



# Statistical approach to identify variables predicting sulphide clay occurrence in southern Finland

Maarit Saesma<sup>1,2</sup> · Monica Löfman<sup>2</sup> · Emilia Kosonen<sup>1</sup> · Antti E. K. Ojala<sup>3</sup> · Leena Korkiala-Tanttu<sup>2</sup>

Received: 22 April 2022 / Accepted: 10 May 2023  
© The Author(s) 2023

## Abstract

Acid sulphate soil and sulphide-bearing sediments cause various challenges in construction projects and land use planning, as well as harmful environmental effects. Fine-grained sulphide sediments were mainly formed in coastal areas during the Litorina Sea water phase at approximately 7000 BP in the capital region of Finland, but not all these sediments contain sulphide clay. In this study, environmental and material property variables related to the depositional conditions of sulphide clay were selected for statistical analyses to find their association with the occurrence of sulphide. The datasets consisted of sulphide investigations by the City of Espoo, the City of Helsinki, and the Geological Survey of Finland. Statistically significant associations were found in the study area between the occurrence of sulphide and enumerative variables (i.e., sediment organic content, total clay depth, topographic class in the Litorina Sea phase, and water depth) in the Litorina Sea phase. Locations where sulphide clay is especially likely to occur consist of organic-rich ( $\geq 2\%$ ) thick clay ( $\geq 15$  m) deposits in a topographically narrow depression with deep Litorina water ( $\geq 30$  m), or where there is a moderate depth clay (3–5 m) in a local depression with shallow Litorina water (10–20 m). The best individual predictor for sulphide clay occurrence in the study area was found to be the sediment organic content, and, together with sediment water content, these variables very accurately predicted the occurrence of sulphide clay. In addition, clay depth is a very good predictor and, together with the topographic class narrow depression and the Litorina water depth or current elevation, can be used to predict sulphide occurrence.

**Keywords** Sulphide · Clay depth · Topography · Litorina water depth · Organic content · Statistics

## Introduction

Sulphide sediments and acid sulphate soil can cause challenges in construction and land use planning, such as an increase in construction costs and a possible delay in the construction schedule. Moreover, construction in areas with soft fine-grained sediments leads to surplus soils totalling several millions of tonnes annually in the capital region of Finland due to mass exchange, i.e., the excavation of poor-quality soil and landfill with material suitable for

construction (Forsman et al. 2013). Part of this geotechnically poor quality soil contains acid sulphate soil material and sulphide-bearing sediment (potential acid sulphate soil). Both of these require special treatment (e.g., neutralization) before storage in a landfill site to avoid harmful environmental effects due to increased acidity in adjacent waters and metal leaching. Awareness and identification of corrosion and microbial corrosion related to construction have increased (e.g., Finnish Transport Infrastructure Agency 2017; Suikkanen et al. 2018; Autiola et al. 2022), meaning that special attention should be paid to the construction design of steel and concrete structures on ground potentially comprising sulphide soil. In addition, the ground improvement method of soil stabilization for sulphide clay is likely to require some adjustment to the mixture of binders (Andersson and Norrman 2004; Autiola et al. 2012). More binder material is often initially needed to neutralize the soil, and only after that can the binding reactions start. Sulphide soils in Sweden are usually classified into same

✉ Maarit Saesma  
maarit.saesma@gtk.fi

<sup>1</sup> Geological Survey of Finland, Vuorimiehentie 5,  
02151 Espoo, Finland

<sup>2</sup> Department of Civil Engineering, Aalto University,  
Rakentajanaukio 4, 02150 Espoo, Finland

<sup>3</sup> Department of Geography and Geology, University of Turku,  
Akatemiankatu 1, 20014 Turku, Finland

category as organic soils, since they often contain a considerable amount of organic material, thus displaying similar geotechnical behaviour (Larsson 1990; Larsson et al. 2007; Westerberg et al. 2015). Generally, in the undrained shear strength evaluation of soil using field vane and fall cone tests, the correction factor may need to be adjusted in the case of organic sulphide soil (Larsson 1990; Larsson et al. 2007; Westerberg et al. 2015).

Acid sulphate soils are known worldwide and are usually related to fine-grained sediments of marine or brackish water origin in coastal areas (Pons 1973; Dent and Pons 1995). Estimates of the amount of acid sulphate soil in Finland have been presented, for example, by Yli-Halla et al. (1999) based on criteria suitable for Finnish soils. In Finland, most of sulphide sediments have been identified in the coastal areas of the Baltic Sea, specifically in areas confined by the Litorina Sea (Ojala et al. 2013). The adverse effects of acid sulphate soils used for agricultural purposes have been investigated by Erviö (1975), among others. Acid sulphate soils are generally harmful, because they contain an excess amount of sulphur and their pH is very low. The abundant sulphur in these sediments, originating from deposition in saline or brackish water with microbial activity, occurs in a reduced form as pyrite ( $\text{FeS}_2$ ) or metastable iron sulphide ( $\text{FeS}$ ) below groundwater (Boman et al. 2008). When exposed to atmospheric oxygen, sulphides oxidize and the pH drops, causing acidity and metal mobilization (e.g., Boman et al. 2008, 2010). Numerous environmental studies related to acid sulphate soils have been conducted, of which many have focused on sulphide oxidation, geochemistry, and associated metal leaching, which leads to the poor quality of adjacent waters (e.g., Åström and Björklund 1997; Roos and Åström 2005).

The systematic mapping of acid sulphate soils in Finland down to a maximum depth of three metres started in 2009 with the classification of potential acid sulphate soils (Edén et al. 2012a). Mapping projects were encouraged by the Ministry of Agriculture and Forestry and the Ministry of the Environment (2011) to support the national strategic work on the prevention of contamination of natural waters caused by acid sulphate soil. Since then, most of the western, south-western, and southern coastal areas of Finland have been mapped and classified for acid sulphate soils according to the criteria of the Geological Survey of Finland (Acid sulphate soils, Geological Survey of Finland; Edén et al. 2012b). Contemporaneous methods and guidelines have been presented by the Finnish Transport Infrastructure Agency (2012, 2017) and Pousette (2007) for defining aggressive and corrosive soil. Investigations conducted in connection with construction in the cities of Espoo and Helsinki usually follow one or more of these methods and guidelines, irrespective of whether there is sulphide soil at the planned site. There is a need for a unified methodology to conduct

sulphide soil investigations, especially in land use and construction projects. To address this need, a national guide has been published by Autiola et al. (2022), which defines the methods and actions recommended for sulphide soil investigations. Currently, the number of sulphide investigations in Finland is increasing because construction activities in populated coastal cities are extending to areas that contain soft clay deposited in the Litorina Sea phase, which is often potential sulphide clay (Eden et al. 2012a). Despite the high probability that Litorina clay contains sulphides, many of the analysed samples in the Espoo area do not contain sulphide (Hämäläinen 2018; Saresma et al. 2020). To focus the site-specific investigations more efficiently, preliminary studies on the potential areas for sulphide soil are recommended, preferably in the early stage of land use planning.

Spatial modelling and machine learning methods have been applied to predict the occurrence of acid sulphate soils of different extent in southern Finland (Beucher 2015; Estévez et al. 2022). Some of these methods are promising and speed up the process of predicting potential acid sulphate soils compared to conventional mapping. However, the results from supervised machine learning methods are highly dependent on the input parameters and require evaluation and interpretation in order to obtain reliable maps for predicting the occurrence of acid sulphate soil (Estévez et al. 2022).

So far, less attention has been given to the depositional environments of the Litorina Sea phase and the role of environmental variables in the formation and burial of sulphide-containing sediments in the capital region of Finland. Indications of the importance of clay thickness in sulphide sediment generation have been presented, for example, by Ojala et al. (2007, 2018) from the Suurpelto and Rastaala areas in Espoo, where thick clay layers are associated with the occurrence of sulphide. A similar interpretation was made by Saresma et al. (2020) with data from extensive areas in the city of Espoo. The same study revealed that the topographic classification of the Litorina Sea phase together with clay thickness and gytjtja and gytjtja clay areas can be used to predict sulphide occurrence. Hämäläinen (2018) found a correlation between the sediment organic matter and sulphate content, as well as between the sediment water content and sulphate content in fine-grained soil samples taken in the Espoo area.

The present study statistically analysed the association of selected environmental and material property variables to reveal the occurrence of sulphide using investigation data from the Espoo and Helsinki areas. The approach was based on the selection of environmental variables related to terrain characteristics and clay thickness, as well as sediment property variables, as presented in the methods section. The present study builds upon palaeotopographic models created for the Litorina Sea phase of the Baltic Sea basin (BSB) in the Espoo and Helsinki area (Saresma et al. 2020,

2021a, b) and sulphide analysis conducted at a number of points around the capital region from clayey sequences. The main aim of this research was to identify the most significant variables that predict the occurrence of sulphide clay by (1) analysing the association between the occurrence of sulphide and each variable and variable category, (2) finding relationships between variables associated with the occurrence of sulphide, and (3) fitting a statistical model with selected variables.

## Materials and methods

### Geological setting of the study area

The present study area covers the capital region of Finland, including the cities of Helsinki, Espoo, and Kauniainen; the southern parts of the city of Vantaa; and the easternmost parts of the city of Kirkkonummi (Figs. 1–4). The geological characteristics and post-glacial evolution of the area have been described by Ojala et al. (2018) and Saresma et al. (2021a), among others. The most important detail is that the present study area was an open embayment of the BSB until the Litorina Sea phase and has since been uplifting above the BSB as a result of glacio-isostatic rebound (Björck 1995; Ojala et al. 2013). As a consequence, fine-grained (clayey) sediments are the predominant superficial deposits in the area. They cover the underlying crystalline bedrock and glacial tills, and were deposited in the BSB following the retreat of the Fennoscandian Ice Sheet at around 13,000 BP.

The palaeotopographic analysis in the present study followed the classification and methodology presented earlier by Saresma et al. (2021a) for the Espoo area, which was expanded to the Helsinki area by Saresma et al. (2021b). Here, the palaeotopographic analyses of Espoo and Helsinki were combined to cover the entire study area. The palaeotopography of the study area ranges from high hills in the north of Espoo to lowlands in the south, which formed during the Litorina Sea transgression (7000 before present (BP)) (Fig. 1a). The highest shoreline of the Litorina Sea phase is often used to define the upper limit of the acid sulphate soils around the Baltic Sea (Virtanen et al. 2019). In the present study area, this shoreline lies at 32–35 m a.s.l. (Ojala et al. 2013). Since sulphide sediment deposition in the BSB took place in the brackish-water phase where condition became favourably anoxic for sulphate reducing bacteria to convert plant residues and seawater sulphate into sulphides, the bottom topography during the Litorina Sea transgression has an important role (Virtanen et al. 2019; Saresma et al. 2021a). According to previous studies (Kaskela 2017; Virtasalo et al. 2020; Saresma et al. 2021a), the deposition of acid sulphide sediments was probably restricted to patchy areas of continuous sedimentation, which were deep enough (ca.

