Sebastian Schaller

Muristrasse 88, 3006 Bern

Sebastian.schaller@unibe.ch

Conceptual Design Report

Core and wireline log based lithoclasification of unconsolidated sediments from drill cores and boreholes

30 October 2023

Abstract

Gaining insight into the geological subground is a major aim in geosciences. A broad mix of methods and techniques are used to achieve this goal, all with their pro and contra points: i) Geophysics can be used for larger scaled surveys, but they return a physical parameter rather than real geology; ii) Core drilling offers a punctual but high-resolution insight into the real geology and access to undisturbed material, but it is costly and time-consuming; and iii) Destructive flush drilling is fast and relatively inexpensive but offers only a very limited insight into the geology and no access to undisturbed material. This project tries to combine the advantages offered by geophysical and chemical measurements done in the borehole (wireline logging) with the relatively fast and inexpensive destructive flush drilling technique by using advanced statistical analyses in combination with a machine learning approach. The aim is to obtain a data-based lithological borehole profile with a resolution close to that of core drillings, but only at a fraction of the costs and time. The combination of the techniques is not revolutionary per se. In fact, the approach is standard in the oil and gas prospection and exploration industries. However, the combination with advanced data science methods and the application on unconsolidated quaternary sediments is rather new. The project will use a high-resolution data set consisting of a drill core offering high-quality visual description, geophysical core logs, and partly matching geophysical wireline logs from the borehole to develop and test the approach. Further, after successful development, several data sets with similar quality will be used for validation.

Table of Contents

1 Project Objectives and Motivation

Core drilling offers direct and high-resolution access to the in-situ geological underground, in contrast to geophysical methods such as seismic or gravimetry surveys, which return only a (geo)physical parameter, or to the vague impression of the drilled sediments/rocks as provided by destructive flush drilling. However, core drilling is very costly regarding time, money, logistics, and workforce. For example, following the current scientific standards, one meter of core drilling equals ~1000 sFr. Thus, not only the drilling-related costs are high, but also the time investment is great, as it often takes between one and two years from drilling to the first publication of the data. Even if time is slightly less important in science than in private industry, it is still an essential factor for planning projects and acquiring funds. Therefore, it would be of great interest to combine the high speed and low costs of destructive flush drilling with the high resolution and quality of the core drilling in a combined approach without losing too many advantages of both techniques. This may be achieved by combining destructive flush drilling with a wireline logging survey, fast drilling of a cheap hole, and acquiring continuous geophysical and chemical data from the borehole.

Such an approach would allow substituting some expensive core drillings, especially for exploration purposes, to quickly obtain a stratigraphic model by a purely geophysical and chemical log-based classification. It will enable us to gain more information with a still satisfying resolution (> 0.5 m) with the same or even a smaller budget. It would also help to choose optimal locations for the expensive core drillings or even to just score the horizons of interest since core drilling is irreplaceable if access to in-situ material is required. Actually, this approach is well established in the carbon-hydrogen exploration industries[1] (e.g., oil, gas, and coal), mainly targeting well-stratified and homogeneous formal shallow marine sediments from the Mesozoic period[2]. However, this is a terra incognito for the heterogeneous and unconsolidated quaternary sediments.

To explore this tera incognito, we aim to define a catalog of (geo)physical and chemical logs to establish a meaningful geological classification of unconsolidated Quaternary sediments from drill cores and holes from the Northern Alpine Foreland. We try to achieve this in several steps, beginning with the data obtained from an over 250-meter-long drill core, which was recovered in the frame of the "Drilling Overdeepened Alpine Valleys" (DOVE) project[3],[4], as a test data set/case study due to its high quality and resolution in view of sediments and measured data. Later, data from other drill cores from the DOVE project and different data sets with comparable data quality will be used for further testing and validation.

2 Methods

The used infrastructure is local Python installations (Visual Studio Code, Jupiter notebooks, and Anaconda Distribution for Python) on private clients.

The following modules used in the present project are:

1. OS (OperationSystem module)[5] - Navigating to file locations and extracting file names
2. Pandas[6], [7]- Importing and exporting csv files, data manipulation.
3. Numpy[8] - Importing and exporting csv files, data manipulation
4. Matplotlib.pyplot[9] - Plotting figures.
5. Scipy[10] - Statistic analyses
6. Statsmodels[11] - Statistic analyses
7. Scikit-learn[12] - Applying machine learning methods
8. TensorFlow[13], [14] - Further machine learning functionalities
9. Mpltern[15] - Plotting ternary plots

In the first step, after the data is set up and ready for use, the geophysical core logs, measured directly on the cores with the Multi Sensor Core Logger (MSCL) scanner: density, magnetic susceptibility, and natural gamma log are compared with the visually based assigned lithology. This will be done by statistical analyses of the individual logs (CAS-Module 2) and with principle component (PCA) and cluster analyses of the combined MSCL data set (CAS-Module 3). Since in this stage, the data set can be seen as three-dimensional (three components), the results of the PCA and cluster analyses can be compared by simple 3D scatter plots or ternary plots to understand the used codes before moving on to higher dimensional spears by including the wireline log data in a next step.

Loading and setting up the data

i) The individual MSCL data files are loaded.

ii) Adding the needed metadata to link them with the visual-based data (lithology and core quality of the drilled sediments).

iii) The first section of each file, containing the calibration data for additional corrections if needed, is separated, combined into one file, and exported.

iv) The quality and lithology information is added to the individual data files containing the actual log data and combined into one file.

v) The combined MSCL logs are filtered based on core quality data and individual log-specific conditions (e.g., physical implausible values).

vi) A master depth array is created, and corresponding total depth information is added to each data point needed for correlation with wireline data.

vii) Wireline log data are processed and set up[16].

Comparing the geophysical MSCL logs with the visually based lithological classification.

i) Statistical analyses of the individual logs (CAS-Module 2)

ii) PCA

iii) Cluster analyses, comparison between data-based cluster and visual-based classification

iv) Comparison between cluster/PCA analyses and 3D-scatter/ternary plots

Expanding the data set by including wireline data

i) Additional data preparation for compatibility between data sets (e.g., down/upsampling, noise removal, etc.)

ii) PCA

iii) Cluster analyses, comparison between data-based cluster and visual-based classification

iv) Selecting useful data logs for classification

Validation with testing on additional data sets

3 Data

The data set contains i) visual core descriptions of the sedimentological properties, based on which the succession was assigned to one of the four main litho groupes: Diamicts (D), Gravels (G), Sands (S), and Fines (F; Fig. 1); ii) core-log data, measured directly on the cores with the Multi Sensor Core Logger (MSCL) scanner: density, magnetic susceptibility, and natural gamma; iii) proxy data obtained from core samples (Fig. 2); and iv) down-hole wireline logging data. The core-based quality and lithological distributions of the MSCL data points and the impact of a first quality-based cleaning are shown in Figure 3. Since some down-hole logs measured the same parameters as the MSCL core logs, the data set enables a direct high-resolution correlation between core- and borehole logs, visually based sedimentological descriptions, and sampled proxies, despite the differences in sample/data point spacing.

