الجمهورية الجزائرية الديمقراطية الشعبية
Democratic and People's Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة الشهيد الشيخ العربي التبسي – تبسة
Echahid Cheikh Larbi Tebessi University – Tebessa
معـــــــــهد المـــــناجم
Mining Institute
قسم المناجم والجيوتكنولوجيا
Department of Mines and Geotechnology

End of study dissertation

Presented in view of obtaining an academic Master's degree
Sector: Mining engineering

Option: Geotechnics

Indirect Estimation of California Bearing Ratio Using Statistical and Machine Learning Approaches

By
AZRI Hanine

In front of the jury:

BERRAH Yacine	**President**	**MCA**	**University of Echahid Cheikh Larbi Tebessi -Tebessa**
DJELLALI Adel	**Supervisor**	**Pr**	**University of Echahid Cheikh Larbi Tebessi -Tebessa**
HAMDANE Ali	**Examiner**	**MAA**	**University of Echahid Cheikh Larbi Tebessi -Tebessa**

Promotion 2024-2025

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
Et de la Recherche Scientifique
Université Larbi Tebessi – Tébessa
Institut des Mines
Département mines et géotechnologie

وزارة التعليم العالي والبحث العلمي
جامعةالعربي التبسي – تبسة
معهد المناجم
قسم المناجم والجيوتكنولوجيا

Année universitaire : 2024/2025 Tébessa le : 03/06/2025

Lettre de soutenabilité

Noms et prénoms des étudiants :

1 AZRI Hanine.

Niveau : 2éme Année Master Option : Génie Minier

Thème: **Indirect Estimation of California Bearing Ratio with Statistical and Machine Learning Approach**

Nom et prénom de l'encadreur : **Djellali Adel**

Chapitres réalisés	Signature de l'encadreur
Chapter I: **Review on Statistics**	
Chapter II: **Review on Artificial Intelligence**	
CHAPTER III: **Geotechnical Analysis and Soil Mapping**	
Chapter IV: **Development of CBR Formula by Statistical Model and ANN Method**	

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

مؤسسة التعليم العالي : جامعة العربي التبسي – تبسة

تصريح شرفي
خاص بالالتزام بقواعد النزاهة العلمية لانجاز بحث

أنا الممضي ادناه،

السيد (ة) عمروري حسين ..... الصفة : طالب، أستاذ باحث، باحث دائم : طالبة .....

الحامل لبطاقة التعريف الوطنية رقم 118003379006000 و الصادرة بتاريخ 2024/02/13 .....

المسجل بمعهد ..العلوم و التكنولوجيا ..... قسم الهندسة الكهربائية و الالكترونية

و المكلف بانجاز أعمال بحث (مذكرة التخرج، مذكرة ماستر، مذكرة ماجستير، اطروحة دكتوراه)، عنوانها :

Indirect estimation of California Bearing Ratio with statistical and machine learning

اصرح بشرفي أني ألتزم بمراعاة المعايير العلمية و المنهجية و معايير الأخلاقيات المهنية و النزاهة الأكاديمية المطلوبة في انجاز البحث المذكور أعلاه.

التاريخ: 2025/06/03

إمضاء المعني (ة)

Acknowledge

I would like to extend my deepest gratitude to the faculty and professors who have guided me throughout my academic journey. Your dedication to teaching, your commitment to our success, and your unwavering support have been invaluable.

Through your expertise, you've not only imparted knowledge but also inspired me to think critically, challenge myself, and strive for excellence. The lessons you've taught me go far beyond the classroom, shaping not only my career but also my perspective on life.

Thank you for your encouragement, patience, and for always pushing me to be my best. I am truly fortunate to have had the privilege of learning from such exceptional educators.

To the esteemed **Chikh Larbi tbessi faculty**,
I would like to express my deepest gratitude for the guidance, wisdom, and unwavering support you have provided throughout my academic journey. Your dedication to teaching and commitment to our growth have shaped not only my education but also my character.

Each of you has made a lasting impact on my life, and the lessons I have learned in your classrooms will continue to resonate with me long after graduation. Thank you for inspiring me to strive for excellence and for being a source of encouragement and motivation.

I am truly fortunate to have had the privilege of learning from such exceptional educators, and I will always carry your teachings with me into the next chapter of my life.