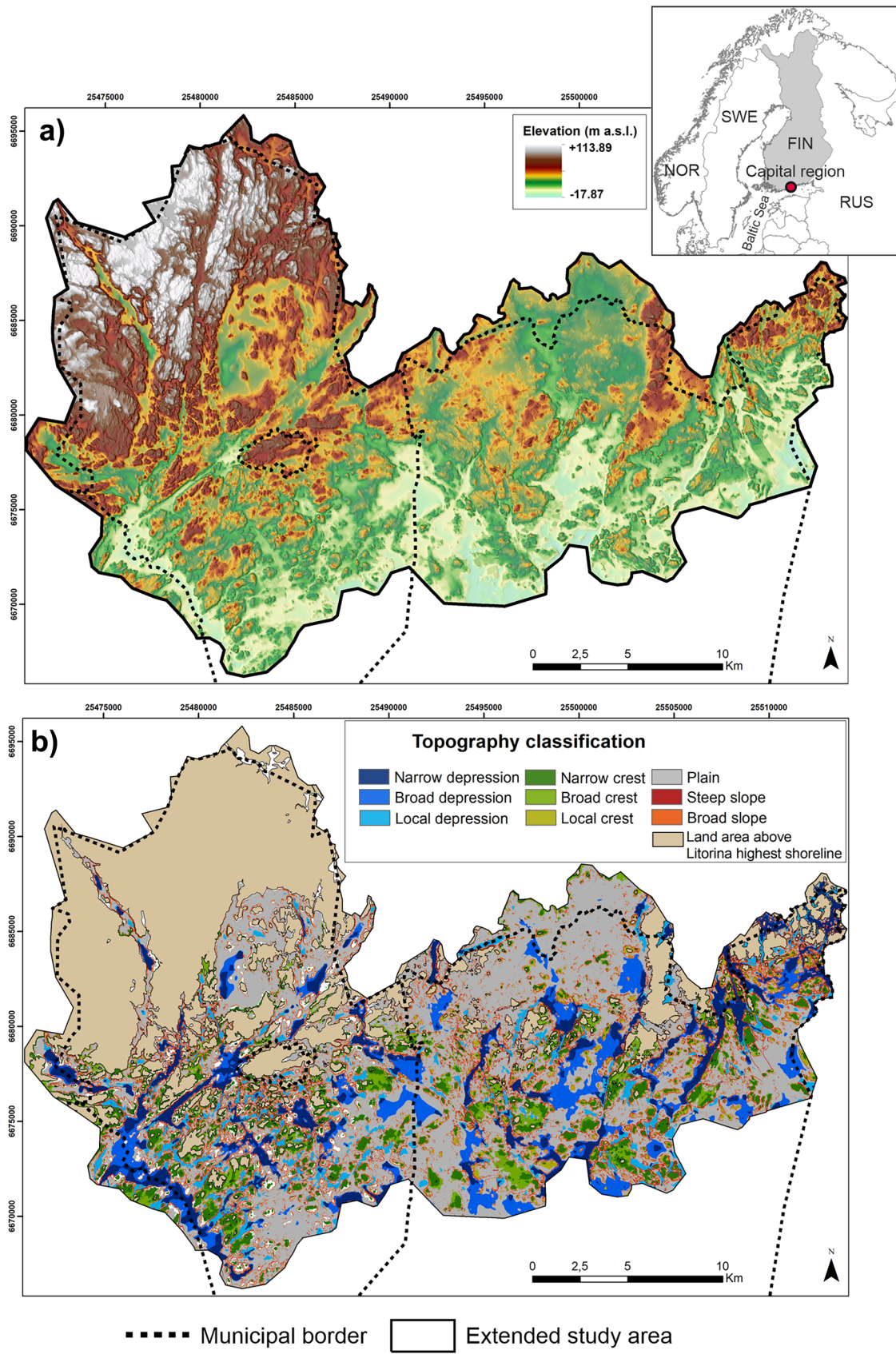
30–50 m) and sheltered from dominant south-westerly winds and waves to maintain near-bottom hypoxia.

At around 7000 BP, the highest parts of the study area in the north were above the Litorina Sea shoreline (Ojala et al. 2013), whereas most of southern sector was below sea level. The submerged topography of the lowest parts is characterized by a versatile terrain, including crests, depressions, flat plains, and low gradient slopes (Saresma et al. 2020, 2021a, b) (Fig. 1b). Narrow depressions are the deepest form of depressions, on average having the thickest layers of fine-grained sediments (Saresma et al. 2021a) (Figs. 1b and 2a). Broad depressions are larger areas that are often adjacent to narrow depressions, while local depressions represent smaller individual depressions with the thinnest clay layers on average (Saresma et al. 2021a) (Figs. 1b and 2a). The deepest water during the Litorina Sea phase at 7000 BP occurred in the southern sector (Fig. 2b), in areas that are currently situated at the lowest elevation near the coastline (Fig. 3). The locations of the deep Litorina Sea water are often associated with thick clay deposits and reveal narrow depressions in their terrain (Saresma et al. 2021a).

## Data and processing

### Available points of sulphide investigations

Sulphide investigations have been conducted at 319 points in the study area in the City of Espoo (City of Espoo, unpublished) and the City of Helsinki (Ground investigations, Geological Survey of Finland) and in the acid sulphate soil mapping data of the Geological Survey of Finland (Acid sulphate soils, Geological Survey of Finland). In addition, investigation points for sediment lithology were included from Ojala et al. (2007), Ojala (2007, 2009) and Saresma et al. (2017). The spatial distribution of the investigation points with classification into sulphide-containing and non-sulphide-containing points is presented in Fig. 3. Here, a sulphide-containing point refers to an investigation point where either potential or actual acid sulphate soil, or both, has been observed or determined at one or more sediment depth(s) below the ground surface. Investigations have been conducted in the cities by different consulting companies over the years and by the Geological Survey of Finland. The sulphide investigation methods and measurements have been manifold, but conclusions regarding whether sulphide clay is present at a study point have been based on guidelines presented by Eden et al. (2012a), the Finnish Transport Agency (2012, 2017), and Pousette (2007). We note that the methods and guidelines are not fully comparable, but the resulting conclusions from the investigations regarding the occurrence of sulphide in soil samples are comparable. Thereby, we adapted hazardous sulphide content as they were presented in each original study, without any adjustment of a specific



**Fig. 1** The study area in the capital region of Finland, including the cities of Espoo, Kauniainen, and Helsinki. Palaeotopographic maps of **a** surface elevation during the Litorina transgression at 7000 BP, and **b** BTM classification of terrain characteristics (Saresma et al. 2021a) below the highest Litorina shoreline (Ojala et al. 2013), when northern parts of the study area were already above sea level

threshold value. The present study used this enumerative data, i.e., ‘sulphide yes’ or ‘sulphide no’, as a dependent variable.

### Processing of the sulphide data

The total of 319 sulphide investigation points included 97 points for which the sample depth was at least 3 m. For the rest of the data points, samples analysed for sulphides were taken from a depth of less than 3 m. These included samples mainly taken by the Geological Survey of Finland in acid sulphate soil mapping. Overall, 123 investigation points were found to contain sulphide soils (39% of the total number of points in the study area) and 196 points (61%) did not contain sulphide. At 52 points, sulphide was detected when samples were taken from a depth of 3 m or more. At these points, sulphide-containing clay layers occurred at various depths, and the overall data indicate that sulphide clay at the sulphide-containing investigation points can occur between depths of 0.1–9.0 m in the sequences. The thinnest layers of sulphide clay occurred at points where detailed lithological descriptions had been conducted. At these points, samples were taken at 0.1-m depth intervals. Usually, samples for sulphide analysis were taken at intervals of 0.5 or 1.0 m. One investigation point in the study area entirely consisted of sulphide clay with a depth of up to 9 m, exceeding the estimated thickness of clay deposited during the Litorina Sea phase. In fact, the thickness of sulphide clay exceeded the estimated thickness of clay deposited during the Litorina Sea phase at 17 points (Fig. 4a). In most cases with sulphide-containing sediments, the sulphide was located in the topmost part of the sediment profile. There were 22 sample points (44%) with sulphide occurring in topmost parts of the sample profile, 11 points (21%) with sulphide mixed at various depths of the sample profile, 11 points (21%) with sulphide occurring at all sample depths, and 3 points (6%) with sulphide occurring at the bottom of the sample profile (Fig. 4b). In addition, at five sulphide-containing points (10%), only one depth was sampled.

### Environmental and material property variables

Sulphide investigation points were combined with environmental variables from the study area. These variables included the total clay depth at each point, the topographic class in the Litorina Sea phase, the water depth of the

Litorina Sea during the Litorina transgression at 7000 BP, and the present surface elevation at each investigation point. Material property variables included the sediment water content and organic content, which were combined with sample depth at each investigation point. Because for many of the samples analysed the loss on ignition (LOI) was measured instead of the organic content, LOI was transformed into organic content to obtain comparable variables. The transformation is typically carried out using a formula that is based on the clay fraction of the sample (Finnish Geotechnical Society 1985), but because this information was unavailable for sulphide clay samples, the transformation was performed by using the average ratio between LOI and the organic content for each fine-grained soil type. This average ratio was derived from the database of Finnish clays compiled by Löfman (2021), and a multiplier of 0.3 was used for clay soil, 0.5 for gyttja clay, 0.8 for gyttja, and 0.7 for silty clay. In addition, the sample depths of the soil profile at each investigation point were collated.

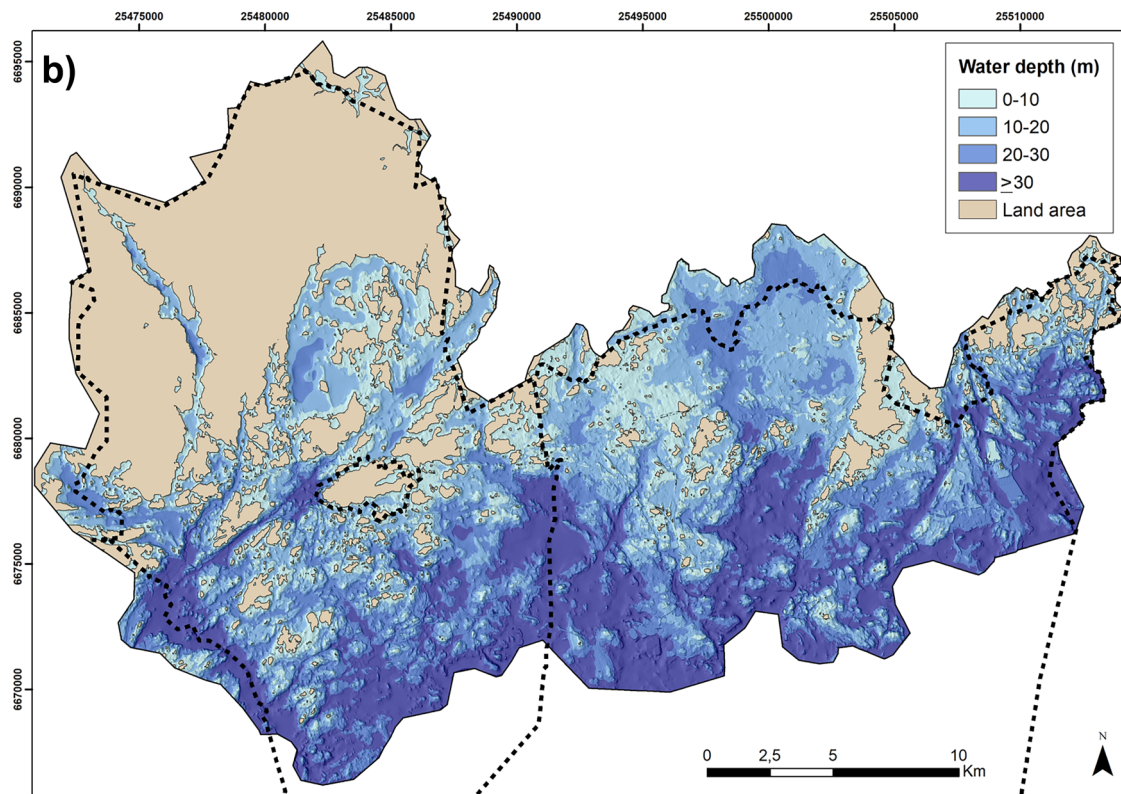
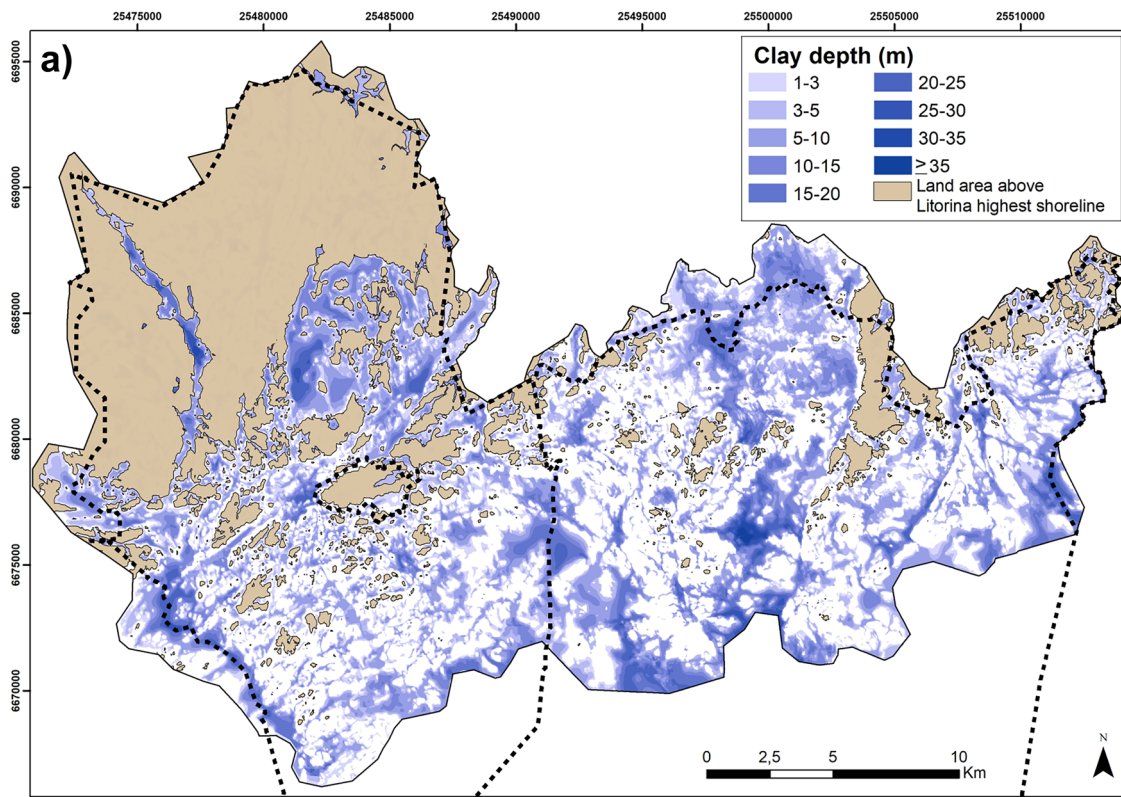
Environmental variables were manually classified into different categories following a principle according to which the nearest values were classified into the same category (Figs. 1 and 2; Tables 2, 3, 4, and 5). The organic content was classified into three categories, 1 (0–2%), 2 (2–6%), and 3 ( $\geq 6\%$ ), according to the ‘GEO’ classification standards (Korhonen et al. 1974) at each point and sample depth. Following a risk-based approach, the highest organic content was chosen to represent the amount of organic material at each point. For some of the samples, information was only available on the soil type (i.e., clay, gyttja clay), and in these cases, the organic content category was defined based on the soil type.

