

Article

Levee Soil Stratification Based on PAM Cluster Analysis of Measured Soil Samples from Multiple Probe Drilling Sites

Haitong Zhang ^{1,2}, Xin Wang ^{1,2}, Lei Su ^{3,*}, Yuan Wei ^{4,5} and Wenhong Dai ^{6,7}

¹ School of Civil Engineering, Institute of Disaster Prevention, Sanhe 065201, China; zhanghaitong@cidp.edu.cn (H.Z.); wangxin@cidp.edu.cn (X.W.)

² Hebei Technology Innovation Center for Multi-Hazard Resilience and Emergency Handling of Engineering Structures, Sanhe 065201, China

³ Yellow River Institute of Hydraulic Research, Zhengzhou 450003, China

⁴ Yellow River Engineering Consulting Co., Ltd., Zhengzhou 450003, China; weiyuan@yrec.cn

⁵ Key Laboratory of Water Management and Water Security for Yellow River Basin, Ministry of Water Resources (Under Construction), Zhengzhou 450003, China

⁶ College of Hydraulic and Civil Engineering, Xinjiang Agricultural University, Urumqi 830052, China; wdai@hhu.edu.cn

⁷ College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China

* Correspondence: sl667651@163.com

Abstract: Accurate soil stratification is crucial for levee safety evaluation, yet limited field sampling often hinders comprehensive analysis. This study applies the Partitioning Around Medoids (PAM, also known as K-Medoids) clustering approach for levee soil stratification using data from multiple probe drilling sites. Focusing on a Yellow River levee section in China as a study case, the PAM clustering approach effectively identifies its distinct soil types and reconstructs its soil stratification by analyzing key soil properties relevant to levee seepage and stability characteristics, including coefficient of permeability, angle of internal friction, and cohesion. The resulting soil stratification, when applied to seepage and stability analyses of the levee section, yields relatively high safety factors, indicating low failure risks under design flood conditions. These analytical results align with recent monitoring records, validating the effectiveness of the approach. A sensitivity analysis on the number of clusters, the key parameter in the PAM clustering approach, demonstrates the typical existence of an optimal value balancing computational accuracy and practical interpretability. A comparison with a hierarchical clustering approach further confirms the robustness of the PAM clustering approach. This study contributes to improving levee soil stratification methodology and enhancing levee safety evaluation, particularly when dealing with limited and spatially distributed sampling data.

Keywords: soil stratification; cluster analysis; PAM; levee; soil samples; multiple sites; probe drilling; seepage analysis; stability analysis



Citation: Zhang, H.; Wang, X.; Su, L.; Wei, Y.; Dai, W. Levee Soil Stratification Based on PAM Cluster Analysis of Measured Soil Samples from Multiple Probe Drilling Sites. *Water* **2024**, *16*, 3056. <https://doi.org/10.3390/w16213056>

Academic Editors: Laura Bulgariu and Didier Orange

Received: 23 August 2024

Revised: 16 October 2024

Accepted: 23 October 2024

Published: 24 October 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As global climate change intensifies, the increased frequency of extreme weather events and flood risks has necessitated safety evaluations for numerous levees in China. A crucial component of such evaluations involves analyzing the seepage and stability characteristics of levees through numerical procedures. These procedures generally require the input of soil stratification information to allocate appropriate soil properties to each computational unit. However, in practice, field sampling often yields only limited and sparsely distributed data for a given levee. Consequently, this leads to the additional challenge of inferring comprehensive soil stratification information for the entire levee based on these discrete data.

Traditional methods of soil stratification, primarily relying on visual inspection and expert judgment, are often subjective and may overlook subtle variations in soil properties. To address these limitations, researchers have explored various advanced statistical and computational techniques. Cao and Wang (2013) and Wang et al. (2020) resorted to Bayesian frameworks [1,2], offering a probabilistic approach to soil classification. Wang et al. (2018), Hu and Wang (2020), and Gong et al. (2020) investigated the use of random field methods and stochastic approaches [3–5], aiming to capture spatial variability and uncertainties in soil properties. Shi and Wang (2021) developed a machine learning approach based on the radial basis function network [6] to identify complex patterns in soil data. However, these advanced techniques often share common limitations: the need for extensive site investigation data, complex modeling or training processes, and high computational intensity.

Cluster analysis has emerged as a particularly promising approach for soil stratification. This method, capable of working with either small-scale or large-scale datasets, aims to group similar soil samples together based on their measured properties, potentially revealing natural divisions in a soil profile. It is especially suitable for soil stratification problems due to its ability to handle multidimensional data and identify inherent structures without prior assumptions about the nature of soil strata. Several researchers, including Hegazy and Mayne (2002), Liao and Mayne (2007), Cai et al. (2009), and Molina-Gómez et al. (2021), have investigated hierarchical clustering approaches for soil stratification [7–10], demonstrating the potential to identify soil strata objectively and detect subtle transitions.

Building upon these insights, this study also employs a clustering approach for levee soil stratification but introduces the use of Partitioning Around Medoids (PAM, also known as K-Medoids) clustering instead of hierarchical clustering. PAM offers several advantages over hierarchical clustering. Unlike hierarchical clustering, which requires either agglomerative or divisive processes, PAM provides greater flexibility in finding optimal cluster configurations [11]. This fundamental difference allows PAM to overcome limitations associated with the step-wise nature of hierarchical approaches, potentially leading to more globally optimal solutions. Moreover, PAM offers better interpretability of clustering results, as it provides actual data points as cluster centers (so-called “medoids”). It also demonstrates less sensitivity to outliers and noise [12,13], enhancing its robustness in dealing with real-world soil data. Additionally, various techniques can be employed to accelerate PAM [14–16], making it more efficient for larger datasets. This novel application of PAM clustering to levee soil stratification aims to overcome the limitations of previous approaches, potentially offering a more robust and practical approach for engineering applications.

In addition to the application of PAM clustering, this study differs from previous works by Hegazy and Mayne (2002), Liao and Mayne (2007), Cai et al. (2009), and Molina-Gómez et al. (2021) [7–10] in two other significant aspects:

1. This study examines the effectiveness of the PAM clustering approach using data from multiple sites, aiming to achieve cross-site soil stratification. Rather than focusing on a single engineering site, this multi-site consideration enhances the applicability of the approach to larger-scale levee systems and provides insights into the spatial variability of soil properties across different locations.
2. The data used in this study are not derived from in-situ penetration tests but from laboratory tests on samples obtained through probe drilling. Despite involvement with some sample disturbance, this method is common in engineering practice and provides a wide variety of data. Verifying the feasibility of the PAM clustering approach for this type of data source is thus highly significant for practical applications in levee engineering.

This study applies the PAM clustering approach to levee soil stratification, and the present paper details its methodology and findings using a section of the right-bank levee along the Yellow River in Zhongmu, Zhengzhou City, Henan Province, China, as a study case. The paper begins by introducing the problem and materials and detailing the study

case and the collected soil data. It then elaborates on the PAM clustering approach for levee soil stratification and its implementation. Following this, the results of the PAM clustering analysis are presented, including the identified soil types and their spatial distribution. The practical application of these stratification results is demonstrated through seepage and stability analyses of the levee. The paper also includes a discussion on the sensitivity of the approach to its key parameter (number of clusters) and compares the results with those from a hierarchical clustering approach. Finally, the study concludes by summarizing the main findings and their implications for levee engineering practice. Through this comprehensive analysis, this study aims to contribute to the ongoing research in levee soil stratification methodology and potentially provide useful insights for improving levee safety evaluations.

2. Problem and Materials

To illustrate the challenge inherent in inferring comprehensive soil stratification from limited field sampling data, a real-world case is presented. This case concerns a section of the right-bank levee along the Yellow River in China, specifically the Zhongmu segment near Zhengzhou City, Henan Province, at chainage 35+000. The Zhongmu segment plays a crucial role in flood protection for the densely populated surrounding area. Like many levees along the Yellow River, this section exhibits a complex geological profile resulting from historical flooding patterns and sediment deposition. The soil composition at this location is characterized by heterogeneous strata, reflecting the dynamic nature of the fluvial environment over time.

