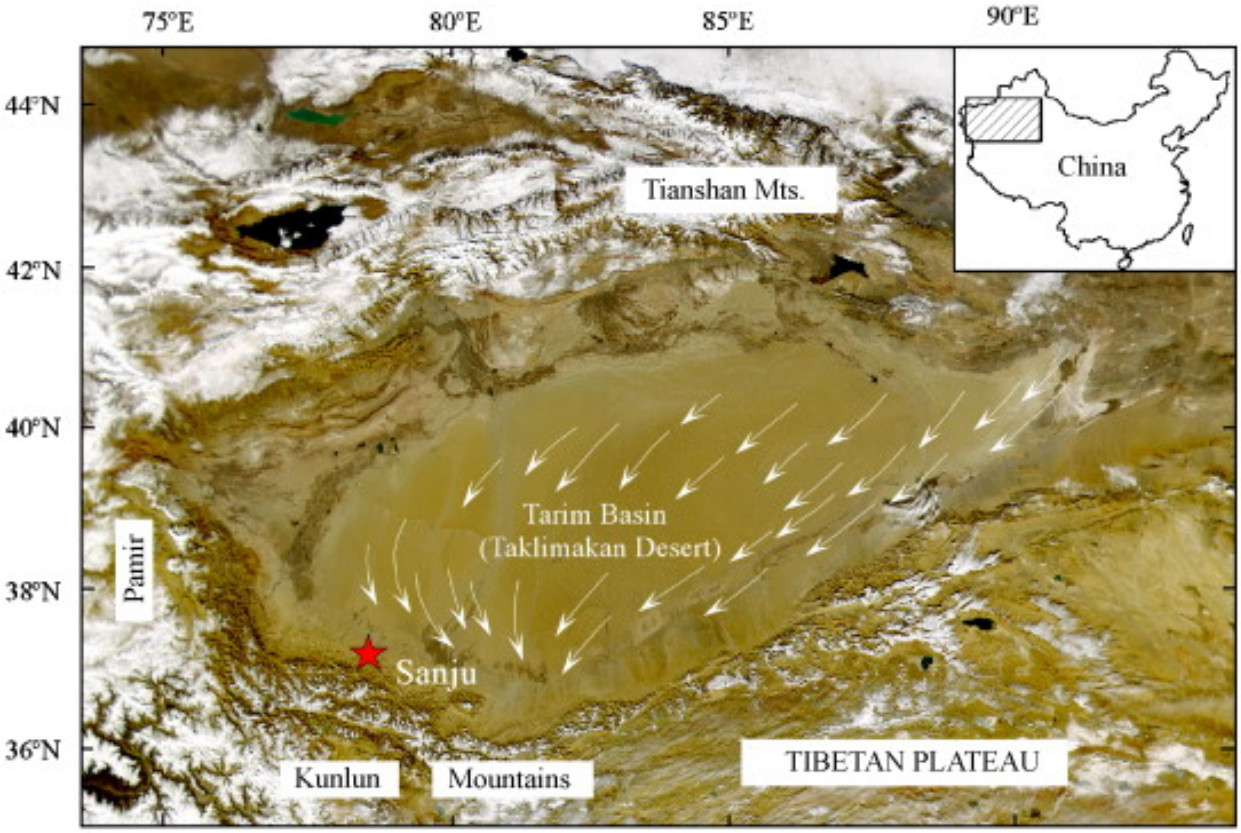


Figure 6: Wind direction in Tarim Basin The arrows indicate directions of the present near-surface winds (dominated by northeasterly winds within the basin),.

.5 Experts Estimates

.5.1 Literature Review

In tables 3 and 4 the soil characteristic mentioned in scientific papers for respectively the Tarim and Qaidam Basin are shown.

Table 3: Soil characteristics in Tarim Basin

Tarim basin			
Soil Characteristic	Min	Max	Mean
SOC	2.26[a], 2.25[b]	17.65[a], 8.57[b]	6.61[a], 4.6[b]
pH	7.2 [c]	-	-
Sand	-	-	-
Silt	-	-	-
Clay	-	-	-

[a]Xu, E.; Zhang, H.; Xu, Y. Effect of Large-Scale Cultivated Land Expansion on the Balance of Soil Carbon and Nitrogen in the Tarim Basin. Agronomy 2019, 9, 86

[b]Zhou, H. H., Chen, Y. N., & Li, W. H. (2010). Soil properties and their spatial pattern in an oasis on the lower reaches of the Tarim River, northwest China. Agricultural Water Management, 97(11), 1915-1922.