For statistical analyses, the data used in this study are described as follows. Point data are the data represented by variables at a specific geographic location (i.e., sulphide investigation point). Depth data are the third dimension of point data and represent material property variables at each sample depth of a soil profile. Data and statistical methods including implementation procedure are visualized with a flow chart in Fig. 5.

## Statistical methods

### Test of independence

The association between the occurrence of sulphide and the variables organic content, clay depth, topography, and Litorina water depth was determined using Pearson’s chi-squared test of independence. The test was performed between the enumerative values of sulphide and organic content ( $n = 302$ ), following sulphide and clay depth ( $n = 315$ ), sulphide and topography ( $n = 313$ ), and sulphide and Litorina water depth ( $n = 315$ ). The chi-squared test

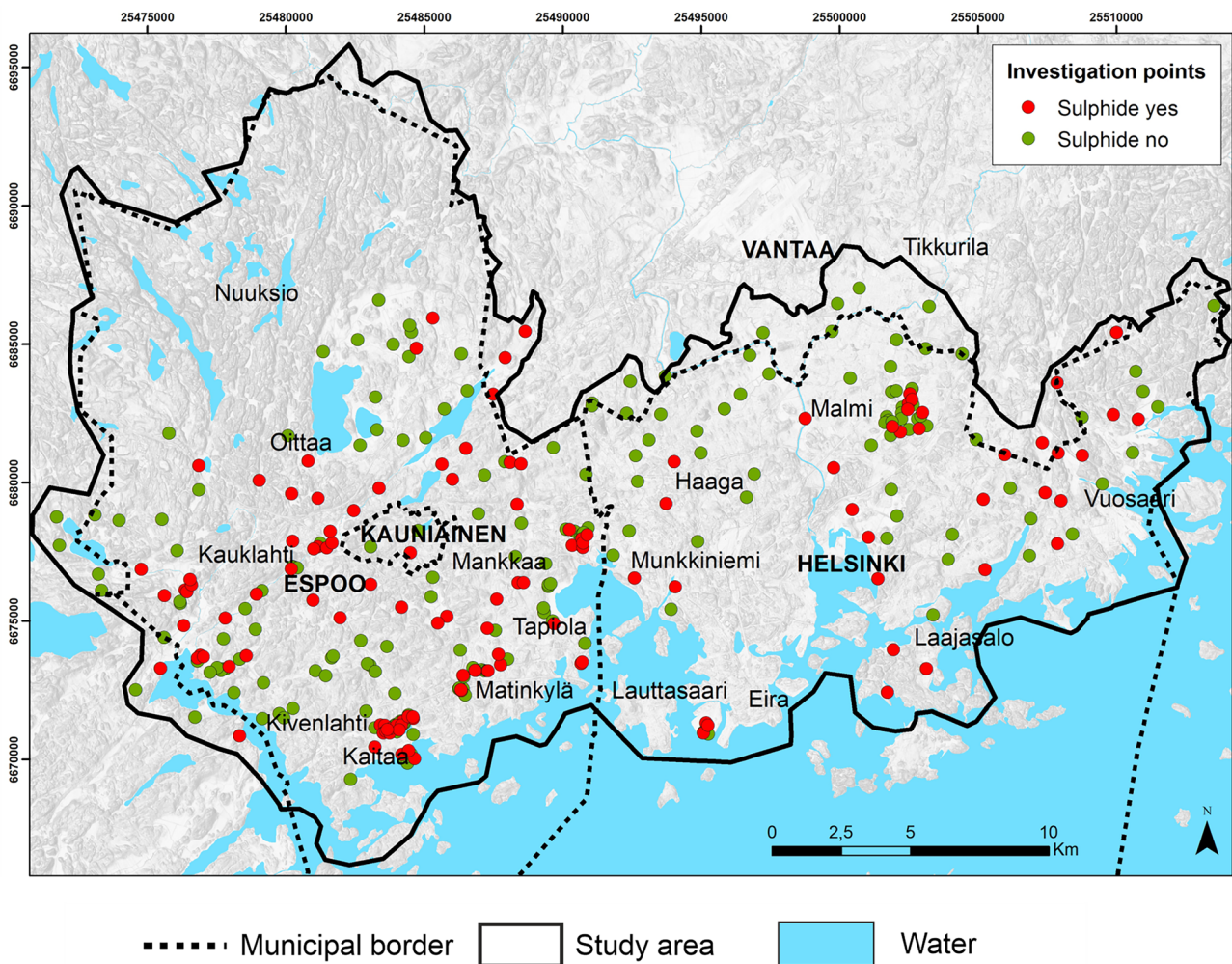


----- Municipal border     Extended study area

**Fig. 2** Maps of **a** the total clay depth presently observed in areas below the highest Litorina Sea shoreline (Ojala et al. 2013), and **b** water depth in the study area during the Litorina transgression around 7000 BP (Saresma et al. 2021a, b). Clay depth is modified after dataset Clay areas and depths © Geological Survey of Finland

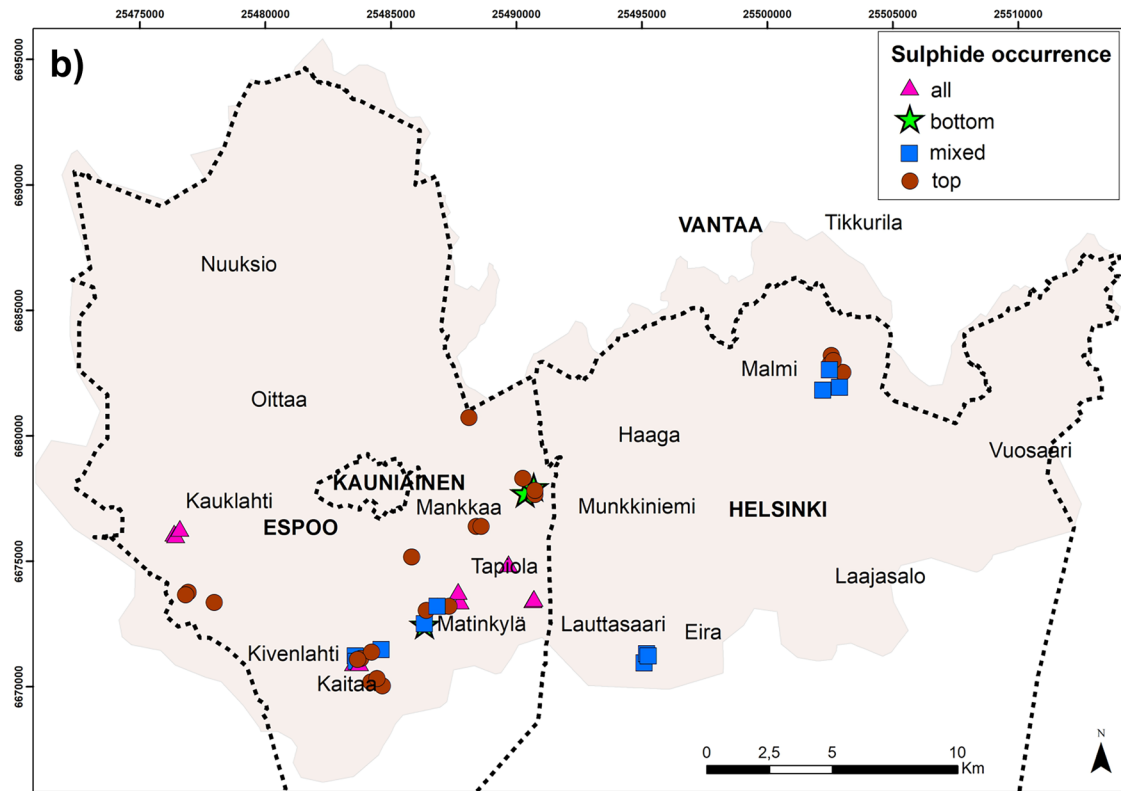
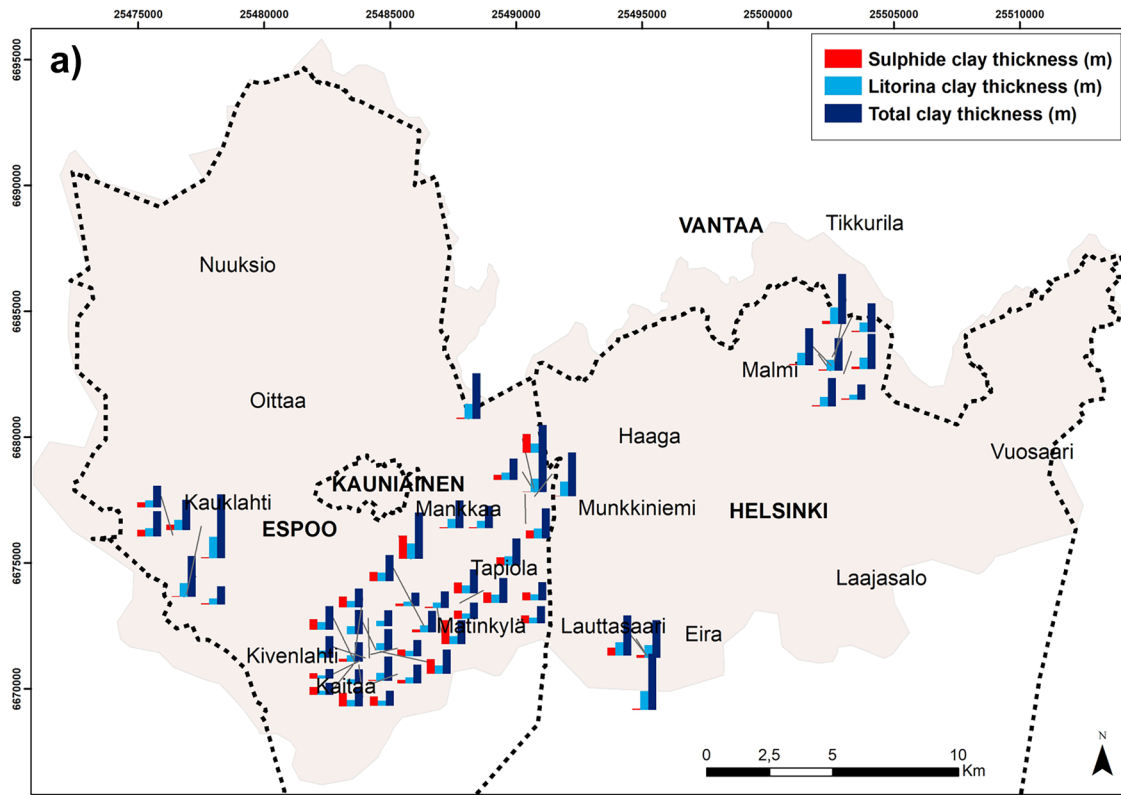
includes both a null hypothesis and an alternative hypothesis (Bilder and Loughin 2014). The null hypothesis for the sulphide data was that the variables are independent and a change in organic content, clay depth, topography, or Litorina water depth will not affect the occurrence of sulphide. The null hypothesis can be rejected if the occurrence of sulphide is statistically significantly associated with the variables. Contingency tables were utilized to classify data frequencies. The sample size was adjusted to meet the common criteria of Pearson’s chi-squared test (Bilder and Loughin 2014). To enable this, no 0 or 1 count for any

cell was allowed. Clay depth categories 5 (15–20 m) and 6 (20–25 m) were combined to form a category labelled as 5 ( $\geq 15$  m), and the topographic categories narrow crest, broad crest, and broad slope were combined to form a category named ‘other’. The categorization of ‘other’ was based on the depositional environment of topographic classes other than plains and depressions, which are the most interesting topographic locations regarding the occurrence of sulphide (e.g., Haavisto and Kukkonen 1975; Visuri et al. 2021). The test of independence was performed with a 95% confidence level by default, and the *p*-value was calculated from the test statistic to evaluate the significance of the association. Contingency table visualization was carried out using the R packages *vcd* and *mosaic*. Tests of independence and all statistical analyses in the study were carried out using the R software version 4.1.2.