The profile of the levee section in this case is illustrated in Figure 1. To analyze the seepage and stability characteristics of the levee, it was necessary to determine the properties of the levee soil. For this purpose, three probe drilling sites were established, labeled I, II, and III, as shown in Figure 1. Field sampling was conducted at multiple points of varying depths along each site, denoted by local labels (1), (2), (3), etc., also depicted in Figure 1. Laboratory testing of the samples revealed that the dry unit weight (γ_d) and porosity (p) of the levee soil were relatively consistent across different locations, with $\gamma_d \approx 15.2 \text{ kN/m}^3$ and $p \approx 42.7\%$. Given this uniformity, the seepage and stability characteristics of the levee can be primarily attributed to three key parameters: the coefficient of permeability (k), the angle of internal friction (ϕ), and the cohesion (c). The measured values of these parameters for the sampling points across all sites are presented in Table 1.

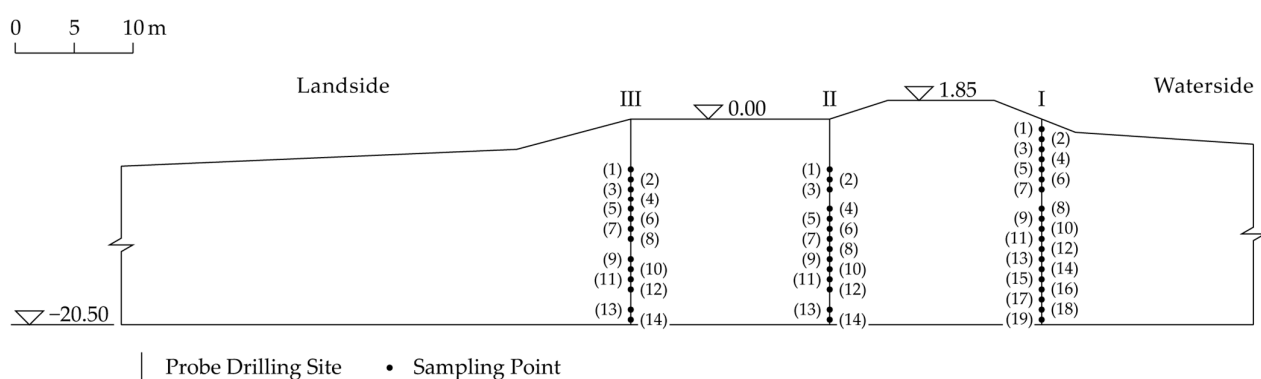


Figure 1. Profile of the levee section with sampling locations in the study case.

To ensure the representativeness and reliability of the soil data used in this study, several measures were taken during the sampling and testing processes. The three probe drilling sites (I, II, and III) were strategically located to cover different parts of the levee cross-section, providing a comprehensive representation of the soil profile. The sampling depths were carefully selected to capture variations in soil properties throughout the vertical profile, with more frequent sampling in zones where rapid changes in soil properties were anticipated based on prior geological knowledge of the area.

Table 1. Measured coefficient of permeability (k), angle of internal friction (φ), and cohesion (c) of the soil samples in the study case.

Global Sample No.	Site No.	Local Sample No.	Sampling Depth (m)	Soil Properties		
				k (cm/s)	φ (deg)	c (kPa)
1	I	(1)	1.0	2.03×10^{-6}	24.0	11.2
2		(2)	2.0	2.03×10^{-6}	24.3	11.0
3		(3)	3.0	3.68×10^{-7}	24.1	11.4
4		(4)	4.0	2.57×10^{-7}	24.8	12.2
5		(5)	5.0	2.98×10^{-7}	25.2	13.3
6		(6)	6.0	2.68×10^{-7}	24.3	13.1
7		(7)	7.0	3.47×10^{-7}	24.5	13.2
8		(8)	9.0	7.70×10^{-6}	24.5	13.2
9		(9)	10.0	3.51×10^{-7}	24.5	13.5
10		(10)	11.0	3.49×10^{-7}	24.5	13.4
11		(11)	12.0	2.98×10^{-7}	24.5	15.3
12		(12)	13.0	3.99×10^{-7}	24.5	14.6
13		(13)	14.0	4.87×10^{-7}	18.6	14.5
14		(14)	15.0	4.29×10^{-6}	18.7	15.5
15		(15)	16.0	3.65×10^{-6}	19.2	16.3
16		(16)	17.0	5.99×10^{-6}	23.2	14.3
17		(17)	18.0	3.81×10^{-6}	15.7	27.4
18		(18)	19.0	2.08×10^{-7}	15.6	27.6
19		(19)	20.0	3.56×10^{-7}	22.4	12.2
20	II	(1)	5.0	2.67×10^{-7}	24.8	12.2
21		(2)	6.0	3.69×10^{-7}	25.2	13.3
22		(3)	7.0	6.40×10^{-6}	24.3	13.1
23		(4)	9.0	2.30×10^{-6}	24.8	12.2
24		(5)	10.0	3.48×10^{-7}	21.6	11.7
25		(6)	11.0	3.52×10^{-7}	23.4	11.9
26		(7)	12.0	4.19×10^{-7}	20.2	12.2
27		(8)	13.0	4.68×10^{-7}	21.1	13.8
28		(9)	14.0	4.07×10^{-7}	23.4	11.9
29		(10)	15.0	3.61×10^{-7}	19.1	12.4
30		(11)	16.0	4.17×10^{-7}	18.7	15.5
31		(12)	17.0	3.92×10^{-7}	23.2	13.2
32		(13)	19.0	3.99×10^{-7}	15.7	28.4
33		(14)	20.0	5.60×10^{-6}	20.0	13.4
34	III	(1)	5.0	3.81×10^{-6}	15.7	28.6
35		(2)	6.0	4.82×10^{-6}	11.5	17.8
36		(3)	7.0	4.68×10^{-6}	22.0	14.1
37		(4)	8.0	2.65×10^{-7}	22.0	14.1
38		(5)	9.0	3.89×10^{-7}	21.5	13.1
39		(6)	10.0	6.10×10^{-6}	21.2	13.7
40		(7)	11.0	3.69×10^{-7}	21.2	13.7
41		(8)	12.0	3.79×10^{-7}	18.5	30.1
42		(9)	14.0	4.59×10^{-6}	18.5	30.1
43		(10)	15.0	2.98×10^{-7}	18.2	13.5
44		(11)	16.0	4.20×10^{-7}	23.0	12.8
45		(12)	17.0	3.11×10^{-6}	16.3	12.8
46		(13)	19.0	3.88×10^{-6}	18.5	29.5
47		(14)	20.0	3.88×10^{-6}	16.3	12.8

The soil samples were collected using standardized drilling and sampling techniques to minimize disturbance and ensure consistency. Upon retrieval, the samples were immediately sealed to preserve their in-situ moisture content and transported to the laboratory under controlled conditions. Laboratory tests were conducted in accordance with stan-

dards. Each test was performed in triplicate to assess repeatability, with the average values reported in Table 1. The coefficient of variation for each property was typically below 10%, indicating good consistency in the measurements.

While the dataset used in this study is relatively small (47 samples), it represents the typical level of data availability in many practical levee engineering projects. Yet, we acknowledge that a larger dataset, potentially including data from additional sites or non-invasive geophysical surveys, could further enhance the robustness of the analysis.

To assign reasonable values of k , ϕ , and c to locations within the levee section where no samples were obtained, it is necessary to infer the soil stratification for the entire levee section based on the measured k , ϕ , and c data presented in Table 1. This inference process, which aims to extrapolate from limited point measurements to a comprehensive understanding of the internal structure of the levee section, represents the challenge that this study addresses. The successful resolution of this challenge is crucial for accurate seepage and stability analyses of the levee system.