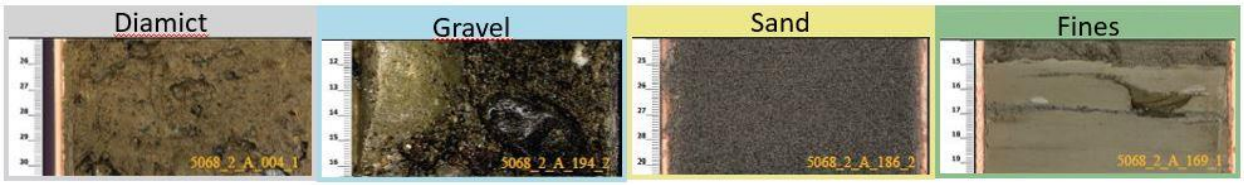


Figure 1: Examples of the four main lithologies encountered in the core.

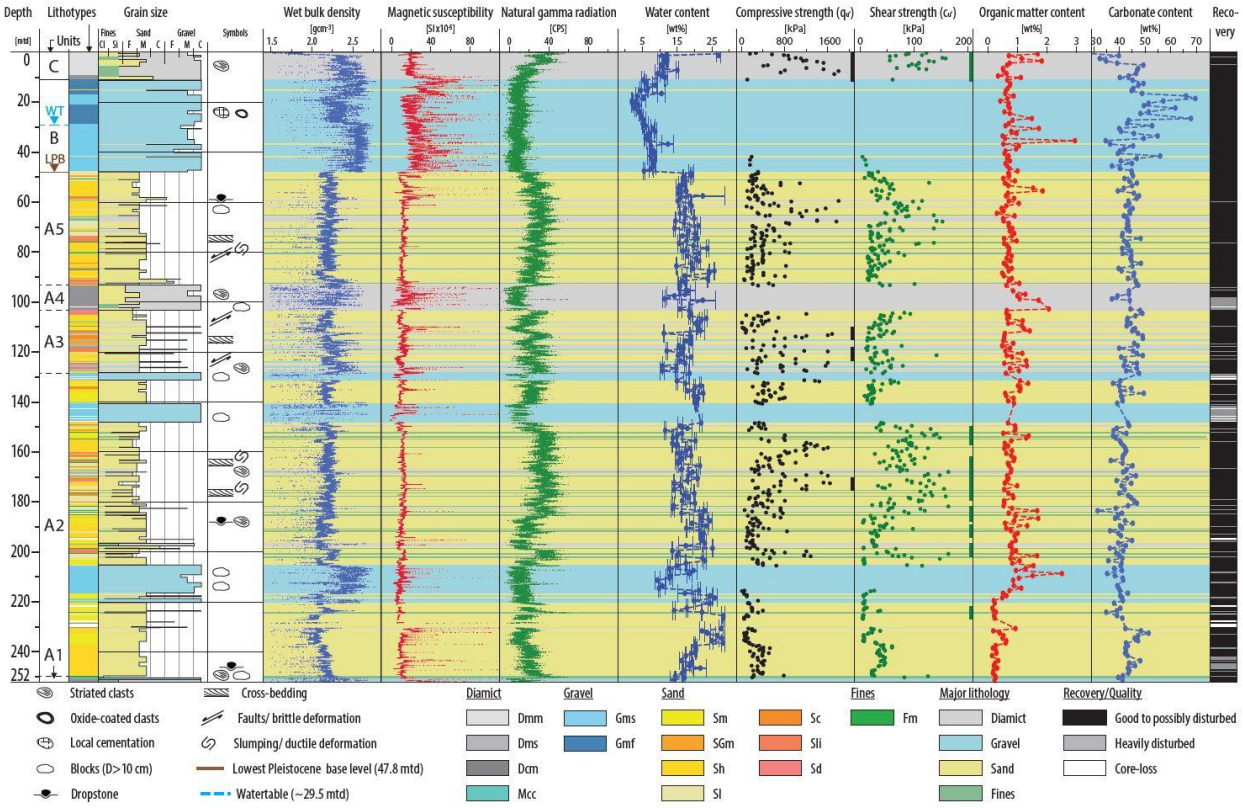


Figure 2: Lithological and petrophysical data versus depth. Columns from left to right: depth-scale [mtd], stratigraphic units (labeled with A1–A5, B, and C; WTDwater table; LPBDlowest Pleistocene base level), lithotypes, dominant grain size with indicated main lithotypes, symbols of prominent observations, wet bulk density (g cm□3), magnetic susceptibility (SI_10□5), natural gamma radiation (CPS, counts per second), water content with indicated standard deviation (wt %), undrained uniaxial compressive strength (qu) (kPa), undrained shear strength (cu) (kPa), organic matter content (wt %), carbonate content (wt %), and the recovery. Main lithotypes are indicated as semitransparent color codes over the plot's width. (Schaller et al., 2023, [3]).