To the **Mining Institute**,
I would like to extend my sincerest gratitude for the invaluable education and training I have received throughout my time here. The knowledge, skills, and hands-on experience provided by this institute have equipped me with the tools necessary to succeed in the field of mining and beyond.

The dedication of the faculty and staff, as well as the resources available, have made a profound impact on my academic and professional growth. The experiences I've had here will undoubtedly shape my future, and I am grateful for the opportunities to learn, challenge myself, and grow within such a distinguished institution.

Thank you for preparing me for the challenges and opportunities ahead, and for being an integral part of my journey.

To the esteemed **professors of the Mining Institute**, particularly...

to the Geotechnical Engineering field professors,**Amrani Dounia, KarabatiRahouadja Nour, Mabrouk Faouzi, Berreh Yacine, Hamdane Ali, Benghazi Zied and Brahmi Sarhane**in his absence

I would like to express my deepest appreciation for the exceptional guidance, knowledge, and mentorship you have provided throughout my academic journey. Your expertise in geotechnical engineering has not only broadened my understanding of the field but has also inspired me to pursue this discipline with passion and dedication.

The rigorous coursework, coupled with your encouragement and invaluable insight, has given me the confidence and foundation to tackle the complexities of geotechnical challenges in the real world. Your patience, professionalism, and unwavering commitment to our success have been a driving force in shaping my education and future career.

Thank you for the dedication you've shown to my growth as a student and future engineer. The lessons I've learned in your classes will continue to guide me as I embark on my professional journey in mining and geotechnical engineering.

To my esteemed supervisor,**Djalali Adel**
I would like to extend my heartfelt gratitude for your exceptional guidance, support, and mentorship throughout my academic journey. Your expertise, thoughtful advice, and unwavering encouragement have been invaluable in shaping both my personal and professional growth.

You have not only helped me navigate the challenges of my studies but have also inspired me to approach my work with a greater sense of purpose and dedication. The knowledge and skills I have gained under your supervision will continue to guide me as I take the next steps in my career.

Thank you for believing in me, challenging me, and providing the direction I needed to succeed. It has truly been an honor to work under your leadership, and I will always carry the lessons I've learned from you forward.

Your beautiful student: Hanine

Thank you

As I close this chapter and prepare to step into a new one, I find myself reflecting on the journey that brought me here. Graduation isn't just a celebration of academic achievement—it's a moment filled with gratitude for all the love, support, and encouragement I've received along the way.

Before anything else, I want to take a moment to express my deepest thanks to those who walked beside me—through the late nights, the challenges, the small wins, and the big milestones. This accomplishment is not mine alone. It belongs to everyone who believed in me, lifted me up, and helped me grow.

To my dad,**Mouhamed**
Thank you for being my greatest protector, my quiet strength, and my constant source of support. Your sacrifices, your hard work, and your belief in me—even when I doubted myself—mean more than words could ever express.You taught me how to stay strong, how to keep going when things get tough, and how to always give my best. Every step I've taken, I've carried your love and guidance with me. This milestone is as much yours as it is mine.

I hope I've made you proud—because I'm endlessly proud to be your little girl.

To my mom,**Nadjet**
Thank you for being my heart, my home, and my greatest source of love. Your strength, patience, and unwavering support have carried me through more than you'll ever know.You were there for every late night, every tear, every doubt—and you met it all with nothing but love and encouragement. You believed in me when I couldn't believe in myself. You've been my biggest cheerleader, my best friend, and the reason I kept pushing forward.

This achievement is yours too, Mom. I am who I am because of you—and I will carry your love with me wherever I go.

To my sisters,**Romaissa, Ikram and Insaf**
Thank you for being my built-in best friends, my secret keepers, and my biggest supporters. Through every high and low, you've been by my side with love, laughter, and the kind of understanding only sisters can give.

You've inspired me, challenged me, and reminded me who I am when I started to forget. Your presence has been my comfort, your words my motivation, and your love my strength.

I'm so grateful to have grown up with you—and even more grateful to move forward in life knowing you'll always be there. This one's for us.

To my little brother,**Hamza**

Thank you for being the unexpected light in my darkest days and the quiet strength I didn't know I needed. Watching you grow has been one of the greatest joys of my life, and your innocence, humor, and love have meant more to me than you'll ever realize.Even when you didn't understand everything I was going through, your presence reminded me to smile, to breathe, and to keep going. You gave me a reason to stay strong—and someone to make proud.