**Fig. 3** The spatial distribution of sulphide clay investigation points extended over municipal borders. The points indicate the classification results, i.e., whether they contain sulphide clay (red) or not

(green). Sulphide investigation points © City of Espoo, City of Helsinki, Geological Survey of Finland



----- Municipal border    Study area

**Fig. 4** The occurrence of sulphide clay at the sampling points investigated in the present study where the sample depth was  $\geq 3$  m. The investigated points are presented by **a** comparing the thickness of the sulphide clay section with the thickness of the total clay section and the clay section deposited during the Litorina Sea phase, and **b** according to the depth at which sulphide occurred in each section when present in the sample profile

### Log of odds ratios and odds ratios

Odds ratios were calculated to determine the most prevalent combination of variable categories related to the most frequent outcome of sulphide. If the null hypothesis is rejected, odds, as a way of representing probability, can be used to measure the sources and strength of an association (Bilder and Loughin 2014). Log of odds ratios and odds ratios were used to analyse the association between the categories of a variable and in identifying statistically significant categories related to the occurrence of sulphide. Log odds ratios were calculated with 95% confidence by default using the R package *vcd*.

### Association rules

The data mining method association rule mining was used to measure the association between the enumerative variables organic content, clay depth, topography, and Litorina water depth in sulphide-bearing sediments. Sulphide-containing points ( $n = 123$ ) were selected for the analyses. Association rule mining is a method for finding associations and frequent patterns in the data (Shekhar and Chawla 2003). This dependency is expressed with a rule  $A \rightarrow B$  ( $s, c$ ). Support ( $s$ ) for the rule is the relative frequency of all transactions in the data and confidence ( $c$ ) is the strength of the rule. Association rules utilize the well-known Apriori algorithm due to its efficiency in mining large databases (Agrawal et al. 1993, Agrawal and Srikat 1994). To find important associations, which are called strong rules, the minimum support and minimum confidence must be defined (Agrawal et al. 1993; Koperski and Han 1995). The minimum support threshold was set to 0.001 and the minimum confidence threshold to 0.001. Low thresholds were set to find all useful rules from a relatively small dataset. Association rule mining was conducted using the R packages *arules* and *arulesViz*, while the package *RColorBrewer* was used to present data frequencies.

### Logistic regression

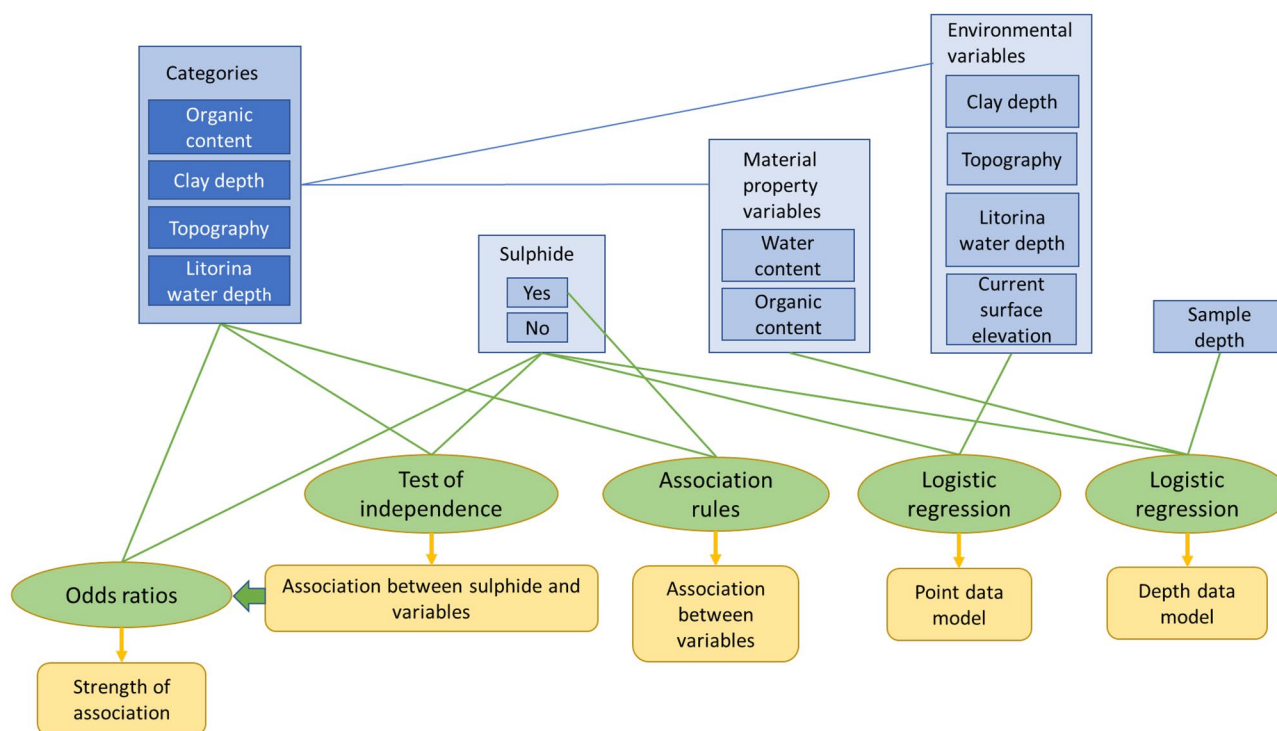
A logistic regression model was created for two complete datasets to estimate the probability of occurrence of sulphide. In contrast to linear regression, logistic regression with the generalized linear model function can be used for enumerative data with sulphide probabilities being between

0 and 1. According to James et al. (2014) and Bilder and Loughin (2014), the logistic regression function can be written as follows:

$$p(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_i x_i}}, \quad (1)$$

where the probability of sulphide occurrence is expressed as  $p$ , the predictor variable (i.e., clay depth) as  $x$  and the regression coefficients, log of odds or logit, as  $\beta$ . Logistic regression models were created for continuous variables instead of enumerative variables except topography due to its qualitative nature. The first dataset consisted of point data ( $n = 300$ ), including the continuous variables clay depth, Litorina water depth, current surface elevation, and the discrete variable topography. The second dataset consisted of depth data ( $n = 654$ ) and contained the variables organic content, water content, and sample depth from each data point. First, a full model was created for both datasets with all variables included, following the backward elimination approach (Bilder and Loughin 2014). One variable at a time was then removed to improve the model performance. Model generation was repeated four times for the point dataset and three times for the depth data. This process enabled the optimization of a model containing the most useful variables for predicting the occurrence of sulphide.

The principle of evaluating the model outcome is basically the same as in linear regression and follows the idea of assessing how much impact an independent variable has on a dependent variable. Logistic regression is based on the log-likelihood function, with comparison of the observed values of a specific variable with the predicted values obtained from the model (Hosmer et al. 2013). There are several methods to assess the fit of a model including evaluation of the model performance, tests of the significance of each variable, and measure of goodness of fit (Peng et al. 2002; Hosmer et al. 2013; Bilder and Loughin 2014). Models were created with a 95% confidence level. Model accuracy evaluation included classification of the predicted probabilities into two groups using a threshold of 0.5. A probability less than 0.5 means that there is no sulphide, whereas a probability of more than 0.5 means that sulphide occurs in a specific point or sample. Model variable performance was evaluated based on the  $p$ -value. Interpretation of the variable coefficients is based on the coding of variable levels when the model contains qualitative predictors of multiple levels. However, the prediction of the coefficients is the same, regardless of the coding used (James et al. 2014). For the variable topography, a broad depression represented the baseline and was coded 0 by default, against which the other categories were compared. The goodness of fit was assessed with the Akaike information criterion (AIC), a relative measure between



**Fig. 5** A flow chart of the data and methods used in the study

models for a given set of data (Bilder and Loughin 2014), and McFadden's pseudo  $R^2$  evaluated by the p-value. The R package *tidyverse* was used to calculate model accuracy with the function 'mean' (James et al. 2014). The R package *ggplot2* was used to plot the logistic regression models.

## Results and interpretation

### Test of independence

The results from Pearson's chi-squared test of independence are presented in Table 1. All the  $p$ -values indicating statistical significance are below 0.05, which provides evidence against independence. The null hypothesis can be rejected for all the variables, and we conclude that there is an association between each of the variables and the occurrence of sulphide. Specifically, the strongest association was found between the organic content and sulphide occurrence. The relationships between clay depth and sulphide and topography and sulphide are strong, while the association between the Litorina water depth and sulphide displays moderate significance according to Pearson's chi-squared test. Enumerative data frequencies are presented in contingency Tables 2, 3, 4 and 5 and visualized with mosaic plots in Fig. 6.

### Log of odds ratios and odds ratios

Log of odds ratios, odds ratios, and their statistical significance are presented in Table 6. An illustration of the log of odds ratios between adjacent variable categories is displayed in Fig. 7. Log of odds are normalized odds for making the data symmetrical with positive and negative values within the same order of magnitude. Statistically significant log of odds ratios and odds ratios were found for the variables organic content between categories 1 (0–2%) and 2 (2–6%) and clay depth between categories 2 (3–5 m) and 3 (5–10 m) and categories 4 (10–15 m) and 5 ( $\geq 15$  m). Minor significance of the log of odds ratios and odds ratios was recorded for the variables organic content between categories 2 (2–6%) and 3 ( $\geq 6$ ) and Litorina water depth between categories 1 (0–10 m) and 2 (10–20 m). In interpretation, odds ratios are often used to describe an increase in the likelihood of an event occurring if the odds ratio is greater than 1 or a decrease if the odds ratio is less than 1 (Bilder and Loughin 2014). We found all the odds ratios to be greater than 1, except for the variables clay depth between categories 1 (1–3 m) and 2 (3–5 m) and topography between categories 4 (narrow depression) and 5 (other), meaning that the odds of occurrence of sulphide are increasing. In the case of the variable organic content, the occurrence of sulphide is 4.3 times

**Table 1** Chi-squared test results

Variable	$X^2$	df	<i>p</i> -value
Sulphide and organic content	46.7	2	$7.39 \times 10^{-11}$
Sulphide and clay depth	17.8	5	0.0014
Sulphide and topography	13.7	4	0.0085
Sulphide and Litorina water depth	10.5	3	0.0145

$X^2$  goodness-of-fit, *df* degrees of freedom, *p*-value statistical significance

**Table 2** Combinations of categories for the variable organic content

Sulphide	Organic content (%)			Total
	1 (0–2)	2 (2–6)	3 (≥ 6)	
No	112	48	26	186
Yes	25	46	45	116
Total	137	94	71	302

as likely in category 2 as compared to category 1, while in the case of clay depth, the occurrence of sulphide is 2.7 times as likely in category 3 as compared to category 2, and 2.5 times as likely in category 5 as compared to category 4. In the case of Litorina water depth, the occurrence of sulphide is 3.4 times as likely in category 2 as compared to category 1. Finally, for topographic class, the odds ratio shows no statistically significant difference between adjacent categories, but there is an odds ratio of less than 1 and a large negative log of odds ratio between categories 4 (narrow depression) and 5 (other), indicating that very little sulphide occurs in category 5 compared to category 4.