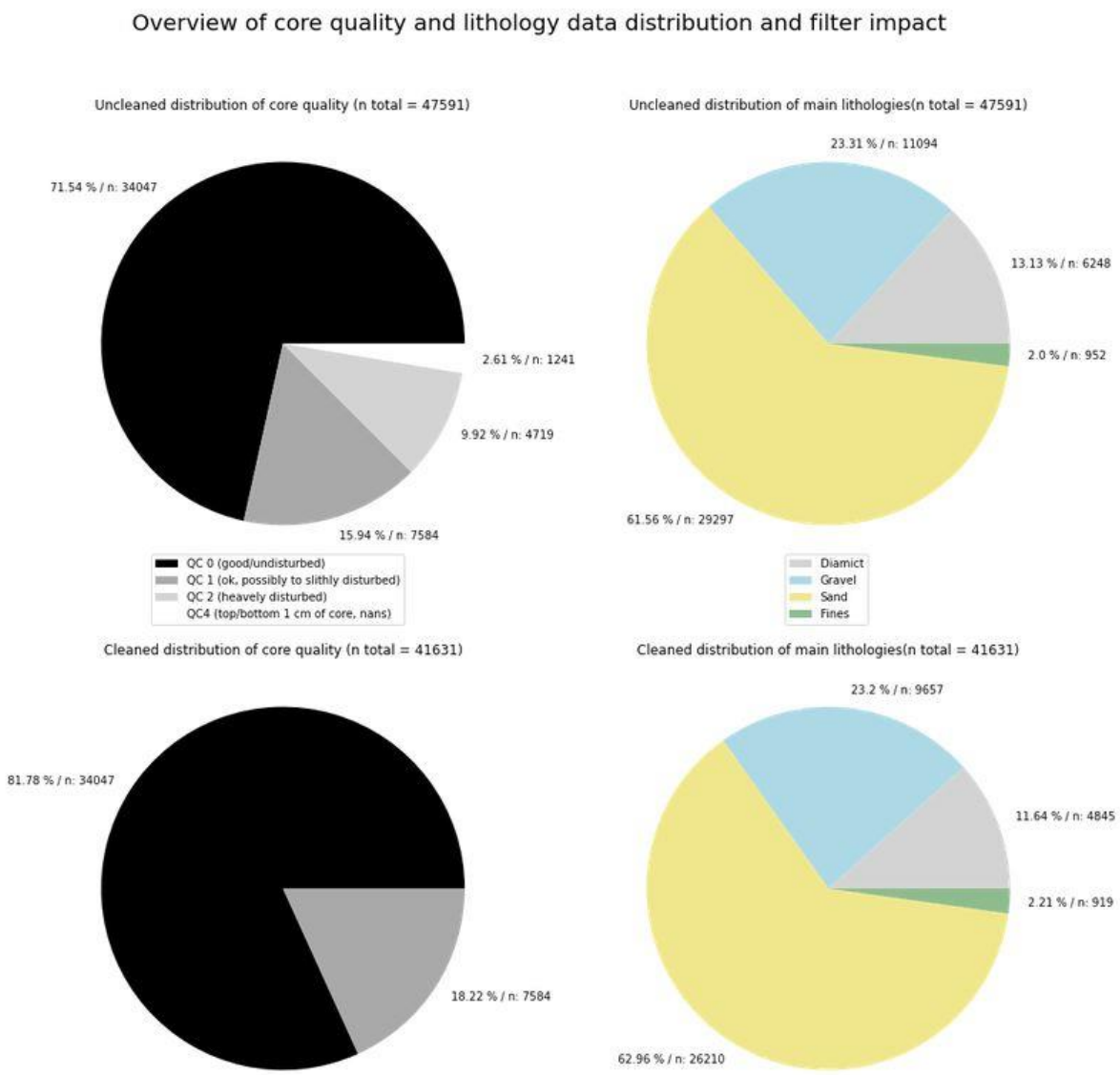


Figure 3: The unfiltered (top) and filtered (bottom) distribution of visual-based quality and lithology classification of MSCL data points of the drill core.

4 Metadata

Each core section is assigned a unique combined ID, and top and bottom depths are assigned after drilling. This core section ID and driller depth build the link between the different types of data (visual core data, MSCL logs, core samples, and wireline logs). Further, to link the MSCL data with the core sections, the ID of the measurement session (filename of the raw data file) and the number of the measured core section (numerical label, starts at 1 when beginning with a new measurement session), is added (Table 1). The three types of metadata (core-related, measurement-related, and depth-related) allow the correlation and combination of different data sets with different resolutions.

Table 1: First five rows of the used metadata table to link the MSCL data with the visual-based core description. From left to right: raw MSCL data file name/ measurement session ID, measurement nr of core section (cal = calibration section), nr of drill core, nr of core section, unique core section label (Project-Nr_Site-ID_drill-core _core-section), top depth of core section, bottom depth of core section, section type (0 = calibration, 1 = actual data).

measurement ID	Section Nr.	Core run	Core-Section	Combined-ID	top depth	bottom depth	used data
5068_2_A_20210526_1	cal	std	0	cal	0	0	0
5068_2_A_20210526_1	1	1	1	5068_2_1_1	0	1	1
5068_2_A_20210526_2	cal	std	0	cal	0	0	0
5068_2_A_20210526_2	1	2	1	5068_2_2_1	1	2	1
5068_2_A_20210526_2	2	3	1	5068_2_3_1	2	3	1

All data used for the analysis, except the raw MSCL and wireline data and some metadata used in processing, are accessible via the ICDP-DOVE project page (https://www.icdp-online.org/projects/by-continent/europe/dove-switzerland). Further, the used dataset[17] is published along with the operational report[16].

5 Data Quality

The data quality is determined by i) the quality of the drilled sediments and ii) the quality of the measurements. The quality of the drilled sediments is visually classified based on their degree of disturbance. A slightly to partly disturbed section may have lost part of its structural information but still preserved some of its lithological and physical properties. Standard calibration procedures control the quality of the measurements and are relatively robust. Therefore, the data quality is mainly determined by the quality of the recovered sediments.

Consequently, poor core quality will result in poor data quality and can only provide limited information gain. The size of the data set is determined by the number of different logs (number of columns/features), the length of the drill core/depth of the drill hole, and the spacing/number of data points of the individual logs (number of rows). The diversity/types of the measured parameters (number of features) is more crucial than the number of data points. Reducing the number of data points decreases the resolution, but missing a critical component leads to losing a whole dimension/aspect of the dataset. Therefore, a larger data set with respect to features (columns) and data points (rows) of good data quality is crucial for developing, testing, and validating the model and approach in a decent resolution. Smaller gaps in the data representing core loss or heavily disturbed sections should not be a larger issue in a decent-sized data set. Still, they should be considered using the stratigraphic position or the distance between neighboring data points as a feature or for further data processing (e.g., rolling mean). The used data originates from a drill core with overall

good to very good quality (Figs. 2 and 3), with only minor gaps and some disturbed sections. Measurement data are of a robust quality and a high sample interval resolution, allowing further data cleaning, manipulation, and downsampling if needed without any significant loss of resolution.

6 Data Flow

The data flow can be separated into four parts, accumulating together in the final project (Fig. 4). The pre-CAS stage consists of three main steps: i) Data acquisition ("digitalizing the physical world"); ii) Data preparation and cleaning ("making the data useful"); and iii) Plotting overview of all. The blocks of module 2 (descriptive statistical properties) and module 3 (principal component (PCA) and cluster analyses) are first used to investigate only the MSCL logs to establish and test the approach. In the next step, modules 2 and 3 will be repeated with additional data until satisfying results are reached ("Checkpoint 1") with the case study data set. After, validation with other data sets with similar quality is done until satisfying results are achieved again ("Checkpoint 2"). The flow after testing the principle with the MSCL logs can be seen as a loop that expands its used data with each iteration until, in the ideal case, a catalog of useful features for data-based classification of unconsolidated quaternary sediments in drill cores and boreholes is created.