[c]Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L.& Chen, D. (2013). A China data set of soil properties for land surface modeling. Journal of Advances in Modeling Earth Systems, 5(2), 212-224.

Table 4: Soil characteristics in Qaidam Basin

Qaidam basin			
Soil Characteristic	Min	Max	Estimate
SOC	-	-	11.46[a]
pH	7.2 [b]	-	8[a]
Sand	-	-	-
Silt	-	-	-
Clay	-	-	-

[a]Qiji, W., Wenying, W., & Fagang, W. (2004). Forming factors and saline-geochemical features of deserted farmland in Qaidam basin. Acta Pedologica Sinica, 41(1), 44-49.

[b]Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L.& Chen, D. (2013). A China data set of soil properties for land surface modeling. Journal of Advances in Modeling Earth Systems, 5(2), 212-224.

.5.2 Expert Estimations

We did not ask Expert A to give only estimations for our target soil property: SOC but also for the soil properties: pH, sand, silt, and clay content. The estimations including the highest and lowest possible value for these soil properties in the study areas Tarim and Qaidam basin can be found in table 5 till 14

Table 5: Soil Organic Carbon in Tarim Basin

Tarim basin SOC					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	0.5	1.5	0.3	0.1	1.3
20 - 40	0.3	1	0.15	0.1	0.8
40 - 60	0.1	0.5	0.1	0.04	0.2
60 - 100	0.05	0.2	0.05	0.01	0.1
100 - 200	0	0.1	0	0	0.08

Table 6: Soil Organic Carbon in Qaidam Basin

Qaidam basin SOC					
Depth in cm	min(g/kg)	max(g/kg)	estimate(g/kg)	Lower(g/kg)	Upper(g/kg)
0 - 20	0	40	8	6	30
20 - 40	0	35	7	5	25
40 - 60	0	30	4	1	20
60 - 100	0	20	1	0.5	15
100 - 200	0	10	0	0	8

Table 7: pH in Tarim Basin

Tarim basin pH					
Depth in cm	min	max	estimate	Lower	Upper
0 - 20	7.5	9	7.8	7.6	8.5
20 - 40	7.5	9	8.2	7.6	8.5
40 - 60	7.5	9	8.5	7.6	8.6
60 - 200	8.0	8.5	7.9	8.2	8.4

Table 8: pH in Qaidam Basin

Qaidam basin pH					
Depth in cm	min	max	estimate	Lower	Upper
0 - 20	7	9.4	8.5	7.8	9.0
20 - 40	7	9.0	8.2	8.0	8.8
40 - 60	7	9.0	8.0	7.8	8.5
60 - 200	6.9	8.5	7.5	7.7	8.4

Table 9: Sand Content (%) in Tarim Basin

Tarim basin Sand Content					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	60	100	90	80	98
20 - 40	60	100	90	75	95
40 - 60	60	100	85	65	90
60 - 200	55	100	75	65	95

Table 10: Sand Content (%) in Qaidam Basin

Qaidam basin Sand Content					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	60	100	85	65	88
20 - 40	60	100	85	65	90
40 - 60	60	100	80	65	85
60 - 200	55	100	75	60	85

Table 11: Silt Content (%) in Tarim Basin

Tarim basin Silt Content					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	3	8	4	4	9
20 - 40	3	8	6	4	9
40 - 60	3	8	8	4	9
60 - 200	3	10	6	5	10

Table 12: Silt Content (%) in Qaidam Basin

Qaidam basin Silt Content					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	2	15	10	5	8
20 - 40	2	15	10	5	8
40 - 60	2	15	10	5	8
60 - 200	2	20	15	7	10

Table 13: Clay Content (%) in Tarim Basin

Tarim basin Clay Content					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	1	8	4	2	6
20 - 40	1	8	4	2	6
40 - 60	1	10	6	2	6
60 - 200	5	15	8	3	8

Table 14: Clay Content (%) in Qaidam Basin

Qaidam basin Clay Content					
Depth in cm	min(%)	max(%)	estimate(%)	Lower(%)	Upper(%)
0 - 20	1	15	7	3	6
20 - 40	1	15	7	3	6
40 - 60	1	15	8	3	6
60 - 200	5	20	10	7	12