**Association rules**

Rules produced by association rules mining were listed according to confidence to determine the high probability rules and according to support to extract the most frequent items. The selected, significant rules are presented below with the support for the rule shown as the first number between parentheses and the confidence as second number.

*Rule 1: Clay = 5(≥ 15), Topo = Narrow depression, Org = 2(2–6) → Litorina water depth = 4(≥ 30) (4.1%; 100%).*

**Table 3** Combinations of categories for the variable clay depth

Sulphide	Clay depth (m)					Total
	1 (1–3)	2 (3–5)	3 (5–10)	4 (10–15)	5 (≥ 15)	
No	12	47	84	38	11	192
Yes	6	12	57	28	20	123
Total	18	59	141	66	31	315

*Rule 2: Clay = 2(3–5), Topo = Local depression → Litorina water depth = 2(10–20) (4.1%; 100%)*

*Rule 3: Litorina water depth = 4(≥ 30) → Clay = 3(5–10) (27.6%; 50.0%).*

*Rule 4: Clay = 3(5–10) → Org = 2(2–6) (18.7%; 40.4%)*

*Rule 5: Org = 3(≥ 6) → Topo = Narrow depression (16.3%; 44.4%).*

*Rule 6: Topo = Other → Org = 1(0–2) (0.8%; 50%)*

Rules 1 and 2 have a very high confidence. In fact, according to rule 1, when the clay depth is more than 15 m and the organic content varies between 2 and 6% in a narrow depression, there is a 100% probability that the Litorina water depth is more than 30 m at the sulphide-containing point. Rule 2 states that when the clay depth is 3–5 m in a local depression, there is a 100% probability of having a Litorina water depth of 10–20 m. The frequency is not high in these cases, which means that they are not the most common cases in the data, but these variable combinations are most likely to appear in sulphide-containing locations. Rules 3, 4, and 5 represent the rules with the highest support. According to rule 3, the clay depth is often 5–10 m when the Litorina water depth is more than 30 m. This is obviously concluded from the absolute variable frequencies plot in Fig. 8, where a Litorina water depth of over 30 m and a clay depth of 5–10 m represent the most common variables in the dataset. Rule 4 shows the relationship between a clay depth of 5–10 m and an organic content of 2–6%, and rule 5 verifies the association between a high organic content (≥ 6%) and topography consisting of a narrow depression. Sometimes, a detected negation in the data is of interest, although the support for the rule is very low (Karasová 2005; Karasová et al. 2005). Rule 6 presents a case that very seldom occurs, i.e., the topographic class ‘other’ is rarely met in sulphide-containing locations, but if it appears, it is likely to appear together with an organic content of 0–2%.

**Logistic regression**

Model diagnostics for point data from the four generated models are presented in Table 7. Model 1 is a full model including all available variables and the rest of the models contain varying combinations of the variables. Due to its qualitative nature, topography is presented in model diagnostics with all the classes except for the baseline class, i.e., broad depression, although only some of the classes are considered significant.

**Table 4** Combinations of categories for the variable topography

Sulphide	Topography					Total
	1 Plain	2 Broad depression	3 Local depression	4 Narrow depression	5 Other	
No	77	37	35	35	6	190
Yes	31	22	26	42	2	123
Total	108	59	61	77	8	313

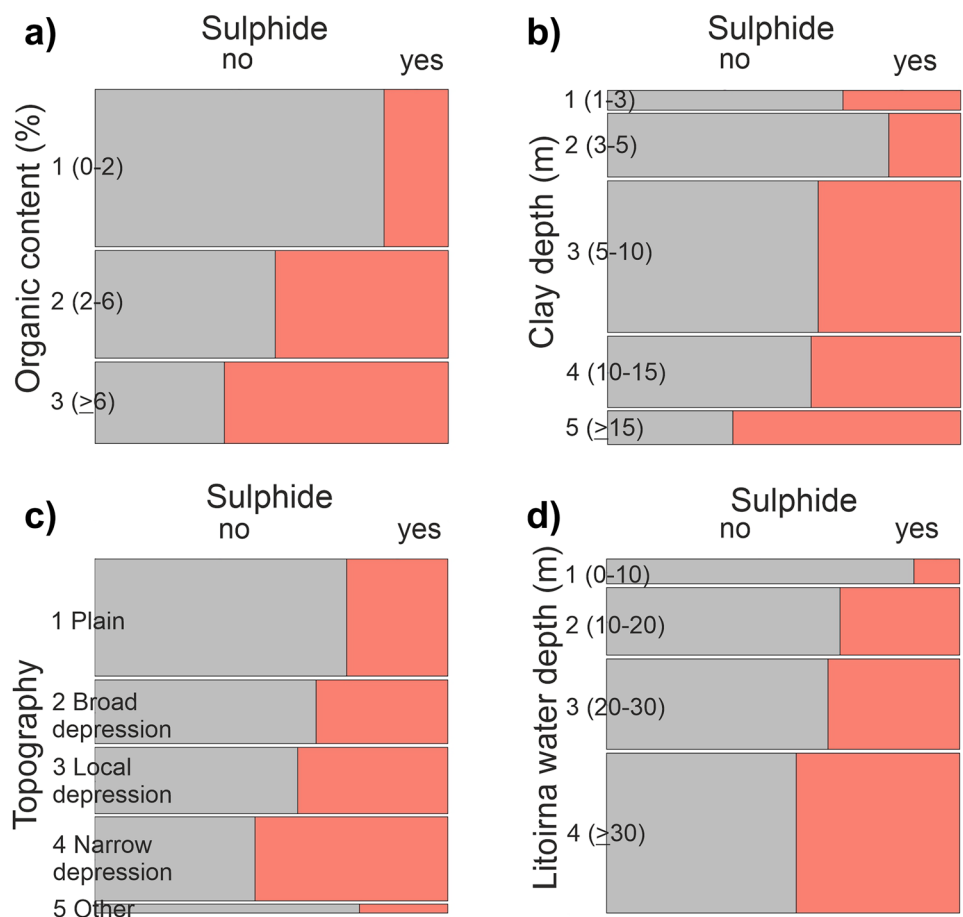
**Table 5** Combinations of categories for the variable Litorina water depth

Sulphide	Litorina water depth (m)				Total
	1 (0–10)	2 (10–20)	3 (20–30)	4 (≥30)	
No	20	41	52	79	192
Yes	3	21	31	68	123
Total	23	62	83	147	315

An overview of the results indicates that there are significant variables in each model and that no model is clearly better than all the others. The variables clay depth and topography, more specifically the class narrow depression, are both related to the occurrence of sulphide ( $p < 0.05$ ) through a positive log of odds in each model. In model 2, the variable elevation is

removed, resulting in an increase in the significance of other variables, such as Litorina water depth and local depression. The prediction accuracy of model 2 is close to good, 69% for the point data set and amount of data, with the misclassification error rate 31%. Model 2 has the smallest AIC (382.2) indicating best fit of the four models (Table 7) and pseudo

**Fig. 6** The division of variable categories for sulphide occurrence in relation to **a** organic content, **b** clay depth, **c** topography, and **d** Litorina water depth. The size of the red bar indicates the amount of sulphide in each variable category and the size of the category bar indicates the relative number of observations in each category



$R^2$  0.089 with statistically significant  $p$ -value ( $7.404 \times 10^{-6}$ ) compared with a null model. Model 3 includes a variable current elevation instead of Litorina water depth. Current elevation has a relationship with the occurrence of sulphide ( $p=0.06$ ). The prediction accuracy of model 3 (68%) is close to that of model 2, with a misclassification error rate of 32%; however, the fit of model 3 (AIC 382.4, pseudo  $R^2$  0.089,  $p=8.255 \times 10^{-6}$ ) is slightly poorer compared to model 2. For model 4, both Litorina water depth and elevation are removed, and the model includes only variables clay depth and topography. The fit of model 4 is the worst of the four models (Table 7). Logistic regression model 2 is presented in Fig. 9a with the predicted probabilities of sulphide with the actual sulphide observation data.

Model diagnostics for the depth data from three generated models are presented in Table 8. Model 1 is a full model with all variables included. For model 2, the variable sample depth is removed, and for model 3, all variables except water content are removed. The variables organic content and water content are clearly related to the occurrence of sulphide ( $p < 0.05$ ) in all models. Sample depth does not have similar impact, and, in fact, excluding it from model 2 improves the significance of the other variables. Keeping only organic content and water content in model 2 improves the model performance. The prediction accuracy for model 2 is very good, 87% with the misclassification error rate 13%. Model 2 has the smallest AIC (434.4) indicating best fit of the three models (Table 8) and pseudo  $R^2$  0.248 with statistically significant  $p$ -value (ca. 0) compared with a null model. The fit of model 3 is poor compared to other depth data models, but it indicates that water content is associated to the occurrence of sulphide. The logistic regression model for depth model 2 is presented in Fig. 9b.