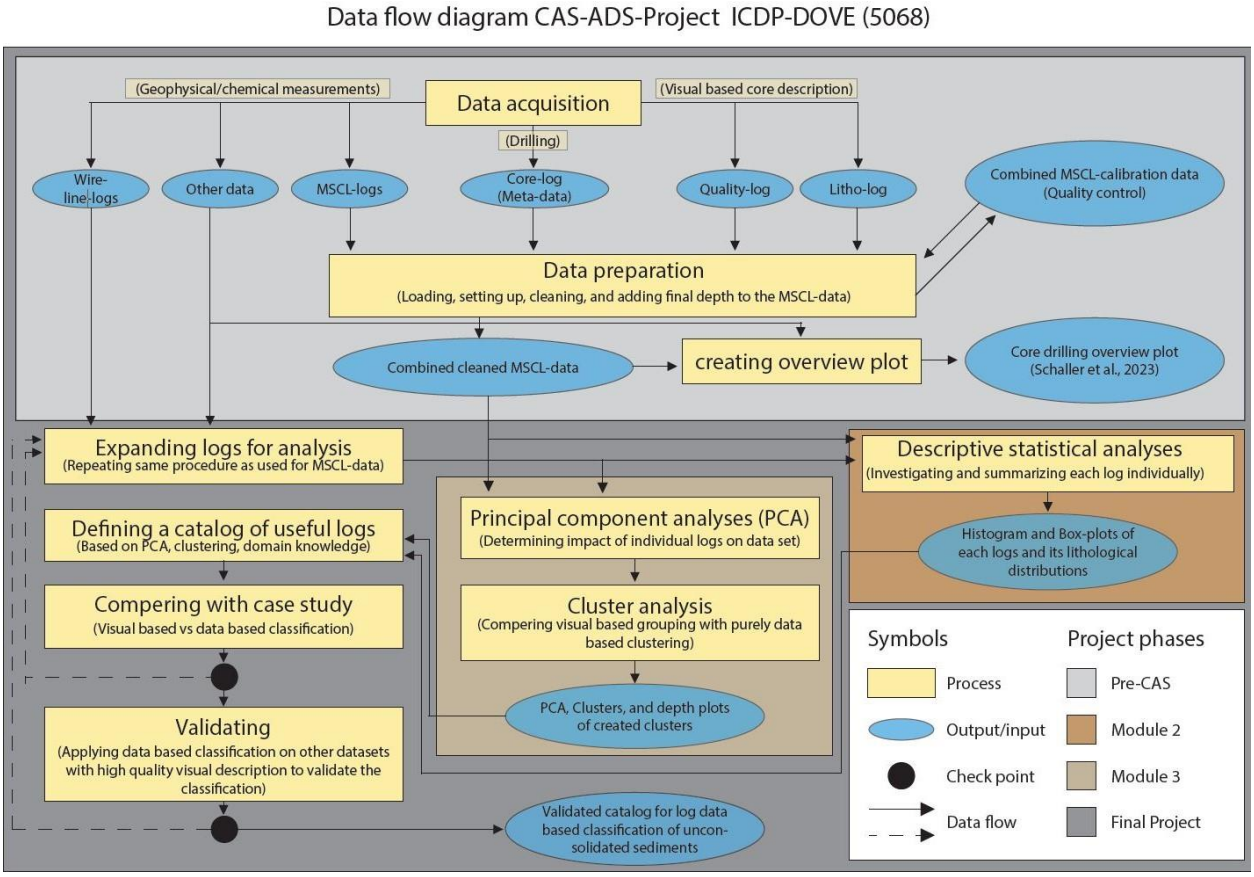


Figure 4: Data flow diagram of the project, with color-indicated stages of the project development and its interaction.

7 Data Model

On a conceptual level, the aim is to use the collected core and wireline data (input) to analyse and compare in order to choose a selection of geophysical and chemical logs for a data-based lithological classification of the drilled/cored sediments (Fig. 5). This will be done by PCA and cluster analyses of the different data sets, the correlation between the different data sets (e.g., visually based core data, MSCL, and wireline logs, …) is possible over the three types of metadata (core-related, measurement-related, and depth-related), the link between the data sets and its features on a logical level is shown in Figure 6. A decent (desktop) computer for data processing with an internet connection and some free memory is sufficient for further data processing and analysis after the data acquisition. The needed equipment and techniques for the data acquisition are described in Schaller et al., 2023[18] and DOVE operational report[16].

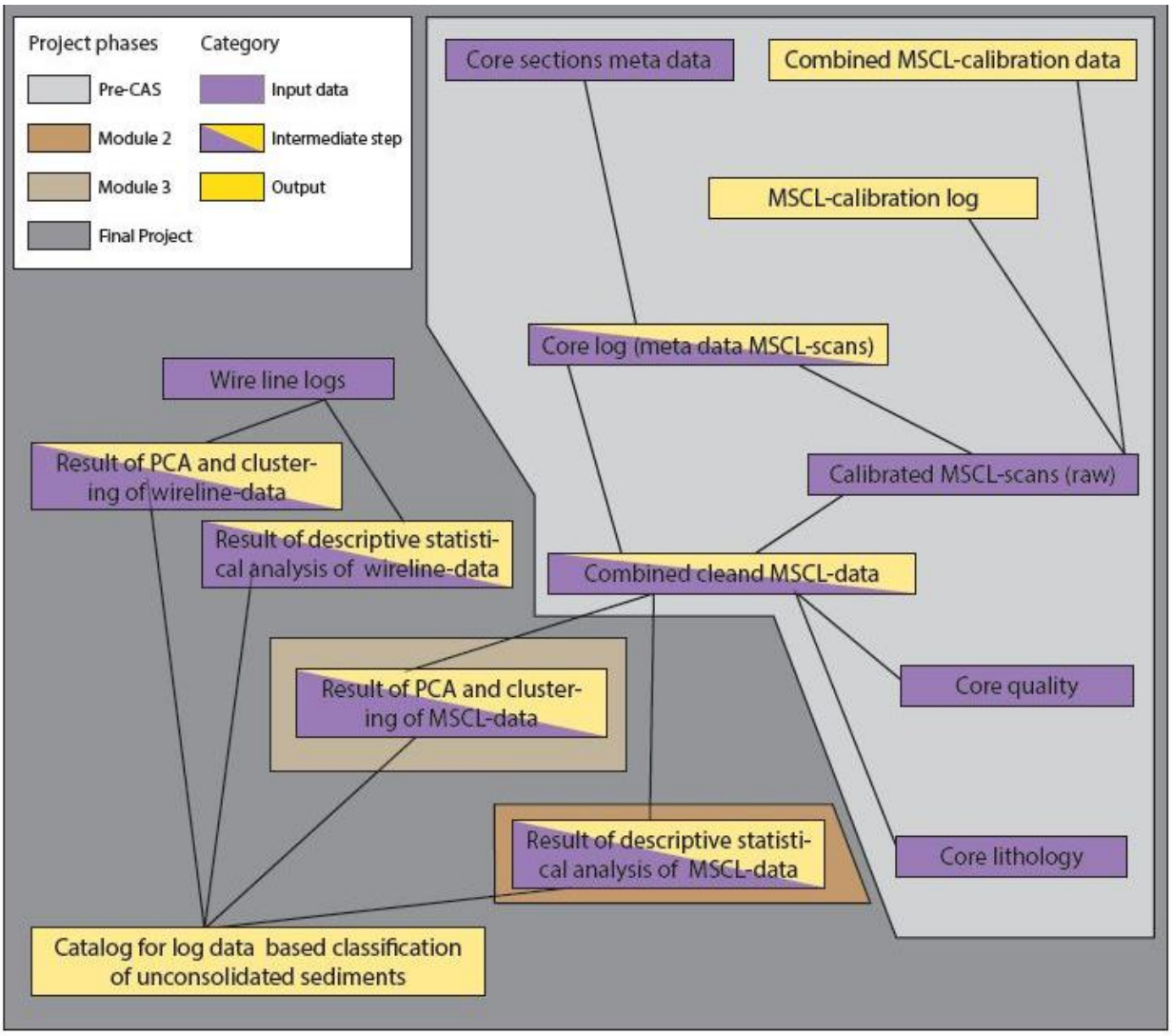


Figure 5: Conceptual data model indicating the relationship between the different inputs/outputs. The color code indicates the different stages of the development/building of the project, which are also indicated in the data flow model (Fig. 4).