No matter how old we get, you'll always be my little brother—and I'll always be here for you, just like you've been there for me.

To My Dear Nephews, **Ammar and Amir**

There's a kind of magic in watching little boys grow into young men—and I've been lucky enough to witness that magic in you.

I still remember the days when your laughter filled the house, when your tiny hands reached for mine, and when your eyes lit up at every new discovery. You were full of mischief, wonder, and so much love. Whether it was messy games, sleepy cuddles, or your endless questions about the world—you brought joy to my life in ways I can never fully explain.

Those moments are etched in my heart. And while time keeps moving forward, a part of me will always see you as those little boys who made everything brighter just by being there.

To My Precious Newborn Niece,**Iline**

From the moment I laid eyes on you, the world felt softer. Brighter. Like everything paused just for a second so we could all take in the miracle that is you.

You are so small, yet you've already filled such a big space in my heart. Your tiny hands, your perfect little face, the way you curl into the safety of love without even knowing it—you're a living reminder of hope, of innocence, and of just how beautiful life can be.

As you sleep peacefully now, unaware of the world, know that you are already so deeply loved. I promise to be there for you—not just as your [aunt/uncle], but as someone who will always cheer you on, protect your light, and remind you of how magical you truly are.

One day, you'll grow. You'll walk, talk, laugh, and chase dreams. But today, you're my tiny miracle—and I will never forget this version of you. So pure. So new. So full of possibility.

Welcome to the world, my sweet girl. We've been waiting for you.

To my best friends,**Hadil, Alia, Issra, Manal and Riham**
Where do I even begin? Thank you for being my rock, my laughter, my comfort, and my constant source of joy through every single moment of this journey. You've been the ones who saw me at my best and my worst, and loved me through it all.

We've shared late-night talks, countless memories, and moments of pure, unfiltered fun. You've been there to celebrate my victories and lift me up through every challenge. I honestly don't know how I would've made it through without you.

This graduation is just as much yours as it is mine, because we've done this together. I am beyond lucky to have you by my side and can't wait to see where life takes us next—because with you, I know it's going to be amazing.

To my college friends,**Ghofrane, Issra, Wafa, Chaima, Marwa, Djihan and Taki**
These past few years have been filled with so much more than just studying and exams—they've been about laughter, late-night talks, unforgettable moments, and the kind of friendship that changes you for the better.

You all have been my support system, my escape, my family away from home. Through every stressful day, every high, and every low, you've been there to share it all with me. We've grown together, learned together, and made memories that will last a lifetime.

Thank you for the endless encouragement, the shared dreams, and most of all, the love. I couldn't have asked for a better group of people to experience college with. This achievement is ours, and I'll forever cherish the bond we've built.

To my high school best friends,**Fatma, Bochra, Wissal and Noussaiba**
Even though time and distance may have separated us, I will forever cherish the memories we created together. You were there for me during some of the most formative years of my life, and the bond we shared back then continues to mean the world to me.

Through all the laughs, the tears, the adventures, and the heart-to-heart talks, we built something unbreakable. Though life may have pulled us in different directions, I will always carry the lessons, the love, and the happiness we shared with me.

You'll always be a part of who I am, and I will never forget the friendship and support you gave me when we were in those high school halls together. Thank you for everything.

To my childhood best friends, **Soundes and Khadija**
From playing in the backyard to staying up all night talking about our dreams, you've been with me through it all. We've grown up together, and through every stage of life, your friendship has been a constant, a reminder of the innocence and joy of those simpler days.

You saw me at my most carefree, my most vulnerable, and through every change, you've always been there, supporting me, laughing with me, and encouraging me to keep going. The memories we've made are treasures I'll hold forever—because they're not just memories, they're pieces of who I am.

Thank you for being the people I've always turned to and for being the ones who truly know me. No matter where life takes us, you'll always be my family, and I'll forever be grateful for the friendship we've shared.