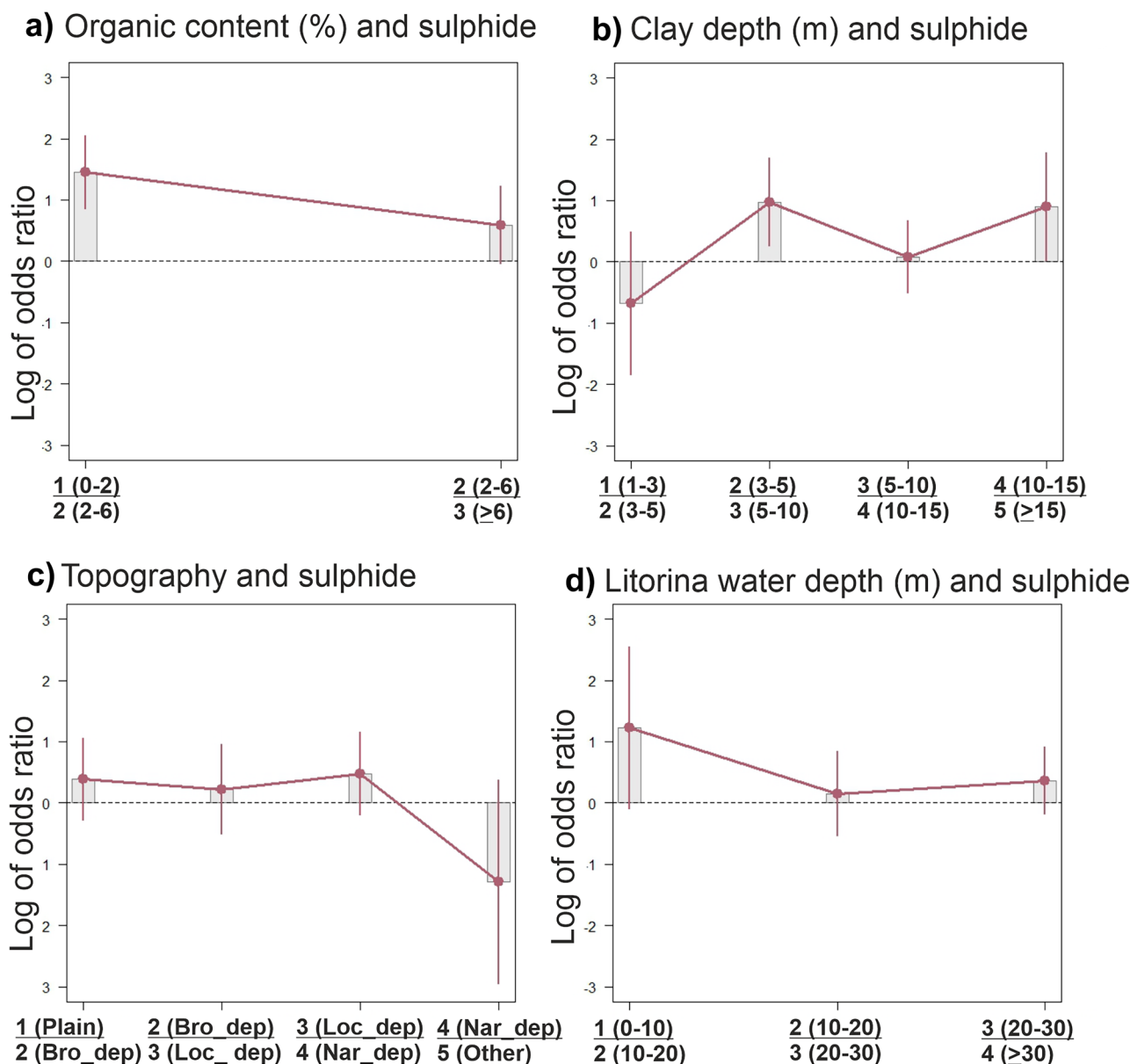
## Discussion

The present results clearly indicate a statistically significant association between the occurrence of sulphide clay and selected enumerative variables. As expected, the occurrence of sulphide was most strongly associated with the organic content of the sediment. Sulphide soils are usually associated with organic soils (e.g., Pons 1973; Erviö 1975; Larsson et al. 2007; Edén et al. 2012b; Visuri et al. 2021), and the data confirm this association for uplifted clayey sediments in southern coastal areas of Finland. More specifically, when fine-grained sediments contain  $\geq 2\%$  organic material, the odds of having sulphide in the sediments increase significantly. Furthermore, an increase in the sediment organic content to  $\geq 6\%$  increases the odds of sulphide occurrence in the sediment. The clay depth at the study location and the characteristics of the terrain topography, as descriptors of the depositional environment, are both good individual variables for predicting sulphide occurrence in the present study area. Two threshold values were identified in the clay depth data above which the odds of sulphide occurrence was found to increase significantly. These are clay depths of  $\geq 5$  m and  $\geq 15$  m, meaning that sulphide is more likely to occur at a study point with clay depth of  $\geq 5$  m compared to  $< 5$  m. Respectively, clay with a depth of  $\geq 15$  m is more likely to contain sulphide than that with a depth of  $< 15$  m. All topographic classes named as a depression (Fig. 1b) are likely to contain sulphide, but a negative association was especially detected between the classes of narrow depression and ‘other’ (e.g. crests), indicating that narrow depressions are more likely to contain sulphide than the adjacent classes. The Litorina Sea water

**Table 6** Log odds ratios, odds ratios, and  $p$ -value for each variable category

Variable	Categories, no/yes	Log of odds ratio	Odds ratio	$p$ -value
Organic content (%)	1 (0–2)/2 (2–6)	1.457	4.293	$1.46 \times 10^{-6***}$
	2 (2–6)/3 ( $\geq 6$ )	0.591	1.806	0.0658
Clay depth (m)	1 (1–3)/2 (3–5)	–0.672	0.511	0.2591
	2 (3–5)/3 (5–10)	0.977	2.658	0.0076**
	3 (5–10)/4 (10–15)	0.082	1.086	0.7853
	4 (10–15)/5 ( $\geq 15$ )	0.903	2.468	0.0450*
Topography	1 Plain/2 Broad depression	0.390	1.477	0.2558
	2 Broad depression/3 Local depression	0.223	1.249	0.5512
	3 Local depression/4 Narrow depression	0.480	1.615	0.1652
	4 Narrow depression/5 Other	–1.281	0.280	0.1309
Litorina water depth (m)	1 (0–10)/2 (10–20)	1.228	3.415	0.0688
	2 (10–20)/3 (20–30)	0.152	1.164	0.6658
	3 (20–30)/4 ( $\geq 30$ )	0.367	1.444	0.1909

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$ ; \* $p < 0.05$ ; .  $p < 0.1$



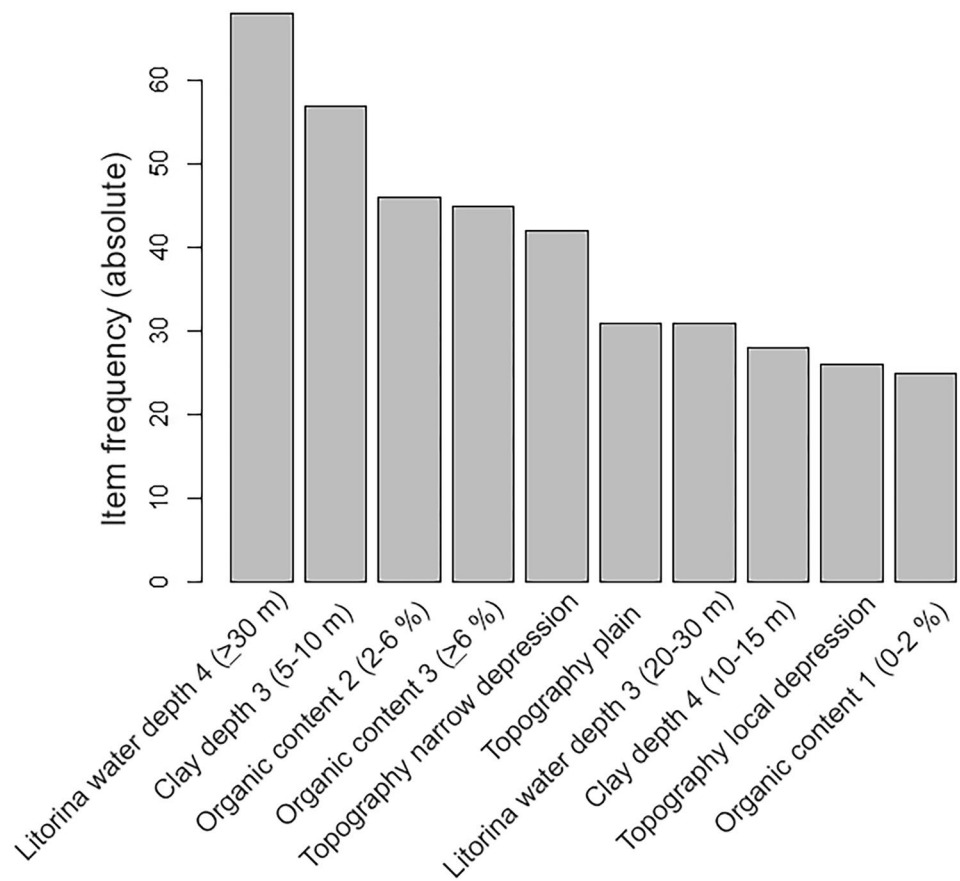
**Fig. 7** Log of odds ratios for the variable categories and sulphide occurrence. The categories for topography are Bro\_dep=broad depression, Loc\_dep=local depression, and Nar\_dep=narrow depression

depth (7000 BP) is moderately statistically significantly associated with sulphide occurrence but is not alone the most reliable predictor. The odds for sulphide occurrence increase as the Litorina water depth increases. The most significant change is between the water depth categories of < 10 m and ≥ 10 m, indicating that sulphide formation seldom occurred in shallow-water environments during the Litorina Sea phase. This observation is in accordance with many studies in the BSB region (e.g., Fonselius 1970; Holmkvist et al. 2014; Ojala et al. 2018), although Nordmyr et al. (2008) and Beucher (2015), among others, demonstrated that significant sulphide formation can also occur

in estuarine shallow waters with a high organic input and sedimentation rate.

The topographic class ‘other’ and the Litorina water depth category 0–10 m both contain only a small number of sulphide observations, and this can cause uncertainty in Pearson’s test of independence, in which a cell size ≥ 5 is usually required. Not only is sulphide clay rarely found in these categories but sulphide investigations are also often focused elsewhere. In fact, previous investigations have not been evenly distributed within the study area, but due to the location of construction projects and preliminary expectations of sulphide occurrence, they have mostly

**Fig. 8** Variable frequency plot at sulphide-containing points



focused on deep clay depressions near the coast, where the Litorina water was deep. However, according to Fig. 2a, lower clay depths appear to be more common across the study area as a whole compared to deep clay depressions.

Nevertheless, according to the data used in the study, sulphide-containing clay occurs at locations where the most frequent variable categories are Litorina water depth  $\geq 30$  m and clay depth 5–10 m. Two combinations of variables and their categories were identified as the most likely associations with sulphide occurrence. The first represents a typical depositional environment in the capital region in which organic-rich sulphide sediments were formed in deep canyons of the terrain having generally thick clay layers (Saresma et al. 2021a). The second combination represents a depositional environment in which a relatively thin layer of clay was deposited in a small depression with moderately shallow water during the Litorina Sea phase. This environment could indicate a location in the central or northern parts of the study area, where the Litorina water depth was shallower (typically 0–20 m) than in the south and some isolated depressions occurred in the terrain, called local depressions. Despite the differences, depositional environments have been favourable for sulphide formation in both cases, and a wind- and wave-sheltered environment may be

the link between these two. Usually, a sheltered environment is needed for sulphide formation, despite the differences in other deposition conditions, as shown, for example, by Kling Jonasson (2020) from western Sweden.

Fitting a regression model to continuous point data confirms the significance of the environmental variables selected for the study in predicting the occurrence of sulphide. The prediction accuracy of the model is quite close to good with the given amount of data but could be improved by more data points. The best-fitting model (model 2) includes the variables for which the coefficients of the estimates display statistical significance: clay depth, topography, more specifically the class narrow depression, and Litorina water depth. A model including the current elevation (model 3) instead of the Litorina water depth provides almost as good a fit of the data and can be used for predicting sulphide occurrence if the Litorina water depth is not known. The similar prediction accuracy of these models is due to the strong negative correlation between elevation and Litorina water depth, indicated by a Pearson correlation coefficient of  $-0.97$  (Fig. 10).

Model 2 for the depth dataset is a well-fitting regression model that results in a predicted sulphide occurrence that is close to the observed occurrence. The accuracy of the model is very good (87%). The sediment organic content

**Table 7** Model diagnostics for point models including variable evaluation and assessment of the model fit between models with the given set of data with the Akaike information criterion (AIC). The model with the lowest AIC value has the smallest prediction error leading to the best model fit

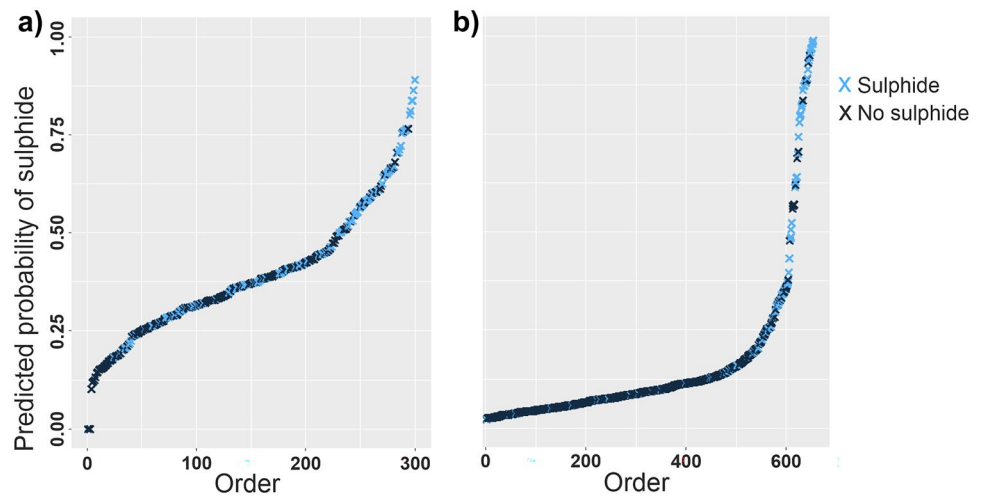
Model (point)	Variable	Coefficient of estimate	Standard error	z value	Pr (> z )	AIC
1	(Intercept)	-3.037	2.808	-1.081	0.2795	384.15
	Clay	0.108	0.046	2.335	0.0195*	
	Topography broad slope	0.033	1.222	0.027	0.9788	
	Topography crest	-13.694	837.947	-0.016	0.9870	
	Topography local depression	0.822	0.426	1.929	0.0537 .	
	Topography narrow depression	0.750	0.378	1.982	0.0474*	
	Topography plain	0.191	0.391	0.490	0.6243	
	Litorina water depth	0.044	0.085	0.520	0.6030	
	Elevation	0.013	0.084	0.152	0.8788	
2	(Intercept)	-2.620	0.618	-4.238	$2.25 \times 10^{-5}$ ***	382.18
	Clay	0.113	0.032	3.498	0.0005***	
	Topography broad slope	0.030	1.222	0.025	0.9803	
	Topography crest	-13.676	837.956	-0.016	0.9870	
	Topography local depression	0.829	0.423	1.959	0.0501 .	
	Topography narrow depression	0.750	0.378	1.983	0.0474*	
	Topography plain	0.195	0.390	0.501	0.6162	
	Litorina water depth	0.032	0.016	1.970	0.0488*	
	Elevation	0.013	0.084	0.152	0.8788	
3	(Intercept)	-1.597	0.460	-3.473	0.0005***	382.43
	Clay	0.126	0.032	3.903	$9.51 \times 10^{-5}$ ***	
	Topography broad slope	0.022	1.224	0.018	0.9855	
	Topography crest	-13.663	838.093	-0.016	0.9870	
	Topography local depression	0.841	0.424	1.985	0.0472*	
	Topography narrow depression	0.750	0.378	1.987	0.0470*	
	Topography plain	0.197	0.390	0.506	0.6128	
	Elevation	-0.030	0.016	-1.907	0.0565 .	
	Elevation	0.013	0.084	0.152	0.8788	
4	(Intercept)	-1.808	0.447	-4.043	$5.28 \times 10^{-5}$ ***	384.17
	Clay	0.123	0.032	3.852	0.0001***	
	Topography broad slope	-0.024	1.206	-0.020	0.9841	
	Topography crest	-14.270	838.572	-0.017	0.9864	
	Topography local depression	0.742	0.419	1.771	0.0765 .	
	Topography narrow depression	0.787	0.375	2.098	0.0359*	
	Topography plain	0.047	0.381	0.122	0.9027	
	Elevation	0.013	0.084	0.152	0.8788	