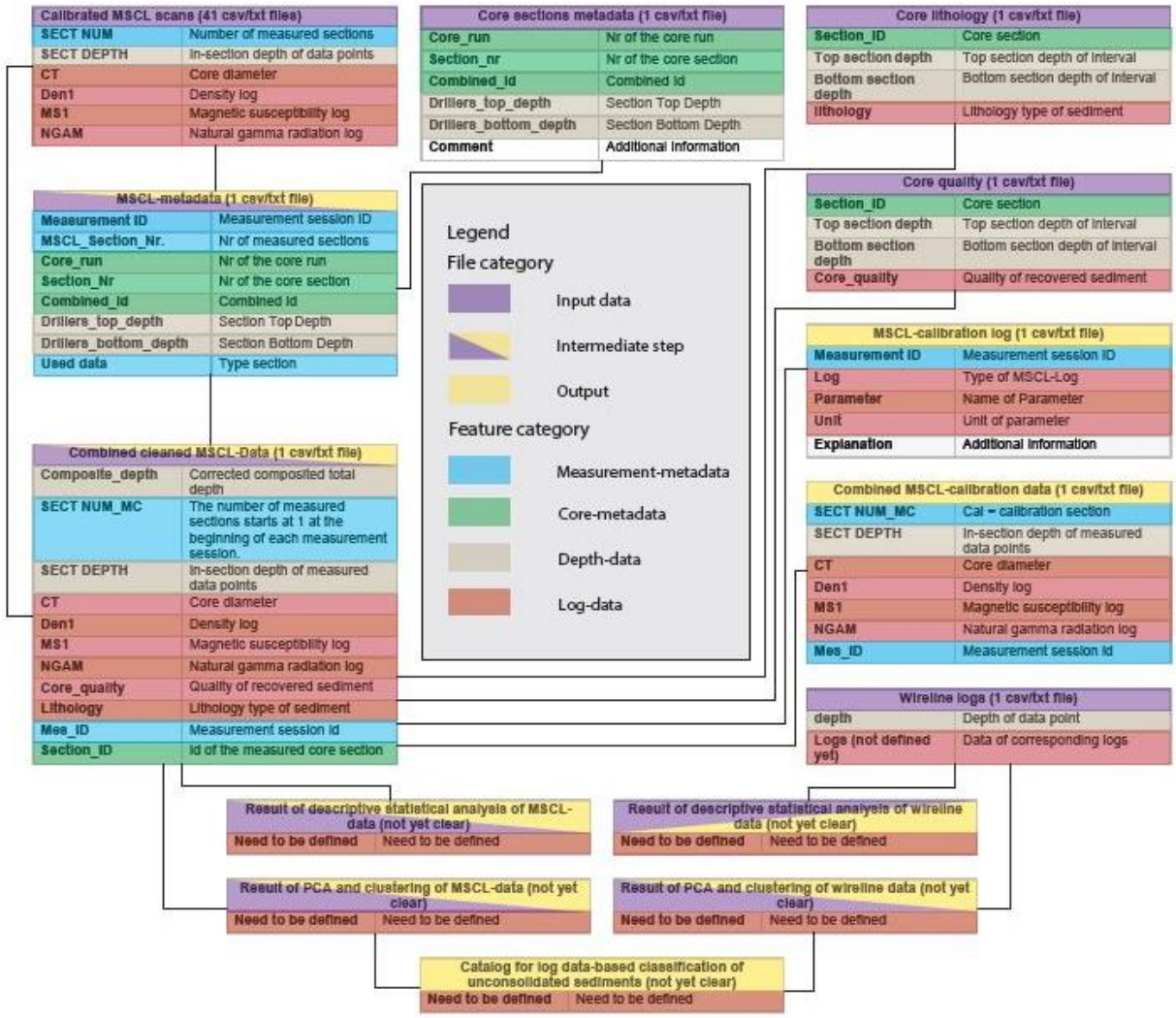


Figure 6: The logical data model indicates the relationship between the different feature classes and the relationship between the different input/output data files. The color code indicates the role of the different files/objects and the category of the individual features.

8 Documentation

The used data and code, including explanatory notes for understanding, will be documented using Jupiter notebooks, which will be published in the appendix or as a separate data publication if the project's findings are included in a scientific publication. Alternatively, it may be included as a chapter in my Ph.D. thesis if not published.

9 Risks

Two aspects could be challenging for the project: i) the data clusters are not clearly separatable due to overlapping properties, some missing features, or the limits of only wireline/core-log-based classification

of unconsolidated sediments (e.g., other aspects needed for precise classification), and ii) the selected feature category is too unflexible to apply to data sets from different unrelated locations. Countermeasure for the first challenge: i) expanding the log-catalog for clustering, ii) using a supervised approach (e.g., random forest tree). However, this would likely cause new challenges (e.g., large data sets of excellent visual and log-based data quality from diverse sites to build a model for broader use). Countermeasure for the second challenge: expanding the catalog to make it more flexible and using, for example, PCA to select the valuable logs for each side individually from the extended catalog, but this would require a broad range of log data to choose from. In general, using a supervised approach likely helps to compare and evaluate the results of the unsupervised one, which would also show the quality of the visual-based classification. Including a supervised approach in the project should not significantly impact the general project schedule. Instead, it would improve the project and should be considered to include.

10 Preliminary Studies

A detailed description of the drilling and data acquisition process, as well as the geological background, motivation, and first interpretation of the data gained from the drill core (Fig 2), which serves as a test data set for this study, is provided in Schaller et al., 2023[18] and operational report[16].

Further, the descriptive statistical values of the three different MSCL logs and the visual-based lithological grouping have been investigated during the work for Module 2. The aim was to gain an overview of the data and see if they support the visually-based classification and make geological sense (Figs. 7-12). Each log was individually analysed for i) the whole log is not normally distributed -> contains more than one population (Table 2; D'Agostino-Pearson-Normality-Test[19]); ii) the sup-datasets representing the four main lithologies do not come from the same population -> are disintegrable from each other (Table 2; Kruskal-Wallis-Test[20], and Pairwise Mann-Whitney-U-Test[21]). Both corresponding 0-hypotheses were rejected. If this had not been the case, the data would not have been usable for the project.