Your sweetheart: Hanine

Summary

Acknowledge

Thank You

List of Figures

List of Tables

Abstract

This study proposes an indirect approach for estimating the California Bearing Ratio (CBR) using geotechnical parameters and machine learning, based on soil samples from the Tebessa region. The methodology consists of three phases. First, soil samples were classified using standard geotechnical properties. Second, Principal Component Analysis (PCA) was applied to identify the most influential variables affecting CBR which are finers, plasticity index (PI), carbonates content (CA), and dry unit weight (γ_d). A multiple linear regression model developed from these variables achieved an R^2 of 78%. In the third phase, an Artificial Neural Network (ANN) was implemented, improving prediction performance with an R^2 of 94% and lower MSE, RMSE, and MAE values. A simplified linear formula using the same variables, with Python program also produced a strong R^2 of 91%. These results demonstrate the potential of combining statistical analysis and machine learning for efficient and accurate CBR prediction.

Résumé

Cette étude propose une méthode d'estimation indirecte du coefficient de portance californien (CBR) à partir des paramètres géotechniques, par les méthodes statistiques et par apprentissage automatique, basée sur des échantillons de sol provenant de la région de Tébessa. La méthodologie se compose de trois phases. Premièrement, les échantillons de sol ont été classés selon des propriétés géotechniques standard. Deuxièmement, une analyse en composantes principales (ACP) a été appliquée afin d'identifier les variables les plus influentes sur le CBR, à savoir le passant, l'indice de plasticité (IP), la teneur des carbonates (CA) et le poids volumique sec (γ_d). Un modèle de régression linéaire multiple développé à partir de ces variables a atteint un coefficient de détermination (R^2) de 78 %. Dans la troisième phase, un réseau de neurones artificiels (RNA) a été mis en œuvre, améliorant la performance des prédictions avec un R^2 de 94 % ainsi que des valeurs plus faibles de MSE, RMSE et MAE. Une formule linéaire simplifiée utilisant les mêmes variables, avec le programme Python, a donné un R^2 de 91 %. Ces résultats démontrent le potentiel de la combinaison de l'analyse statistique et de l'apprentissage automatique pour une estimation efficace et précise du CBR.

الملخص

تقترح هذه الدراسة منهجية غير مباشرة لتقدير نسبة تحمل كاليفورني (CBR) باستخدام المعلمات الجيوتقنية والتعلم الآلي،استناداً إلى عينات التربة من منطقة تبسة. تتكون المنهجية من ثلاث مراحل. أولاً، تم تصنيف عينات التربة باستخدام الخصائص الجيوتقنية القياسية. ثانيًا، تم تطبيق تحليل المكونات الرئيسية (PCA) لتحديد المتغيرات الأكثر تأثيرًا على معدل تراكمي للتربة وهي: نسبة التربة أقل من0,080 مم، ومؤشر اللدونة(PI) ، ونسبة الكربونات(CA) ، والكثافة الجافة (γ_d).وقد حقق نموذج الانحدار الخطي المتعدد الذي تم تطويره من هذه المتغيرات معدل انحدار خطي متعدد (R^2) بنسبة 78%. في المرحلة الثالثة، تم تنفيذ شبكة عصبية اصطناعية(ANN) ، مما أدى إلى تحسين أداء التنبؤ مع نسبة R^2 تبلغ 94% وقيم أقل من MSE و RMSEو .MAEأنتجت أيضًا صيغة خطية مبسطة باستخدام نفس المتغيرات، استخدام برنامج البايثون، نسبة R^2 قوية بلغت 91%. توضح هذه النتائج إمكانات الجمع بين التحليل الإحصائي والتعلم الآلي للتنبؤ الفعال والدقيقلنسبة تحمل كاليفورني.

General introduction

The California Bearing Ratio (CBR) is a fundamental parameter in geotechnical engineering used to evaluate the strength and bearing capacity of subgrade soils, especially for the design of flexible pavements. Traditional methods for determining CBR rely on laboratory or field penetration tests, which, although accurate, are often time-consuming, labor-intensive, and costly. As a result, indirect estimation methods based on easily obtainable soil properties—such as moisture content, dry density, plasticity index, and grain size distribution—have gained significant attention. Statistical techniques, such as multiple linear and nonlinear regression, have been widely used to model the relationship between these soil parameters and CBR. However, these methods are often limited by their assumption of linear relationships and may not adequately capture the complex interactions between variables. To address these limitations, machine learning models have been introduced, offering enhanced flexibility and predictive accuracy. In this study, Bee Colony Optimization (BCO), a nature-inspired metaheuristic algorithm based on the foraging behavior of honey bees, is employed as the machine learning model to estimate CBR values. BCO is particularly effective in handling nonlinear, multidimensional optimization problems and is well-suited for modeling

the complex relationships inherent in geotechnical data. The integration of statistical analysis with BCO provides a robust and efficient framework for the indirect estimation of CBR, enabling faster and more cost-effective soil strength assessments.