3. Approach

To address the challenge of levee soil stratification, this study employs the PAM clustering approach. PAM was chosen for its flexibility in finding optimal cluster configurations, interpretability of results by using actual data points as cluster centers (i.e., medoids), robustness to outliers and noise in the data, and potential for acceleration through various computational techniques, making it particularly suitable for soil parameter analysis. PAM effectively groups similar soil samples based on their measured properties, allowing for a data-driven identification of soil strata without relying on preconceived geological assumptions. The PAM clustering approach for levee soil stratification can be implemented as follows:

Firstly, determine the number of clusters (K). This initial step involves specifying the number of cluster medoids to be selected from the total N data points. The value of K represents the number of distinct soil types expected to be identified within the levee soil samples.

Secondly, standardize the data of the soil samples to provide a uniform scale for different properties of the levee soil. This standardization process ensures that all variables contribute equally to the subsequent clustering analysis, regardless of their original scales and units. The standardization formula applied is as follows:

$$\xi_i^x = \frac{x_i - \mu^x}{\sigma^x}, \quad (1)$$

where ξ_i^x and x_i represent the standardized and original values, respectively, of the property x ($= k, \phi, c$, in this study) for the data point i ($= 1, 2, \dots, N$); μ^x is the average of that property across all samples; and σ^x is the standard deviation. This step is crucial as it prevents properties with larger numerical ranges from dominating the clustering process, thereby allowing for a more balanced and accurate representation of the levee soil characteristics in the analysis.

Then, proceed with the core step of PAM to identify the “optimal” clustering configuration. This step involves calculating the “distance” between a data point assigned to a cluster and the medoid of that cluster. PAM supports various distance measures (which quantify dissimilarity between pairwise data points). Following comprehensive numerical testing, this study adopts the squared Euclidean metric as the distance measure, defined as follows:

$$\delta_{mn} = \sum_x (\xi_m^x - \xi_n^x)^2, \quad (2)$$

where m denotes the medoid of a cluster; and n denotes any other data point assigned to that cluster, excluding the medoid itself (note that the “Global Sample No.” in Table 1 can be used to uniquely identify each data point in the dataset, including the medoids).

The criterion that quantifies the quality of a clustering configuration, termed “cost” (C), is defined as follows:

$$C = \sum_m \left(\sum_n \delta_{mn} \right). \quad (3)$$

The objective of this step is to find a clustering configuration that minimizes the cost.

However, it has been proven that the PAM problem is NP-hard [11], meaning that it is at least as hard as the hardest problems in NP, and no known polynomial-time algorithm exists to solve it optimally. This implies that finding the globally optimal configuration may be computationally infeasible. Nevertheless, a locally optimal configuration, which approximates the globally optimal configuration, can typically be obtained through the following greedy heuristic algorithm [17], which has a time complexity of $O(K(N - K)^2)$ per iteration:

1. Randomly select K data points from the total N data points as the initial cluster medoids.
2. For each non-medoid data point n , calculate its distance δ_{mn} to every cluster medoid m using Equation (2) and assign n to the cluster with the nearest medoid (i.e., the medoid with the smallest δ_{mn}). This process forms the initial clustering configuration.
3. Calculate the cost C of the initial clustering configuration using Equation (3).
4. For each medoid m and each non-medoid data point n :
 - a. Swap the positions of m and n , making n the new medoid.
 - b. Recalculate the distances between each non-medoid data point and all the medoids, then reassign each to its nearest medoid, forming a new clustering configuration (as in Step 2).
 - c. Calculate the cost of this new configuration (as in Step 3).
 - d. If the cost has decreased, accept the swap; otherwise, revert to the previous configuration.
5. Repeat Step 4 until no further improvement in cost can be achieved (note that for big-data problems, a predetermined maximum number of iterations may be set to limit computational time).
6. The final configuration is considered “optimal” and designated as the solution. The algorithm terminates.

The efficiency of this algorithm can be significantly enhanced through various advanced techniques. Schubert and Rousseeuw (2019; 2021) proposed methods that can achieve an $O(K)$ -fold speedup [14,15]. Alternatively, the BanditPAM algorithm introduced by Tiwari et al. (2020) can reduce the time complexity to $O(N \log N)$ per iteration while maintaining result quality with high probability [16]. However, implementing these advanced techniques requires additional computational complexity and expertise. Given the relatively small dataset in the study case, comprising only 47 soil samples, runtime efficiency is not a primary concern. Consequently, the original PAM algorithm was implemented without incorporating any speedup techniques, ensuring straightforward implementation and clear interpretation of results.

Figure 2 depicts the workflow of the above PAM clustering approach in this study, illustrating its key operations and decision points. To execute this approach, an R script was developed by the authors for the present study, employing the “pam” function from the “cluster” package [18].

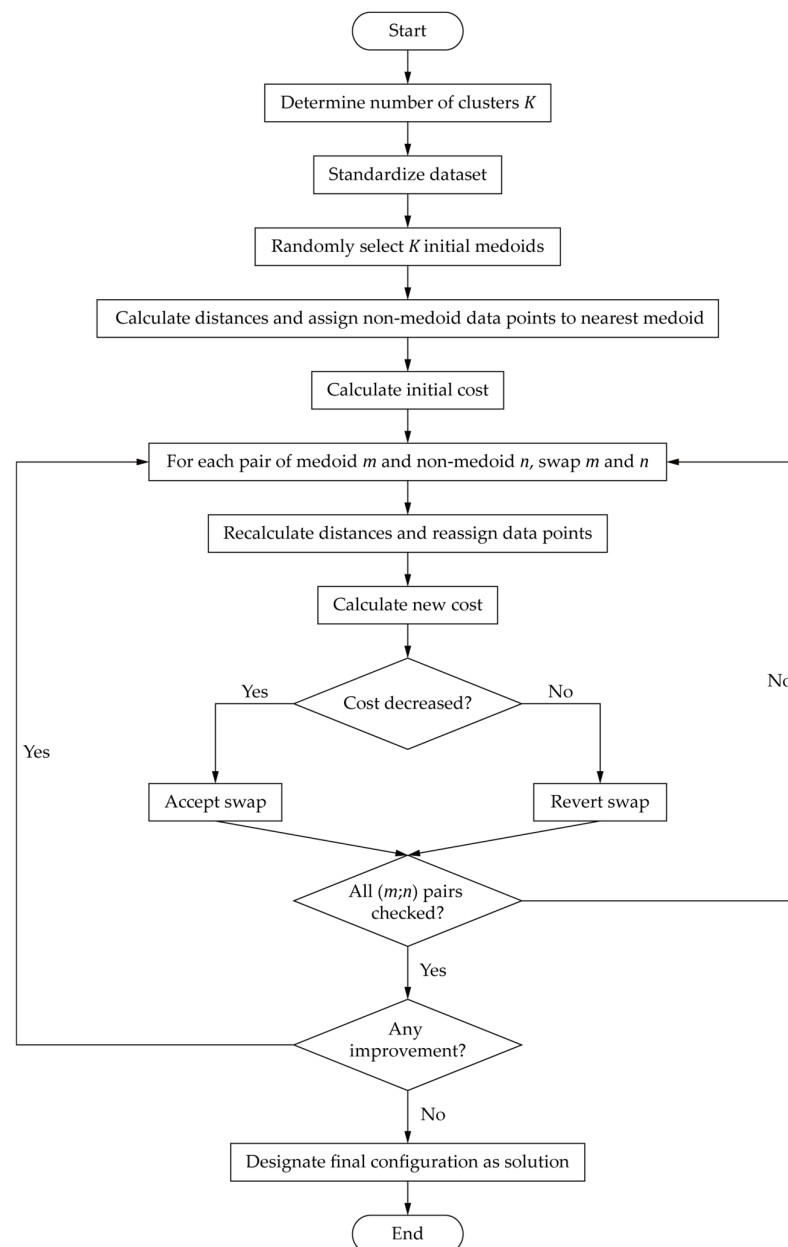


Figure 2. Workflow of the PAM clustering approach.