\*\*\* $p < 0.001$ ; \* $p < 0.05$ ; .  $p < 0.1$

is an excellent predictor of sulphide occurrence in a sample. Water content is a good predictor but requires organic content to be included in the model. That is, if the organic content is not known, the predictability of sulphide occurrence is poor based on water content alone. If both variables are known, an increase in the organic and water contents increases the probability of sulphide occurrence. This especially acts in an inverse way: if the organic content of a sample is low (e.g., < 2%), a high water content will not increase the probability of sulphide occurrence due to the higher importance of organic content in the model. For sample depth, a negative coefficient indicates that the probability of sulphide occurring will decrease when the sample depth

increases. This is reasonable, since sulphide is often located in the Litorina clay layer, which is the topmost layer of clay. Specifically, the Litorina clay section is approximated to comprise one-third of the total clay thickness (Gardemeister 1975; Ojala 2007; Ojala et al. 2007, Ojala 2009; Saresma et al. 2017), but as Fig. 4 a and b illustrate, it can vary to some extent. However, there is no statistical evidence that sample depth predicts the occurrence of sulphide in a sample profile, and thus, sample depth cannot be considered as a significant predictor. This may have been affected by many uncertainties related to sample depth, such as errors due to sampling, or sampling often being focused on the uppermost layers of soil, meaning that deep samples are rare. In

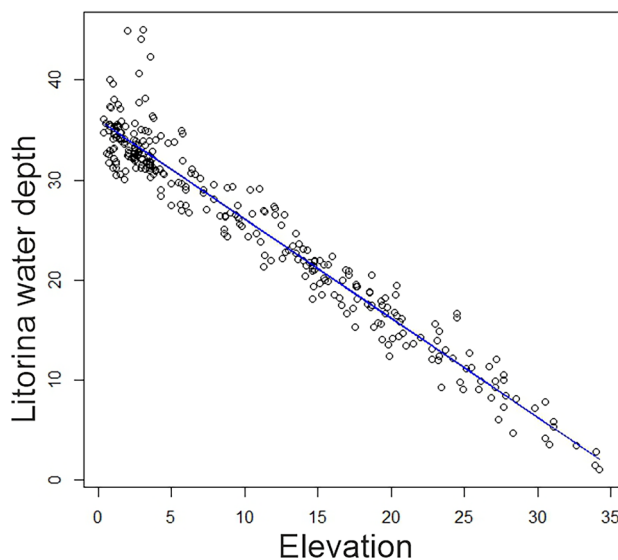
**Fig. 9** Logistic regression model for **a** point data (including clay depth, topography, and Litorina water depth) and **b** depth data (including organic content and water content). The predicted probabilities of sulphide occurrence are presented as crosses in the plot and coloured according to the actual sulphide data



**Table 8** Model diagnostics for depth models, including variable evaluation and assessment of model fit

Model (depth)	Variable	Coefficient of estimate	Standard error	z value	Pr (> z )	AIC
1	(Intercept)	-3.9663	0.5465	-7.258	$3.94 \times 10^{-13}***$	436.09
	Sample depth	-0.0212	0.0418	-0.507	0.6123	
	Organic content	0.6373	0.0791	8.060	$7.63 \times 10^{-16}***$	
	Water content	0.0143	0.0044	3.247	0.0012**	
2	(Intercept)	-4.1493	0.4172	-9.946	$< 2 \times 10^{-16}***$	434.35
	Organic content	0.650	0.0756	8.598	$< 2 \times 10^{-16}***$	
	Water content	0.0149	0.0043	3.489	0.0005***	
3	(Intercept)	-3.3866	0.3467	-9.769	$< 2 \times 10^{-16}***$	541.03
	Water content	0.0204	0.0037	5.543	$2.98 \times 10^{-8}***$	

\*\*\* $p < 0.001$ ; \*\* $p < 0.01$



**Fig. 10** The relationship between the water depth of the Litorina Sea phase and current elevation

addition, the upper sulphide layer may be eroded away in some areas, which would explain why sulphide occurs in the lowest part of the depth profile but is missing from the upper part of the profile (Fig. 4b), or the deposition conditions may have changed during BSB history so that an environment once prone to sulphide formation later changed to become an environment where sulphide no longer forms, for example due to glacioisostatic adjustment and changes in wind- and wave-sheltered environments.

Although the datasets are spatially and temporally limited in relation to the diversity of sedimentation environments of the BSB during the Litorina transgression in southern coastal Finland, they can be used for estimating the significance of variables for sulphide prediction. Knowing the most significant variables and variable combinations associated to sulphide will assist predicting dangerous sulphide soils in other countries with similar deposition environments in the BSB region. Furthermore, results could be used in assessing potential areas for developing hypoxia in the Baltic Sea, a

severe environmental issue related to topographically sheltered deep basins (e.g., Kaskela 2017; Jokinen et al. 2018; Virtanen et al. 2019). The same areas prone to hypoxia are often potential areas for sulphide formation. Based on previous studies (e.g., Virtanen et al. 2019; Virtasalo et al. 2020) and the results of the present study, the following interrelated factors site-specifically influenced the deposition of acid sulphide sediments in the BSB during the Litorina Sea phase. The water depth and bottom roughness that defined the bottom hypoxic configuration. The prevailed bottom dynamics that defined areas of sediment erosion, transportation, and accumulation. Warmer climate and water stratification combined with the appearance of sulphate reducing bacteria in near-bottom water at sedimentary basins.

Today, the estimation of challenging areas in the Baltic Sea is necessary for increased offshore planning and construction due to energy transition. Finally, the identified significant variables associated to sulphide occurrence can act as improved input parameters in machine learning which has shown advantages compared to the conventional acid sulphate soil mapping (Estévez et al. 2022). In the future, the increasing amount of sulphide investigation data from southern Finland could be used to test the model fit. A combined model with the best predictive variables from both point and depth data models, i.e., the sediment organic content, total clay depth, topographic class in the Litorina Sea phase, water depth in the Litorina Sea phase, and water content of the sediment sample, could provide the best predictive model for sulphide occurrence. Here, such combining of point and depth data was not possible because material property variables were not available for all points, and it would therefore have led to a significant decrease in the total amount of data. The accuracy of such prediction model could be further improved by incorporating Bayesian methods to account for prior knowledge of sulphide clay occurrence and their spatial autocorrelation (e.g., Han et al. 2022a). Further, the prediction of sulphide occurrence at unstudied sites could involve comparison with previously studied sites via similarity analysis (e.g., Han et al. 2022b).

## Conclusion

The adverse effects of acid sulphate soils on the environment have been known for a long time, and they are extensively studied. Recently, awareness of their harmful effects on construction and land use planning has increased, leading to an increase in the number of site-specific sulphide investigations. These investigations usually are expensive with laboratory analyses included and time-consuming, especially if the hazardous sulphide

material is found at a construction site. To focus the investigations cost-efficiently and accurately, it is necessary to collect all preliminary data related to sulphide soil in advance. This study provides information on the environmental and material property variables related to depositional environments of Litorina Sea phase, when most of the sulphide-bearing fine-grained sediments were formed. Most of these variables are available from basic geotechnical investigations via spatial analyses. The study identified the most statistically significant variables associated with the occurrence of sulphide clay, with following conclusions:

1. All categorized variables (sediment organic content, total clay depth at each point, topographic class in the Litorina Sea phase, and Litorina Sea water depth) are associated with the occurrence of sulphide clay in the present study area. The following variable categories are most prone to sulphide occurrence:
2. organic content  $\geq 2\%$  (compared to  $< 2\%$ )
3. clay depth  $\geq 5$  m (compared to  $< 5$  m) and  $\geq 15$  m (compared to  $< 15$  m)
4. topographic class narrow depression
5. Litorina water depth 10 m or more (compared to less than 10 m)
6. Sulphide clay is likely to occur with the following combination of variables: an organic-rich, thick clay section in a narrow depression with deep Litorina water or a moderate depth of clay in a local depression with a shallow Litorina water depth.
7. The best individual predictor of sulphide clay occurrence is the organic content of the sediment. Clay depth is a very good predictor, and together with the topographic class narrow depression and Litorina water depth or elevation, it can be used to predict the occurrence of sulphide. The sediment water content together with its organic content is a good sulphide predictor.

**Acknowledgements** The City of Espoo and the City of Helsinki are gratefully acknowledged for providing the sulphide investigation data used in this study. Jaakko Auri is thanked for assembling the acid sulphate soil mapping data for the study area. Jari Rantanen is thanked for professional guidance with the statistical analyses and support with software R. Two anonymous reviewers are acknowledged for their valuable comments on the manuscript.

**Author contribution** All authors participated to the study conception and design. Data collection and processing were performed by Maarit Saresma, Monica Löfman, and Emilia Kosonen. Statistical analyses were conducted by Maarit Saresma. First draft of the manuscript was written by Maarit Saresma. All authors commented the manuscript. The final manuscript was read and approved by all the authors.

**Funding** Open Access funding provided by Geological Survey of Finland (GTK, GTK Mintec).

**Data Availability** The sulphide data from the City of Helsinki and Geological Survey of Finland that support the findings of this study are openly available in Ground investigations at <https://gtkdata.gtk.fi/Pohjatutkimukset> and Acid sulphate soils at <https://gtkdata.gtk.fi/hasu>. The sulphide data from the City of Espoo is available from the City of Espoo upon reasonably request.