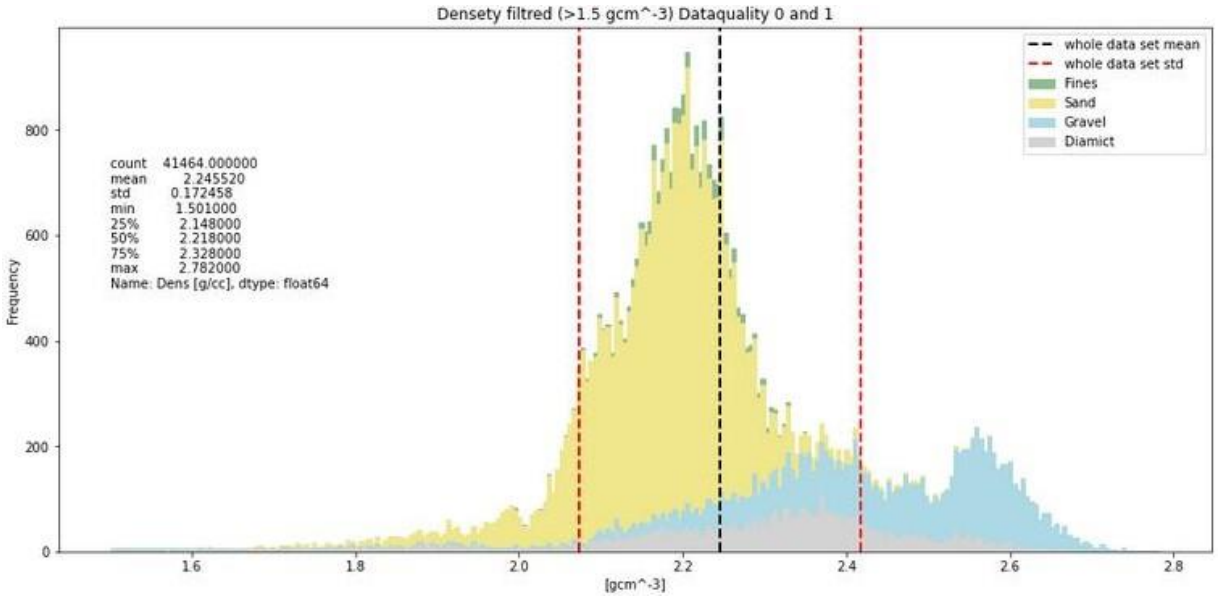


Figure 7: Histogram of the four sub-datasets representing the four main lithologies of the cleaned density log.

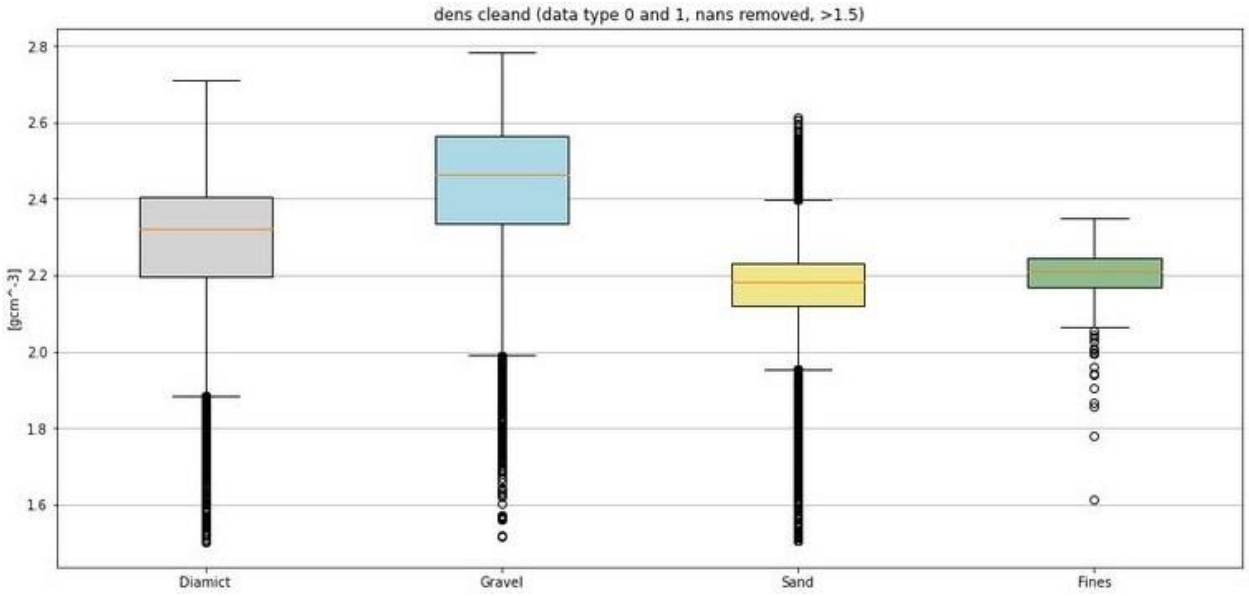


Figure 8: Q-Q whiskers plot of the four sub-datasets representing the four main lithologies of the cleaned density log.