4. Results

4.1. PAM Clustering Results and Derived Levee Soil Stratification

The PAM clustering approach described in Section 3 was applied to the soil data presented in Table 1, which were obtained through field sampling and laboratory testing at the Zhongmu segment of the Yellow River levee, as detailed in Section 2. The number of clusters (K) was set to six, a value determined to optimize the balance between computational accuracy and practical interpretability, which will be further elaborated in Section 5.1. The cluster analysis aimed to identify distinct soil types within the levee section based on the three key parameters: coefficient of permeability (k), angle of internal friction (φ), and cohesion (c).

The resulting clusters, representing distinct soil types, are designated by the letters A through F. The attributes of these clusters are summarized in Table 2, presenting their respective medoids, the data points (including the medoid) assigned to each cluster, and the arithmetically averaged values of k , φ , and c across all soil samples within each cluster. These averaged values of k , φ , and c are adopted to represent the properties of correspond-

ing soil types. This outcome provides data-driven insights into the spatial distribution and characteristics of different soil types for the study case.

Table 2. Attributes of the clusters ($K = 6$) in the study case.

Cluster (Soil Type)	Medoid *	Data Points *	Averaged Soil Properties		
			k (cm/s)	φ (deg)	c (kPa)
A	7	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 19, 20, 21, 23, 25, 28, 31, 44	6.24×10^{-7}	24.2	12.7
B	16	8, 16, 22, 33, 36, 39	6.08×10^{-6}	22.5	13.6
C	26	13, 24, 26, 27, 29, 30, 37, 38, 40, 43	3.82×10^{-7}	20.2	13.5
D	32	18, 32, 41	3.29×10^{-7}	16.6	28.7
E	34	17, 34, 42, 46	4.02×10^{-6}	17.1	28.9
F	47	14, 15, 35, 45, 47	3.95×10^{-6}	16.4	15.0

* Identified by the “Global Sample No.” in Table 1.

Based on the clustering results presented in Table 2, a soil stratification for the levee section in the study case was reconstructed, as illustrated in Figure 3. This reconstruction assumes that the boundary between two vertically adjacent soil bodies along each site, represented by the corresponding soil samples taken from different depths, lies midway between the sampling depths. Adhering to this principle, soil bodies of the same type across different sites were interconnected to establish continuous strata where feasible.

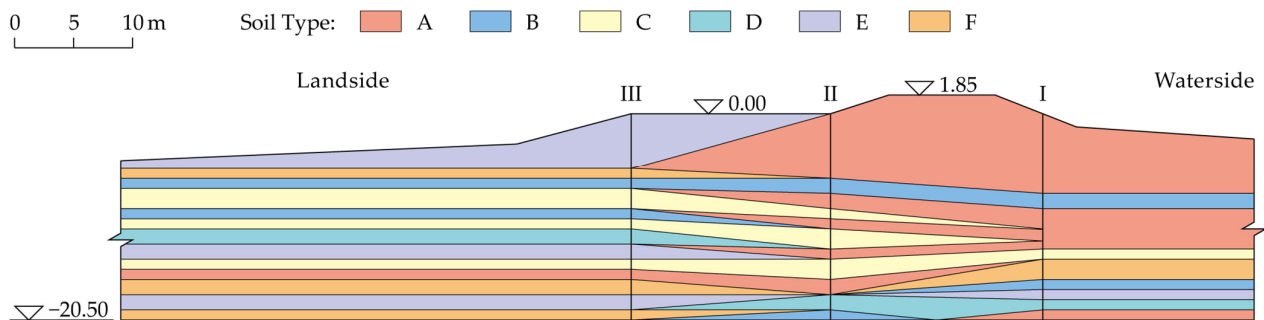


Figure 3. Soil stratification for the levee section in the study case.

Note that the reconstruction of soil stratification based on discrete sampling points, as shown in Figure 3, inevitably involves some degree of subjective judgment. One might interpret the data slightly differently, potentially resulting in minor variations in the exact boundaries between soil types or in the continuity of certain strata. However, sensitivity analysis indicates that such small variations in the interpreted soil stratification do not significantly impact the seepage and stability analyses of the levee section. The seepage and stability behavior of the levee depends primarily on the clustering results presented in Table 2, which form the foundation of the soil stratification. These clustering results, derived from the PAM approach, are objective and data-driven, independent of subjective interpretations. They capture the essential properties and distribution of soil types within the levee.

Moreover, the consistency of the clustering results across different numbers of clusters and their agreement with an alternative hierarchical clustering method (see Section 6) further demonstrate the robustness of the soil type identification. This stability in the core results ensures that, while minor subjective judgments may be necessary for visualizing the complete soil stratification, the fundamental understanding of the internal structure of the levee and the subsequent engineering analyses remain reliable and consistent.

This soil stratification shown in Figure 3 unveils a complex internal structure that significantly influences the seepage and stability behavior of the levee. Key findings include the following:

1. Type-A soil, the predominant material comprising the majority of the levee body, exhibits moderate permeability ($k = 6.24 \times 10^{-7}$ cm/s) and strength parameters ($\phi = 24.2^\circ$, $c = 12.7$ kPa). This widespread distribution contributes substantially to the overall structural integrity of the levee.
2. Type-B soil, dispersed throughout the levee with a continuous layer in the upper-middle portion, possesses the highest permeability ($k = 6.08 \times 10^{-6}$ cm/s) among all identified types. This characteristic potentially creates preferential seepage pathways, necessitating particular attention in seepage analyses.
3. Type-C soil, concentrated in the middle section and forming a relatively impermeable continuous layer in the lower-middle part, demonstrates the second-lowest permeability ($k = 3.82 \times 10^{-7}$ cm/s) among all identified soil types. This layer likely serves as a natural barrier to downward seepage, potentially influencing the pore water pressure distribution within the levee.
4. Type-E soil, primarily forming a cover layer on the landside of the levee, features relatively high permeability ($k = 4.02 \times 10^{-6}$ cm/s) and cohesion ($c = 28.9$ kPa). These characteristics enhance drainage and resistance to sliding soil on the landside face, potentially contributing significantly to the overall stability of the levee.
5. Type-D and Type-F soils, though less extensive and appearing mainly as isolated interlayers, play contrasting roles. Type-D soil has the lowest permeability ($k = 3.29 \times 10^{-7}$ cm/s) and the second-highest cohesion ($c = 28.9$ kPa) among all identified soil types, enhancing the permeation resistance and stability of the levee. Conversely, Type-F soil, characterized by relatively high permeability ($k = 3.95 \times 10^{-6}$ cm/s) and low strength parameters ($\phi = 16.4^\circ$, $c = 15.0$ kPa), may introduce localized vulnerabilities in terms of seepage control and structural integrity.

The alternating or interlayered arrangement of high and low permeability soil strata (particularly Type-B and Type-C) likely results in complex seepage patterns and pore water pressure distributions within the levee. This heterogeneously stratified structure underscores the critical importance of considering soil variability in comprehensive seepage and stability analyses of the levee system.

4.2. Engineering Analyses Based on Levee Soil Stratification

The soil stratification derived from the PAM clustering approach, as illustrated in Figure 3, provides a comprehensive representation of the internal structure of the levee in the study case. To demonstrate the practical significance of this result, we applied it to analyze the seepage and stability characteristics of the levee under design flood conditions. The computational analysis was performed using the AutoBank software (version 7.7) [19], a specialized tool for levee and riverbank engineering calculations. The following subsections present the results of the analysis, highlighting the influence of the identified soil strata on the performance of the levee during flood events.

4.2.1. Seepage Analysis

The seepage analysis in AutoBank is based on solving Laplace's equation, derived from a combination of Darcy's law and the continuity equation, for the potential head. AutoBank employs the Finite Element Method (FEM) to solve this equation. The computational mesh used for the FEM is shown in Figure 4.