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Acid sulfate soils, Map Services. Geological survey of Finland. <https://gtkdata.gtk.fi/hasu/index.html> Accessed on 27 Apr 2021
- Agrawal R, Imielinski T, Swami A (1993) Mining association rules between sets of items in large databases. *ACM SIGMOD Rec* 22(2):207–216. <https://doi.org/10.1145/170036.170072>
- Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th int. conf. very large data bases, VLDB Conference Santiago, Chile 487–499
- Andersson M, Norrman T (2004) Stabilisering av sulfidjord, En litteratur- och laboratoriestudie. Arbetsrapport 33 (2004-06), Swedish Deep Stabilization Research Centre, Linköping, Sweden
- Åström M, Björklund A (1997) Geochemistry and acidity of sulphide-bearing postglacial sediments of western Finland. *Environ Geochim Health* 19:155–164
- Autiola M, Suonperä E, Suvanto S, Napari M, Nylund M, Vienonen S, Järvinen K, Forsman J, Koivulahti M, Suikkanen T, Auri J, Boman A, Mattbäck S (2022) Happamien sulfaattimaiden kansallinen opas rakennushankkeisiin: Opas happamien sulfaattimaiden huomioimiseen ja vaikutusten hallintaan. *Publ Minist Environ* 2022:3
- Autiola M, Hakanen T, Kaarankainen V, Lindroos N, Mäkelä A, Ratia K (2012) Excavation practises in sulfide clay areas, in project highway no. 8 Sepänkylä bypass Vaasa-Mustasaari. In: Österholm P, Yli-Halla, M, Edén P (eds) 7th IASSC in Vaasa, Finland 2012, Towards Harmony between Land Use and the Environment, Guide 56, Geological Survey of Finland 16
- Beucher A (2015) Spatial modeling techniques for mapping and characterization of acid sulfate soils. Dissertation, Åbo Akademi
- Bilder CR, Loughin TM (2014) Analysis of categorical data with R. Taylor & Francis Group (2015), Boca Raton
- Björck S (1995) A review of the history of the Baltic Sea, 13.0-8.0 ka BP. *Quatern Int* 27:19–40. [https://doi.org/10.1016/1040-6182\(94\)00057-C](https://doi.org/10.1016/1040-6182(94)00057-C)
- Boman A, Åström M, Fröjdö S (2008) Sulfur dynamics in boreal acid sulfate soils rich in metastable iron sulfide — the role of artificial drainage. *Chem Geol* 255(1–2):68–77. <https://doi.org/10.1016/j.chemgeo.2008.06.006>
- Boman A, Fröjdö S, Backlund K, Åström ME (2010) Impact of iso-static land uplift and artificial drainage on oxidation of brackish-water sediments rich in metastable iron sulfide. *Geochim Cosmochim Acta* 74:1268–1281. <https://doi.org/10.1016/j.gca.2009.11.026>
- Dent DL, Pons LJ (1995) A world perspective on acid sulphate soils. *Geoderma* 67:263–276
- Edén P, Auri J, Rankonen E, Martinkauppi A, Österholm P, Beucher A, Yli-Halla M (2012a) Mapping acid sulfate soils in Finland: methods and results. In: Österholm P, Yli-Halla, M, Edén P (eds) 7th IASSC in Vaasa, Finland, towards harmony between land use and the environment, Guide 56, Geological Survey of Finland 31–33
- Edén P, Rankonen E, Auri J, Yli-Halla M, Österholm P, Beucher A, Rosendahl R (2012b) Definition and classification of Finnish acid sulfate soils. In: Österholm P, Yli-Halla, M, Edén P (eds) 7th IASSC in Vaasa, Finland, towards harmony between land use and the environment, Guide 56, Geological Survey of Finland 29–30
- Ervio R (1975) Cultivated sulphate soils in the drainage basin of river Kyrönjoki. *J Sci Agric Soc Finland* 47:550–561
- Estévez V, Beucher A, Mattbäck S, Boman A, Auri J, Björck K-M, Österholm P (2022) Machine learning techniques for acid sulfate soil mapping in southeastern Finland. *Geoderma* 406:115446. <https://doi.org/10.1016/j.geoderma.2021.115446>
- Finnish Geotechnical Society (1985) GLO-85 Geotekniset laboratoriohjeet, 1. Finnish Geotechnical Society (Suomen Geoteknillinen Yhdistys SGY) and Rakentajain kustannus Oy, Luokituskoeket
- Finnish Transport Agency (2012) Sillan geotekninen suunnittelu, Sillat ja muut taitorakenteet. *Liikenneviraston ohjeita* 11/2012
- Finnish Transport Agency (2017) Eurokoodin soveltamisohje - Geotekninen suunnittelu - NCCI 7 Siltojen ja pohjarakenteiden suunnitteluohjeet 21.4.2017. *Liikenneviraston ohjeita* 13/2017
- Fonselius SH (1970) On the stagnation and recent turnover of the water in the Baltic. *Tellus* 22(5):533–544. <https://doi.org/10.3402/tellusa.v22i5.10248>
- Forsman J, Kreft-Burman K, Lindroos N, Hämäläinen H, Niutanen V, Lehtonen K (2013) Experiences of utilising mass stabilised low-quality soils for infrastructure construction in the capital region of Finland — case absoils project. In: Proceedings of 28th International Baltic Road Conference, Vilnius
- Gardemeister R (1975) On engineering-geological properties of fine-grained sediments in Finland. Dissertation, University of Turku
- Ground investigations, Map Services. Geological Survey of Finland. <https://gtkdata.gtk.fi/Pohjatutkimukset/index.html> Accessed on 4 Oct 2021
- Haavisto M, Kukkonen E (1975) Quaternary deposits in the Helsinki map-sheet area including sea bottom. *Maaperäkartan selitykset*, 2034 Helsinki, Suomen geologinen kartta, 1:100000, Geological Survey of Finland
- Hämäläinen K (2018) Sulfidisavitutkimukset Espoossa – tulosten koonti ja vertailu tavanomaisiin maanäytetutkimustuloksiin. Master's thesis, Saimaa University of Applied Sciences
- Han L, Wang L, Zhang W, Geng B, Li S (2022a) Rockhead profile simulation using an improved generation method of conditional random field. *Journal of Rock Mechanics and Geotechnical Engineering* 14(3):896–908. <https://doi.org/10.1016/j.jrmge.2021.09.007>
- Han L, Wang L, Ding X, Wen H, Yuan X (2022b) Similarity quantification of soil parametric data and sites using confidence ellipses. *Geosci Front* 13(1). <https://doi.org/10.1016/j.gsf.2021.101280>
- Holmkvist L, Kamysnyy A, Brüchert V, Ferdelman TG, Jørgensen BB (2014) Sulfidization of lacustrine glacial clay upon Holocene marine transgression (Arkona Basin, Baltic Sea). *Geochimica et Cosmochimica Acta* 142:75–94. <https://doi.org/10.1016/j.gca.2014.07.030>
- Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, 3rd edn. John Wiley & Sons, New York

- James G, Witten D, Hastie T, Tibshirani R (2014) An introduction to statistical learning with applications in R. Springer, London
- Jokinen SA, Virtasalo JJ, Jilbert T, Kaiser J, Dellwig O, Arz HW, Hänninen J, Arppe L, Collander M, Saarinen T (2018) A 1500-year multiproxy record of coastal hypoxia from the northern Baltic Sea indicates unprecedented deoxygenation over the 20th century. *Biogeosciences* 15:3975–4001. <https://doi.org/10.5194/bg-15-3975-2018>
- Karasová V, Krisp J M, Virrantaus K (2005) Application of spatial association rules for development of a risk model for fire and rescue services. In: Hauska H, Tveite H (eds) Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (ScanGIS), Stockholm 182–192
- Karasová V (2005) Spatial data mining as a tool for improving geographical models. Master's thesis, Helsinki University of Technology
- Kaskela A (2017) Seabed landscapes of the Baltic Sea: geological characterization of the seabed environment with spatial analysis techniques. Doctoral dissertation, Department of Geosciences and Geography, University of Helsinki, Geological Survey of Finland, Espoo 41
- Kling Jonasson I (2020) Acid sulphate soil in Falkenberg on the west coast of Sweden — the first discovery of active acid sulphate soil outside the Baltic basin. Master's thesis, University of Gothenburg
- Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. In: Proceedings of 4th International Symposium on Large Spatial Databases 47–66
- Korhonen K-H, Gardemeister R, Tammirinne M (1974) Geotekninen maaluokitus. Valtion teknillinen tutkimuskeskus, Geotekniikan laboratorio, Tiedonanto No. 14, Espoo
- Larsson R, Westerberg B, Albing D, Knutsson S, Carlsson E (2007) Sulfidjord – geoteknisk klassificering och odränerad skjuvhållfasthet. Report 69, Swedish Geotechnical Institute
- Larsson R (1990) Behaviour of organic clay and Gytta. Report 38, Swedish Geotechnical Institute
- Löfman MS (2021) Documentation for the clay database. Internal report 5.2.2021 (unpublished), Aalto University
- Ministry of Agriculture and Forestry and Ministry of the Environment (2011) Guidelines for mitigating the adverse effects of acid sulphate soils in Finland until 2020. Ministry of Agriculture and Forestry and Ministry of the Environment 2/2011
- Nordmyr L, Österholm P, Åström M (2008) Estuarine behaviour of metal loads leached from coastal lowland acid sulphate soils. *Mar Environ Res* 66:378–393. <https://doi.org/10.1016/j.marenvres.2008.06.001>
- Ojala AEK, Palmu J-P, Åberg A, Åberg S, Virkki H (2013) Development of an ancient shoreline database to reconstruct the Litorina Sea maximum extension and the highest shoreline of the Baltic Sea basin in Finland. *Bull Geol Soc Finl* 85:127–144
- Ojala AEK, Saresma M, Virtasalo JJ, Huotari-Halkosaari T (2018) An allostratigraphic approach to subdivide fine-grained sediments for urban planning. *Bull Eng Geol Env* 77:879–892. <https://doi.org/10.1007/s10064-016-0981-4>
- Ojala AEK, Ikävalko O, Palmu J-P, Vanhala H, Valjus T, Suppala I, Salminen R, Lintinen P, Huotari T (2007) Espoon Suurpellon alueen maaperän ominaispiirteet. Open file Report P22.4/2007/39, Geol Surv Finland
- Ojala AEK (2007) Espoon Äijänpellon savikon stratigrafia ja geokemialliset piirteet. Open file Report P22.4/2007/26, Geol Surv Finland
- Ojala AEK (2009) Hienorakeisten maalajien kerrosjärjestys: Perkkaa ja Mustalahti. Open file Report P22.4/2009/58, Geol Surv Finland
- Peng C-YJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. *J Educ Res* 96(1):3–14. <https://doi.org/10.1080/00220670209598786>
- Pons LJ (1973) Outline of genesis, characteristics, classification and improvement of acid sulphate soils. In: Dost H (ed) Acid sulphate soils, Publication 18 Vol1, Proceedings of the International Symposium, Wageningen, 3–27
- Pousette K (2007) Råd och rekommendationer för hantering av sulfidjordsmassor. Vägverket, Publikation 2007:100
- Roos M, Åström M (2005) Hydrochemistry of rivers in an acid sulphate soil hotspot area in western Finland. *Agric Food Sci* 14:24–33
- Saresma M, Kosonen E, Ojala AEK, Kaskela A, Korkiala-Tanttu L (2021a) Characterization of sedimentary depositional environments for land use and urban planning in Espoo, Finland. *Bull Geol Soc Finland* 93:31–51. <https://doi.org/10.17741/bgsf/93.1.003>
- Saresma M, Ojala AEK, Ikävalko O (2017) Hienorakeisten maalajien kerrosjärjestys ja ominaisuudet Helsingin Malmin lentokentän kaava-alueella. Open file Report GTK/957/03.02/2016, Geol Surv Finland
- Saresma M, Kosonen E, Kähkölä N, Ojala AEK, Auri J, Huusko A (2020) Espoon sulfidisavien todennäköiset esiintymisalueet. Research report GTK/391/03.02/2019, Geol Surv Finland
- Saresma M, Kosonen E, Hornborg N, Auri J, Ojala AEK (2021b) Helsingin sulfidisavien todennäköiset esiintymisalueet. Research report GTK/705/03.02/2020, Geol Surv Finland
- Shekhar S, Chawla S (2003) Spatial databases: a tour. Prentice Hall, Upper Saddle River, NJ
- Suikkanen T, Lindroos N, Autiola M, Napari M, Taipale T, Laine J, Forsman J, Auri J, Boman A (2018) Esiselvitys happamien sulfaattimaiden kartoitusmenetelmistä ja suosituksia toimenpiteiksi infrahankkeissa pääkaupunkiseudulla. Selvitys, Ramboll
- Virtanen EA, Norkko A, Sandman AN, Viitasalo M (2019) Identifying areas prone to coastal hypoxia — the role of topography. *Biogeosciences* 16(16):3183–3195. <https://doi.org/10.5194/bg-16-3183-2019>
- Virtasalo JJ, Österholm P, Kotilainen AT, Åström ME (2020) Enrichment of trace metals from acid sulfate soils in sediments of the Kvarken Archipelago, eastern Gulf of Bothnia, Baltic Sea. *Biogeosciences* 17:6097–6113. <https://doi.org/10.5194/bg-17-6097-2020>
- Visuri M, Nystrand M, Auri J, Österholm P, Nilivaara R, Boman A, Räisänen J, Mattbäck S, Korhonen A, Ihme R (2021) Maastokäyttöisten tunnistusmenetelmien kehittäminen happamille sulfaattimaille. Tunnistus-hankkeen loppuraportti, Suomen ympäristökeskuksen raportteja 43/2021
- Westerberg B, Müller R, Larsson S (2015) Evaluation of undrained shear strength of Swedish fine-grained sulphide soils. *Eng Geol* 188:77–87. <https://doi.org/10.1016/j.enggeo.2015.01.007>
- Yli-Halla M, Puustinen M, Koskiahio J (1999) Area of cultivated acid sulphate soils in Finland. *Soil Use Manag* 15:62–67