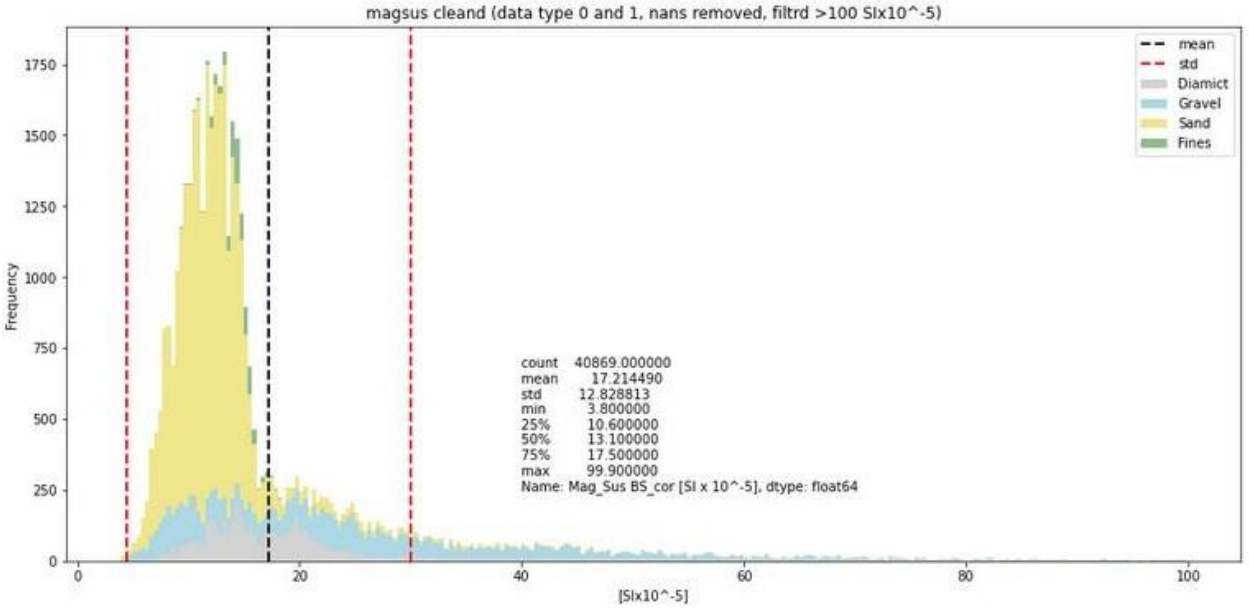
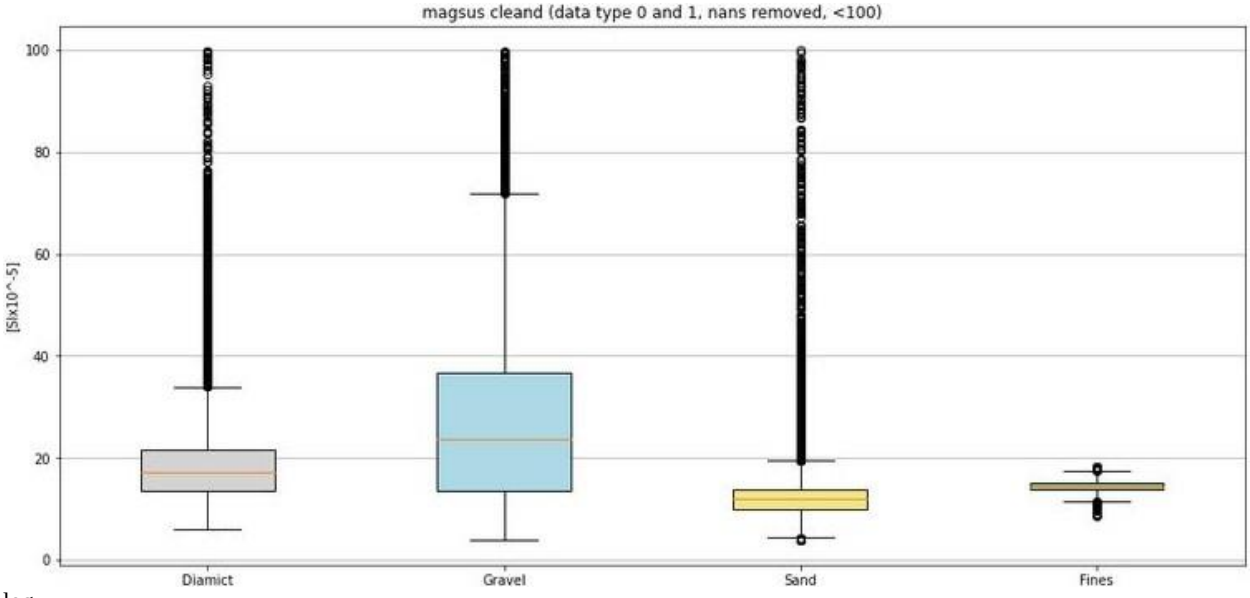


Figure 9: Histogram of the four sub-datasets representing the four main lithologies of the cleaned magnetic susceptibility log.

Figure 10: Q-Q whiskers plot of the four sub-datasets representing the four main lithologies of the cleaned magnetic susceptibility log.



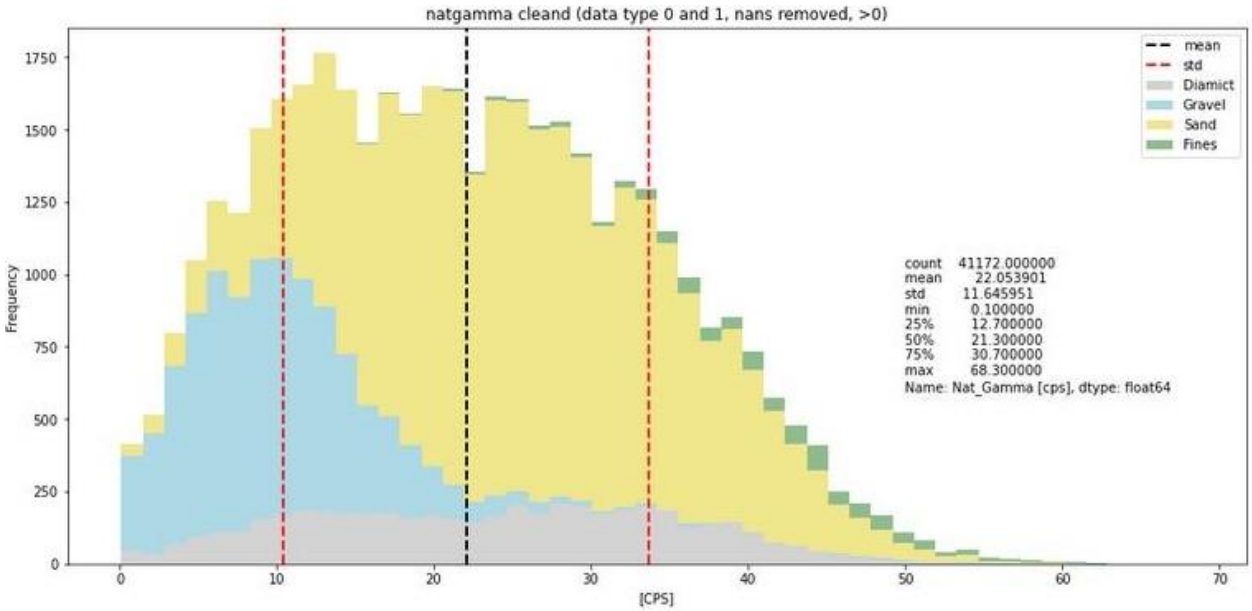


Figure 11: Histogram of the four sub-datasets representing the four main lithologies of the cleaned natural gamma radiation log.