The seepage analysis yielded reasonable results, as presented in Figures 5 and 6, showing the equipotential line map and velocity field, respectively, of the seepage flow. Note that in these figures, the design flood level has been marked out and the corresponding phreatic line has been highlighted in blue.

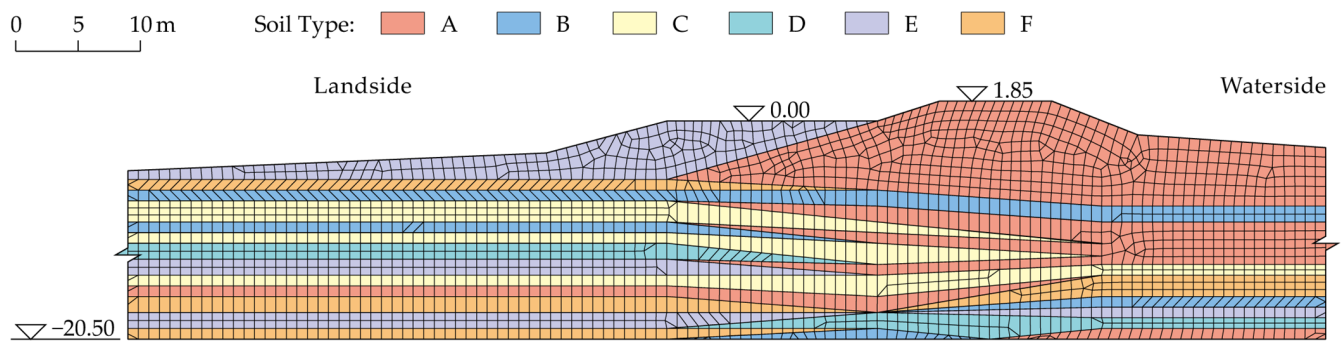


Figure 4. FEM mesh for the seepage analysis.

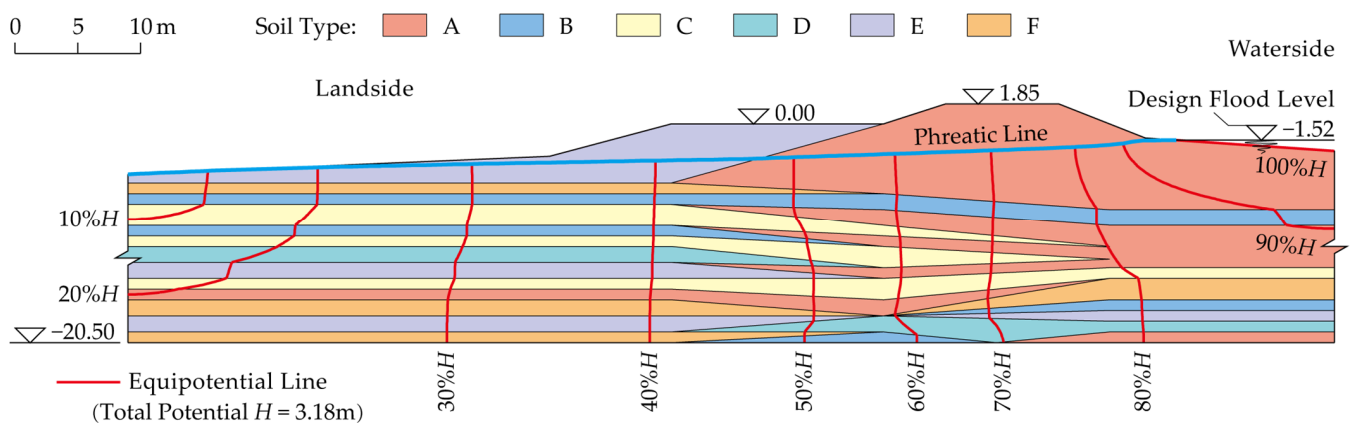


Figure 5. Equipotential line map identified by the seepage analysis.

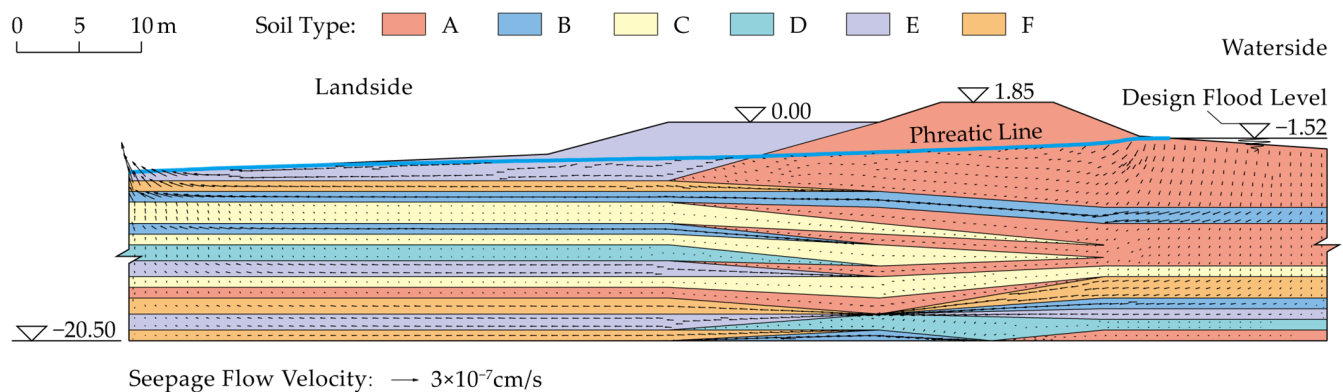


Figure 6. Velocity field identified by the seepage analysis.

Based on the results, the hydraulic gradient at the seepage exit point was calculated to be $J = 0.086$. Given that the levee soil is primarily clay, the potential form of seepage failure would be soil flow. Terzaghi's formula [20] was used to determine the critical hydraulic gradient (J_{cr}) for soil flow as follows:

$$J_{cr} = (G_s - 1)(1 - p), \quad (4)$$

where p , as mentioned in Section 2, is the porosity of soil; and G_s is the specific gravity of soil particles. Considering that $p \approx 42.7\%$ and G_s can be derived from the dry unit weight $\gamma_d \approx 15.2 \text{ kN/m}^3$ as $G_s = (\gamma_d/g)/(1 - p) \approx 2.7$ (g is the gravity acceleration) in the present study case, the critical hydraulic gradient for soil flow was determined to be $J_{cr} = 0.974$ through Equation (4).

Consequently, the factor of safety against seepage failure can be calculated as $FS_{\text{seepage}} = J_{cr}/J = 0.974/0.086 = 11.3$. This safety factor significantly exceeds the minimum allowable value ($=2.0$) specified by the relevant standards [21,22]. Therefore, the risk of seepage failure in the levee under the design flood conditions is considered low.

4.2.2. Stability Analysis

Under design flood conditions, the landside slope of the levee is particularly susceptible to sliding. To assess this risk, the simplified Bishop method [21] was employed, assuming a circular slip surface for the levee soil mass. This method, which considers the equilibrium of forces and moments on individual slices of the potential sliding mass, is widely used in geotechnical engineering for stability analysis of soil due to its balance of accuracy and computational efficiency.

Note that only the portions of Type-A and Type-E soils above the phreatic line required specification of their moisture content (ω): for Type-A soil, $\omega = 15.1\%$; and for Type-E soil, $\omega = 23.8\%$. The remaining portions of these soil types, as well as all other soil types in the levee, were considered saturated for the analysis.

After an automatic search by AutoBank, the most critical slip surface was identified, as illustrated in Figure 7. This represents the potential failure surface with the lowest safety factor.