This study focuses on developing and analyzing indirect estimation models for CBR using a combination of statistical techniques and machine learning algorithms. The research is organized into four chapters, each addressing a specific aspect of the investigation:

- **Chapter 1: Review of Statistical Methods**

This chapter presents a theoretical review of fundamental statistical concepts and tools relevant to geotechnical data analysis.

- **Chapter 2: Review of Machine Learning Techniques**

Chapter 2 provides a detailed review of machine learning methods used in geotechnical engineering, with a focus on metaheuristic algorithms such as Bee Colony Optimization (BCO). It highlights the potential of these models to capture complex nonlinear relationships and improve prediction accuracy in soil property estimation tasks.

- **Chapter 3: Soil Classification Based on Geotechnical Parameters**

In this chapter, the dataset used for the study is described and the soils are classified according to their geotechnical properties. This classification provides a foundation for modeling by grouping similar soil behaviors.

- **Chapter 4: Data Analysis and CBR Prediction Models**

Chapter 4 presents the core of the research. It involves the application of statistical techniques, including correlation analysis and multivariable regression, to identify key influencing factors on CBR. Additionally, machine learning models—especially Bee Colony Optimization (BCO)—are developed and compared to evaluate their performance in predicting CBR. Model validation is conducted using performance metrics such as R^2

Chapter I: Review on Statistics

Chapter I: Review on Statistics

I.1. History of statistics

Statistics is a fascinating scientific field that studies the collection, analysis, interpretation, and presentation of data. It has become an essential tool in the modern world for understanding complex systems, making predictions, and making informed decisions. However, the history of statistics is relatively recent; spanning a few centuries. He origins of statistics can be traced back to ancient civilizations such as Babylon, Egypt, and Rome, where people used basic statistical methods to keep records of population sizes, trade, and taxes. However, the development of modern statistics can be attributed to the European Renaissance, where the scientific method, critical thinking, and empirical observations became the norm.

One of the earliest pioneers of statistics was John Graunt, a British merchant, and statistician. In 1662, he published a book called "*Natural and Political Observations Made upon the Bills of Mortality*" which analyzed patterns of mortality in London. He calculated life expectancies, birth and death rates, and projected the growth of the population. Graunt's work laid the foundation for demographics, vital statistics, and public health. The 18th-century Enlightenment was a period of great scientific and intellectual discovery. Many scholars, including Leonhard Euler, Thomas Bayes, and Pierre Simon Laplace, contributed to the development of statistics. Euler developed mathematical theories of probability and combinatorics, which are relevant to modern statistics. Bayes developed Bayesian statistics, a method for making inferences based on probabilities. Laplace formulated the principle of indifference, which suggests that when there is no prior knowledge or evidence, all hypotheses are equally plausible.

The 19th century saw the widespread adoption of statistics in various fields, including medicine, astronomy, economics, and psychology. In medicine, William Farr, a British epidemiologist, used vital statistics to study the causes of disease and develop preventive measures. In astronomy, Karl Friedrich Gauss developed methods for calculating trajectories of celestial objects using statistical analysis. In economics, Adolphe Quetelet, a Belgian statistician, studied the distribution of human traits such as height, weight, and intelligence.

The 20th century was a century of rapid development in statistics, driven by technological advancements and new applications in fields such as genetics, psychology, and social sciences. In 1925, Fisher introduced the concept of analysis of variance (ANOVA), a

statistical method for testing the significance of differences among multiple groups. ANOVA became a fundamental tool in research design and data analysis.