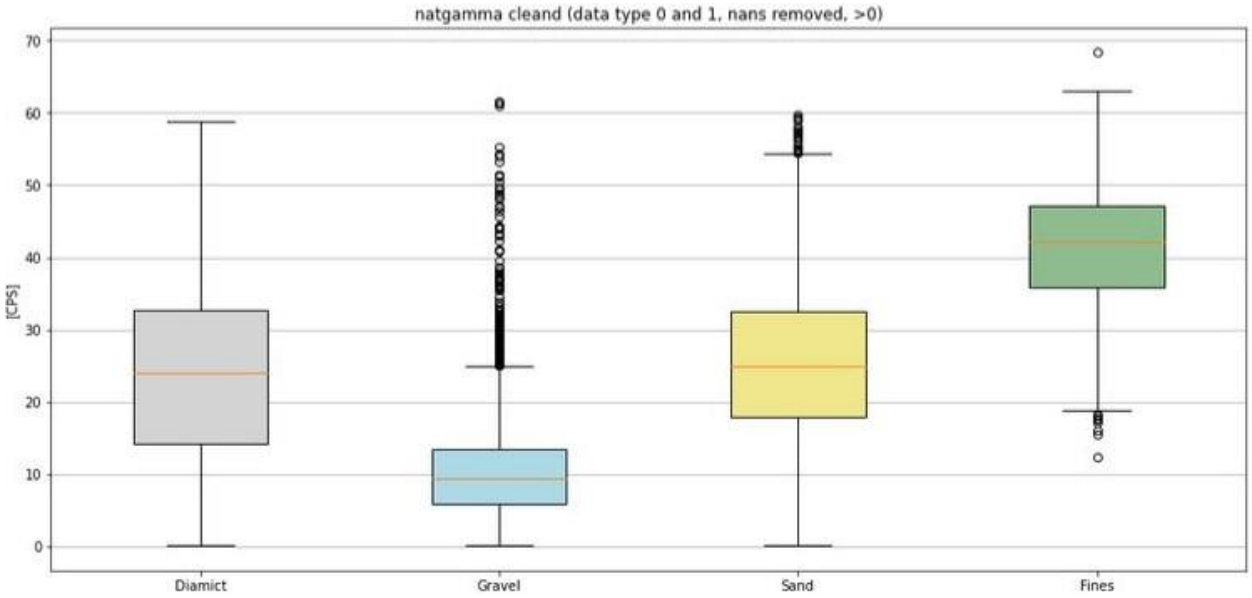


Figure 12: Q-Q whiskers plot of the four sub-datasets representing the four main lithologies of the cleaned natural gamma radiation log.

Table 2: summary of the statistical test: D'Agostino-Pearson-Normality-Test, Kruskal-Wallis-Test, and Pairwise Mann-Whitney-U-Test, Pairwise Mann-Whitney-U-Test with itself for control.

	Density	**Mag. Sus.**	**Nat. Gamma**
	D`Agostino-Pearson-Normality-Test		
	p-value	**p-value**	**p-value**
Whole dataset	0.000	0.000	0.000
Diamict	0.000	0.000	0.000
Gravel	0.000	0.000	0.000
Sand	0.000	0.000	0.000
Fines	0.000	0.000	0.000
	Kruskal-Wallis-Test		
comparison	p-value	p-value	p-value
all	0.000	0.000	0.000
	Pairwise Mann-Whitney-U-Test		
comparison	p-value	p-value	p-value
Diamict-Diamict	1.000	1.000	1.000
Diamict-Gravel	0.000	0.000	0.000
Diamict-Sand	0.000	0.000	0.000
Diamict-Fines	0.000	0.000	0.000
Gravel-Diamict	0.000	0.000	0.000
Gravel-Gravel	1.000	1.000	1.000
Gravel-Sand	0.000	0.000	0.000
Gravel-Fines	0.000	0.000	0.000
Sand-Diamict	0.000	0.000	0.000
Sand-Gravel	0.000	0.000	0.000
Sand-Sand	1.000	1.000	1.000
Sand-Fines	0.000	0.000	0.000
Fines-Diamict	0.000	0.000	0.000
Fines-Gravel	0.000	0.000	0.000
Fines-Sand	0.000	0.000	0.000
Fines-Fines	1.000	1.000	1.000

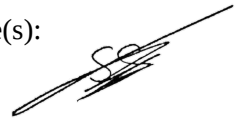
11 Conclusions

Developing a geophysical and chemical log-based approach for lithological classification of unconsolidated Quaternary sediments originating from drill cores and boreholes would be valuable for subsurface investigation. If the results are reliable, it would be a good compromise between costs, time, and information gain. Therefore, the project aims to investigate and, if possible, develop and prove such a tool's application.

Statement

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen."

Date: 26.10.2023 Signature(s):



References

[1] N. H. Mondol, "Well Logging: Principles, Applications and Uncertainties," in *Petroleum Geoscience: From Sedimentary Environments to Rock Physics*, K. Bjørlykke, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 385–425. doi: 10.1007/978-3-642-34132-8_16.

[2] A. Carrasquilla, "Lithofacies prediction from conventional well logs using geological information, wavelet transform, and decision tree approach in a carbonate reservoir in southeastern Brazil," *J. South Am. Earth Sci.*, vol. 128, p. 104431, Aug. 2023, doi: 10.1016/j.jsames.2023.104431.

[3] S. Schaller, M. W. Buechi, B. Schuster, and F. S. Anselmetti, "Drilling into a deep buried valley (ICDP DOVE): a 252\,m long sediment succession from a glacial overdeepening in northwestern Switzerland," *Sci. Drill.*, vol. 32, pp. 27–42, 2023, doi: 10.5194/sd-32-27-2023.

[4] F. S. Anselmetti *et al.*, "Drilling Overdeepened Alpine Valleys (ICDP-DOVE): quantifying the age, extent, and environmental impact of Alpine glaciations," *Sci. Drill.*, vol. 31, pp. 51–70, Oct. 2022, doi: 10.5194/sd-31-51-2022.

[5] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[6] The pandas develop teamment, "pandas-dev/pandas: Pandas." Zenodo, Sep. 2023. doi: 10.5281/zenodo.3509134.

[7] W. McKinney and others, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, 2010, pp. 51–56.

[8] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, Art. no. 7825, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[9] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.

[10] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nat. Methods*, vol. 17, pp. 261–272, 2020, doi: 10.1038/s41592-019-0686-2.

[11] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[12] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[13] TensorFlow Developers., "TensorFlow." Zenodo, 2023. doi: https://doi.org/10.5281/ZENODO.4724125.

[14] Martín Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." 2015. [Online]. Available: https://www.tensorflow.org/

[15] Y. Ikeda, "mpltern: Ternary plots with Matplotlib." Zenodo, 2023. doi: https://doi.org/10.5281/zenodo.3528354.

[16] DOVE-Phase 1 Scientific Team et al, "Drilling Overdeepened Alpine Valleys (DOVE) – Operational Report of Phase 1." GFZ Data Services, 2023. doi: https://doi.org/10.48440/ICDP.5068.001.

[17] DOVE-Phase 1 Scientific Team et al, "Drilling Overdeepened Alpine Valleys (DOVE) – Operational Dataset of Phase 1." GFZ Data Services, 2023. doi: https://doi.org/10.5880/ICDP.5068.001.

[18] S. Schaller, M. W. Buechi, B. Schuster, and F. S. Anselmetti, "Drilling into a deep buried valley (ICDP DOVE): a 252m long sediment succession from a glacial overdeepening in northwestern Switzerland," *Scientific Drilling*, 2023, doi: acepted (in proofreading).

[19] R. D'Agostino and E. S. Pearson, "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and √ b1," *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973, doi: 10.2307/2335012.

[20] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *J. Am. Stat. Assoc.*, vol. 47, no. 260, pp. 583–621, 1952, doi: 10.2307/2280779.

[21] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *Ann. Math. Stat.*, vol. 18, no. 1, pp. 50–60, 1947.