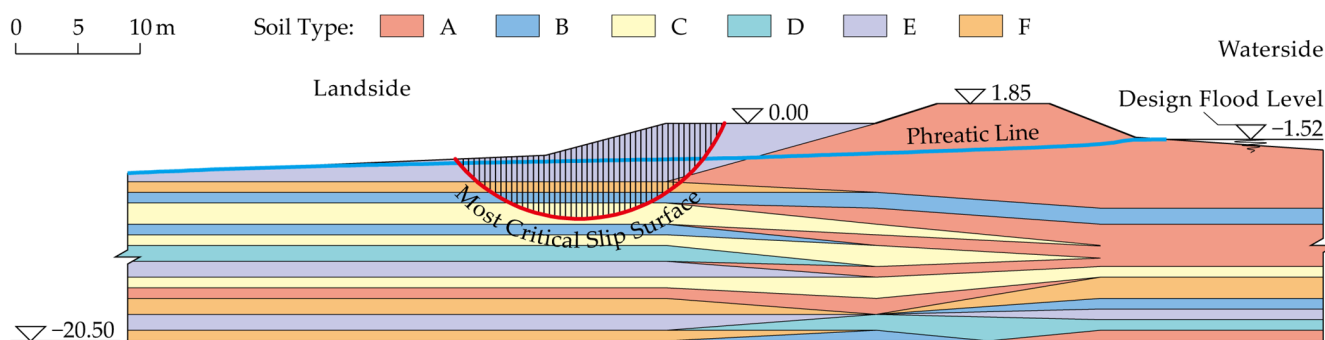


Figure 7. Most critical slip surface identified by AutoBank.

The stability analysis yielded a safety factor $FS_{\text{stability}} = 3.98$ for this critical slip surface. This value, which represents the ratio of resisting forces to driving forces along the slip surface, significantly exceeds the minimum allowable value (1.50 for levees of Grade 1) specified by the relevant standards [21,22]. The high safety factor indicates that the risk of slope instability in the levee under the design flood conditions is relatively low. This result, combined with the favorable seepage analysis outcomes, suggests that the overall structural integrity of the levee is robust under the specified flood conditions.

The above results of seepage and stability analyses are consistent with monitoring of the studied levee section during flood seasons in recent years. No significant seepage or stability issues have been detected in this particular levee section over the past several flood seasons, even during periods of high water levels [23]. This agreement between our analytical results and recent performance records further validates the effectiveness of the PAM clustering approach in characterizing the internal structure of the levee and predicting its behavior under flood conditions.

However, it is important to note that although these results are encouraging, continued monitoring and maintenance of the levee remain essential to ensure its long-term stability and performance, especially in the face of changing environmental conditions and potentially extreme weather events. Regular inspections, as well as updated seepage and stability analyses, should be conducted to account for any changes in soil properties or external conditions over time.

5. Discussion

Section 5 has demonstrated the practical application of soil stratification results obtained through the PAM clustering approach in analyzing the seepage and stability characteristics of levees. This section delves deeper into two critical aspects of the PAM clustering approach: firstly, the sensitivity of its effectiveness to the number of clusters (Section 5.1); and secondly, a comparison of its results with those of an alternative hierarchical clustering approach (Section 5.2).

5.1. Sensitivity Analysis of Number of Clusters

The number of clusters (K) is a critical parameter in the PAM clustering approach that significantly influences the resulting soil stratification, which is commonly selected in practice based on the operator's experience and prior knowledge. To understand the sensitivity of the clustering results to K , a comprehensive analysis was conducted using the case and data presented in Section 2.

PAM cluster analyses were performed for K values ranging from 1 to 20. For each clustering configuration, the cost C , as defined by Equation (3), was calculated. This cost function serves as a measure of the overall dissimilarity within clusters, with lower values indicating more internally consistent and well-defined clusters. Figure 8 illustrates the resulting variation pattern of the cost C versus the number of clusters K . For clarity, the data points for $K = 1$ and $K = 2$ are omitted from this figure, as their significantly large C values (164.8 and 84.5, respectively) would distort the scale of the graph.

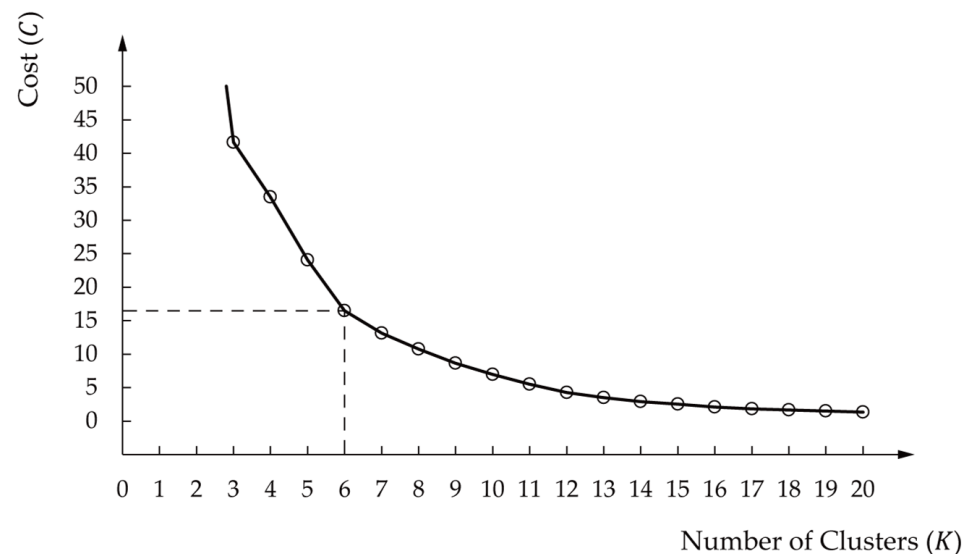


Figure 8. Variation pattern of the cost (C) versus the number of clusters (K).

The sensitivity analysis reveals two key insights:

1. **High sensitivity for $K < 6$:** The cost C shows a sharp decrease as K increases to 6, indicating that the clustering results are highly sensitive to the number of clusters in this range. The increase in the number of clusters significantly improves the internal consistency and distinctiveness of the resulting soil types.
2. **Low sensitivity for $K > 6$:** Beyond $K = 6$, the cost C continues to decrease but at a much more gradual rate, eventually approaching a plateau. This indicates low sensitivity to K in this range, where further subdivision of soil types yields diminishing returns in terms of improving cluster quality and distinctiveness.

The selection of K has significant implications for the resulting soil stratification and thus for the seepage and stability analyses. If K is set too high, it may lead to over-segmentation of the soil profile, potentially creating artificial distinctions among soil strata that are not significantly different in their engineering properties. This can complicate

the interpretation of results and the process of modeling when conducting a seepage and stability analysis without necessarily improving the accuracy. Conversely, if K is set too low, it may oversimplify the soil stratification, potentially merging distinct soil strata and obscuring important variations in soil properties that could affect the performance of a levee. This under-segmentation can lead to a loss of critical information about the internal structure of the levee, resulting in a failure to capture preferential seepage paths or localized weak zones within the levee and thus an underestimation or overestimation of seepage and stability risks.

Based on the above sensitivity analysis, an optimal number of clusters for the study case can be determined. The point at $K = 6$ represents a critical turning point (commonly referred to as the “Elbow Point” in a cluster analysis [24]) where the rate of improvement in cost C significantly slows. This suggests that a clustering configuration of six clusters offers the best balance between computational accuracy and practical interpretability, effectively capturing the main characteristics of the soil data without unnecessarily complicating the soil stratification. This is the reason why $K = 6$ was adopted in the previous sections for analyzing the present study case.

5.2. Comparison with a Hierarchical Clustering Approach

To further evaluate the effectiveness of the PAM clustering approach, a comparative analysis was conducted with a hierarchical clustering approach. As mentioned in the introduction, hierarchical clustering has been applied to soil classification in various geotechnical studies [7–10]. For this comparison, the agglomerative hierarchical clustering approach was employed with Ward’s minimum variance criterion [25], using the same standardized data calculated by Equation (1) and the distance measure represented by Equation (2) as in the PAM clustering approach.

Note that in the hierarchical clustering approach, Ward’s minimum variance criterion aims to merge two clusters in a way that minimizes the increase in the total variance within clusters. This objective is consistent with the cost function Equation (3) used in the PAM approach, as both approaches seek to minimize the overall dissimilarity within clusters. However, the key difference lies in their implementation: PAM optimizes globally across all clusters simultaneously, while the hierarchical approach makes local decisions at each merging step.

Figure 9 presents the dendrogram resulting from the hierarchical clustering analysis. The red rectangles indicate the six clusters obtained, corresponding to the optimal number of clusters determined for the PAM approach in Section 5.1.

Table 3 summarizes the attributes of the six clusters obtained through the hierarchical clustering approach, including the data points assigned to each cluster and the averaged soil properties (k , ϕ , and c) corresponding to each cluster.

Table 3. Attributes of the clusters ($K = 6$) obtained by the hierarchical clustering approach for the study case.

Cluster (Soil Type)	Data Points *	Averaged Soil Properties		
		k (cm/s)	ϕ (deg)	c (kPa)
A	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 20, 21, 23, 25, 28, 31, 44	6.39×10^{-7}	24.3	12.8
B	8, 16, 22, 33, 36, 39	6.08×10^{-6}	22.5	13.6
C	13, 19, 24, 26, 27, 29, 30, 37, 38, 40, 43	3.80×10^{-7}	20.4	13.3
D	18, 32, 41	3.29×10^{-7}	16.6	28.7
E	17, 34, 42, 46	4.02×10^{-6}	17.1	28.9
F	14, 15, 35, 45, 47	3.95×10^{-6}	16.4	15.0

* Identified by the “Global Sample No.” in Table 1.

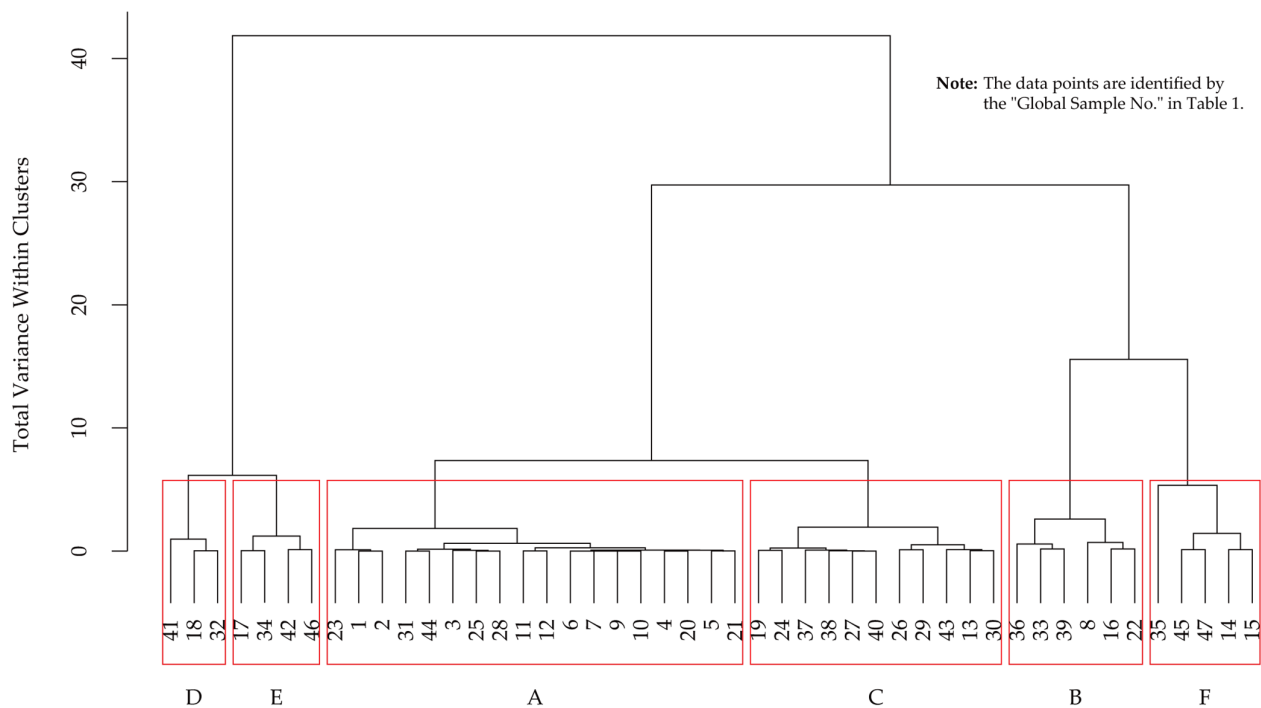


Figure 9. Dendrogram for the hierarchical clustering approach applied to the study case (red rectangles indicate the division into six clusters).

Comparing the results of the hierarchical clustering approach (Table 3) with those of the PAM clustering approach (Table 2) reveals that the clusters obtained by both approaches are largely consistent. The only notable difference is that Data Point 19 (underlined in Table 3) is assigned to Cluster A in the PAM approach but to Cluster C in the hierarchical approach. However, this change does not significantly alter soil properties corresponding to these clusters. The averaged values of k , ϕ , and c for Clusters A and C are quite similar between the two approaches. Moreover, this minor change would not significantly alter the overall soil stratification depicted in Figure 3. Only a small portion in the lower right corner of the levee section would be affected, and this localized change is unlikely to substantially impact the seepage and stability analyses of the levee as a whole.

This high degree of similarity between the results of the two approaches reinforces confidence in the soil stratification derived from the PAM clustering approach. The consistency in cluster compositions and soil properties across both approaches suggests that the identified soil types and their distributions within the levee are robust and not significantly dependent on the clustering approach.

The comparison also highlights the potential advantages of the PAM clustering approach:

1. PAM provides interpretable cluster centers (i.e., medoids) that are actual data points from the dataset. This feature is particularly useful in applications where having representative soil samples for each identified soil type aids in further analysis and interpretation.
2. PAM offers greater flexibility as it can perform clustering for any specified number of clusters. This feature is particularly beneficial when operators have more explicit prior knowledge about the expected soil types. In contrast, the hierarchical clustering approach must proceed agglomeratively (or, divisively).
3. PAM is better suited for handling larger datasets. As the number of data points increases, the computational complexity of hierarchical clustering grows rapidly, which would be reflected in the increasing complexity of the dendrogram. In contrast,

the computational efficiency of PAM clustering can be enhanced through various techniques [14–16], making it more suitable for big-data problems in engineering.

In conclusion, this comparison demonstrates that the PAM clustering approach is as effective as the well-established hierarchical clustering approach for levee soil stratification while offering additional benefits in terms of interpretability and scalability.

6. Conclusions

This study investigated the feasibility and effectiveness of applying the Partitioning Around Medoids (PAM) clustering approach to levee soil stratification using data from multiple probe drilling sites. The main conclusions are as follows:

1. The PAM clustering approach effectively identifies soil types with similar characteristics across multiple sampling locations, achieving reasonable levee soil stratification.
2. The derived soil stratification can be directly incorporated into seepage and stability analyses, capturing complex soil structures and providing accurate input parameters.
3. An optimal number of clusters (K) typically exists, balancing computational accuracy and practical interpretability. This can be determined by analyzing the relationship between clustering configuration cost and number of clusters.

For the study case of the Yellow River right-bank levee section in Zhongmu, Zhengzhou City, Henan Province, China, the following specific conclusions are drawn:

1. The optimal number of clusters (K) is determined to be six, effectively capturing soil characteristic variations while maintaining interpretability.
2. The PAM cluster analysis reveals heterogeneous soil stratification with alternating or interlayered soil strata, significantly impacting seepage and stability characteristics.
3. Seepage and stability analyses based on the soil stratification indicate a low risk of levee failure under design flood conditions, with safety factors far exceeding minimum allowable values specified by relevant standards.

By providing a more accurate and detailed soil stratification, the PAM clustering approach enables engineers to enhance the precision of seepage and stability analyses, leading to more reliable risk assessments and improved levee design. It addresses challenges posed by climate change, aids in identifying potential weak zones or preferential seepage paths within levees, and is particularly useful for large-scale levee systems with limited sampling data. However, limitations exist owing to data quality dependence and discrete sampling nature. Future research could focus on integrating advanced constitutive models to enhance the reliability and comprehensiveness of this approach.

Author Contributions: Conceptualization, H.Z. and L.S.; methodology, H.Z.; software, H.Z. and X.W.; validation, L.S. and Y.W.; formal analysis, H.Z., L.S. and W.D.; investigation, H.Z. and X.W.; resources, L.S. and Y.W.; data curation, L.S.; writing—original draft preparation, H.Z. and L.S.; writing—review and editing, W.D., X.W. and Y.W.; visualization, X.W. and Y.W.; supervision, W.D.; project administration, L.S.; funding acquisition, H.Z., L.S. and W.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Hebei Province, grant number E2024512014; the Science and Technology Project of Hebei Education Department, grant number BJK2023056; the Special Research Fund of the Yellow River Institute of Hydraulic Research, grant number HKY-JBYW-2023-17; the Science and Technology Research Project of Water Resources Department of Jiangsu Province of China, grant number 2021020; and the Fundamental Research Funds for the Central Universities, grant number ZY20220207.

Data Availability Statement: The original contributions presented in the study are included in the article, and further inquiries can be directed to the corresponding author.

Acknowledgments: The authors gratefully acknowledge Nanjing Shuizhun Technology Co., Ltd. for providing the AutoBank software (version 7.7), which was instrumental in conducting the seepage and stability analyses in this study. The software is freely available for download and use from their official website: <http://watersoft.net.cn> (accessed on 22 August 2024).

Conflicts of Interest: Author Yuan Wei was employed by the company Yellow River Engineering Consulting Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Cao, Z.; Wang, Y. Bayesian Approach for Probabilistic Site Characterization Using Cone Penetration Tests. *J. Geotech. Geoenviron. Eng.* **2013**, *139*, 267–276. [CrossRef]
2. Wang, Y.; Hu, Y.; Zhao, T. Cone Penetration Test (CPT)-Based Subsurface Soil Classification and Zonation in Two-Dimensional Vertical Cross Section Using Bayesian Compressive Sampling. *Can. Geotech. J.* **2020**, *57*, 947–958. [CrossRef]
3. Wang, X.; Wang, H.; Liang, R.Y.; Zhu, H.; Di, H. A Hidden Markov Random Field Model Based Approach for Probabilistic Site Characterization Using Multiple Cone Penetration Test Data. *Struct. Saf.* **2018**, *70*, 128–138. [CrossRef]
4. Hu, Y.; Wang, Y. Probabilistic Soil Classification and Stratification in a Vertical Cross-Section from Limited Cone Penetration Tests Using Random Field and Monte Carlo Simulation. *Comput. Geotech.* **2020**, *124*, 103634. [CrossRef]
5. Gong, W.; Zhao, C.; Juang, C.H.; Tang, H.; Wang, H.; Hu, X. Stratigraphic Uncertainty Modelling with Random Field Approach. *Comput. Geotech.* **2020**, *125*, 103681. [CrossRef]
6. Shi, C.; Wang, Y. Non-Parametric Machine Learning Methods for Interpolation of Spatially Varying Non-Stationary and Non-Gaussian Geotechnical Properties. *Geosci. Front.* **2021**, *12*, 339–350. [CrossRef]
7. Hegazy, Y.A.; Mayne, P.W. Objective Site Characterization Using Clustering of Piezocone Data. *J. Geotech. Geoenviron. Eng.* **2002**, *128*, 986–996. [CrossRef]
8. Liao, T.; Mayne, P.W. Stratigraphic Delineation by Three-Dimensional Clustering of Piezocone Data. *Georisk* **2007**, *1*, 102–119. [CrossRef]
9. Cai, G.; Liu, S.; Tong, L.; Du, G. Soil Classification Using CPTU Data Based upon Cluster Analysis Theory. *Chin. J. Geotech. Eng.* **2009**, *31*, 416–424, (In Chinese with English Abstract).
10. Molina-Gómez, F.; Viana da Fonseca, A.; Ferreira, C.; Sousa, F.; Bulla-Cruz, L.A. Defining the Soil Stratigraphy from Seismic Piezocone Data: A Clustering Approach. *Eng. Geol.* **2021**, *287*, 106111. [CrossRef]
11. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1990; ISBN 978-0-471-87876-6.
12. Velmurugan, T.; Santhanam, T. Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. *J. Comput. Sci.* **2010**, *6*, 363–368. [CrossRef]
13. Arora, P.; Deepali; Varshney, S. Analysis of K-Means and K-Medoids Algorithm for Big Data. *Procedia Comput. Sci.* **2016**, *78*, 507–512. [CrossRef]
14. Schubert, E.; Rousseeuw, P.J. Faster K-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. In *Proceedings of the Similarity Search and Applications: 12th International Conference, SISAP 2019, Newark, NJ, USA, 2–4 October 2019*; Amato, G., Gennaro, C., Oria, V., Radovanović, M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; Volume 11807, pp. 171–187.
15. Schubert, E.; Rousseeuw, P.J. Fast and Eager K-Medoids Clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms. *Inf. Syst.* **2021**, *101*, 101804. [CrossRef]
16. Tiwari, M.; Zhang, M.J.; Mayclin, J.; Thrun, S.; Piech, C.; Shomorony, I. BanditPAM: Almost Linear Time k-Medoids Clustering via Multi-Armed Bandits. In *Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: New York, NY, USA, 2020; Volume 33, pp. 10211–10222.
17. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to Algorithms*, 4th ed.; MIT Press: Cambridge, MA, USA, 2022; ISBN 978-0-262-36750-9.
18. Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. *Cluster: Cluster Analysis Basics and Extensions*, R Package Version 2.1.6. 2023. Available online: <https://CRAN.R-project.org/package=cluster> (accessed on 22 August 2024).
19. Nanjing Shuizhun Technology Co., Ltd. *AutoBank, Software Version 7.7*; Nanjing Shuizhun Technology Co., Ltd.: Nanjing, China, 2023; Available online: <http://watersoft.net.cn> (accessed on 22 August 2024).
20. GB 50487-2008; Code for Engineering Geological Investigation of Water Resources and Hydropower. Ministry of Housing and Urban-Rural Development of P.R. China; General Administration of Quality Supervision, Inspection and Quarantine of P.R.China: Beijing, China, 2008. (In Chinese with English Table of Contents)
21. GB 50286-2013; Code for Design of Levee Project. Ministry of Housing and Urban-Rural Development of P.R. China; General Administration of Quality Supervision, Inspection and Quarantine of P.R. China: Beijing, China, 2012. (In Chinese with English Table of Contents)
22. SL/Z 679-2015; Guidelines for Levee Safety Evaluation. Ministry of Water Resources of P.R. China: Beijing, China, 2015. (In Chinese with English Table of Contents)
23. Yellow River Institute of Hydraulic Research. *Safety Evaluation Report for Yellow River Levee Section 34+500 to 35+500 in Zhongmu*; HX-2023-89; Yellow River Institute of Hydraulic Research: Zhengzhou, China, 2023. (In Chinese)

24. Shi, C.; Wei, B.; Wei, S.; Wang, W.; Liu, H.; Liu, J. A Quantitative Discriminant Method of Elbow Point for the Optimal Number of Clusters in Clustering Algorithm. *EURASIP J. Wirel. Commun. Netw.* **2021**, *2021*, 31. [[CrossRef](#)]
25. Murtagh, F.; Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **2014**, *31*, 274–295. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.