In the mid-20th century, a new field called mathematical statistics emerged, which focused on the development of probability and statistical theories. The development of computers in the 1960s revolutionized the field of statistics. It became possible to store, process, and analyze vast amounts of data, leading to the development of statistical software such as SAS, SPSS, and R. In recent years, statistics has become an essential tool in data science, a field that focuses on extracting insights and knowledge from large datasets. Data scientists use statistical methods such as regression, clustering, and classification to model and analyze data. They also use machine learning algorithms to build predictive models and make recommendations (Subhabaha, 2023).

I.2. Statistical Analysis

Statistical analysis is the process of collecting and analyzing data in order to discern patterns and trends. It is a method for removing bias from evaluating data by employing numerical analysis. This technique is useful for collecting the interpretations of research, developing statistical models, and planning surveys and studies.

Statistical analysis is a scientific tool in AI and ML that helps collect and analyze large amounts of data to identify common patterns and trends to convert them into meaningful information. In simple words, statistical analysis is a data analysis tool that helps draw meaningful conclusions from raw and unstructured data.

I.2.3. Types of Statistical Analysis

Given below are the 6 types of statistical analysis:

- **Descriptive Analysis**

Descriptive statistical analysis involves collecting, interpreting, analyzing, and summarizing data to present them in the form of charts, graphs, and tables. Rather than drawing conclusions, it simply makes the complex data easy to read and understand.

- **InferentialAnalysis**

The inferential statistical analysis focuses on drawing meaningful conclusions on the basis of the data analyzed. It studies the relationship between different variables or makes predictions for the whole population.

- **Predictive Analysis**

Predictive statistical analysis is a type of statistical analysis that analyzes data to derive past trends and predict future events on the basis of them. It uses machine learning algorithms, data mining, data modelling, and artificial intelligence to conduct the statistical analysis of data.

- **Prescriptive Analysis**

The prescriptive analysis conducts the analysis of data and prescribes the best course of action based on the results. It is a type of statistical analysis that helps you make an informed decision.

- **Exploratory Data Analysis**

Exploratory analysis is similar to inferential analysis, but the difference is that it involves exploring the unknown data associations. It analyzes the potential relationships within the data.

- **Causal Analysis**

The causal statistical analysis focuses on determining the cause-and-effect relationship between different variables within the raw data. In simple words, it determines why something happens and its effect on other variables. This methodology can be used by businesses to determine the reason for failure.

I.2.4. Statistical Analysis Methods

Although there are various methods used to perform data analysis, given below are the 5 most used and popular methods of statistical analysis:

- **Mean**

Mean or average mean is one of the most popular methods of statistical analysis. Mean determines the overall trend of the data and is very simple to calculate. Mean is calculated by summing the numbers in the data set together and then dividing it by the number of data points. Despite the ease of calculation and its benefits, it is not advisable to resort to mean as the only statistical indicator as it can result in inaccurate decision making.

$$\bar{x} = \frac{x_1 + x_2 \ldots + x_n}{n} \quad (1)$$

- **Standard Deviation**

Standard deviation is another very widely used statistical tool or method. It analyzes the deviation of different data points from the mean of the entire data set. It determines how data of the data set is spread around the mean. You can use it to decide whether the research outcomes can be generalized or not.

- **Regression**

Regression is a statistical tool that helps determine the cause-and-effect relationship between the variables. It determines the relationship between a dependent and an independent variable. It is generally used to predict future trends and events.

- **Hypothesis Testing**

Hypothesis testing can be used to test the validity or trueness of a conclusion or argument against a data set. The hypothesis is an assumption made at the beginning of the research and can hold or be false based on the analysis results.

- **Sample Size Determination**

Sample size determination or data sampling is a technique used to derive a sample from the entire population, which is representative of the population. This method is used when the size of the population is very large. You can choose from among the various data sampling techniques such as snowball sampling, convenience sampling, and random sampling (Aditya, 2024).

I.3. What Is Statistics?

Statistics is a branch of applied mathematics that involves the collection, description, analysis, and inference of conclusions from quantitative data. The mathematical theories behind statistics rely heavily on differential and integral calculus, linear algebra, and probability theory.

People who do statistics are referred to as statisticians. They're particularly concerned with determining how to draw reliable conclusions about large groups and general events from the behavior and other observable characteristics of small samples. These small samples

represent a portion of the large group or a limited number of instances of a general phenomenon (Chappelow, 2024).

I.4. Descriptive and Inferential Statistics

The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions. Descriptive statistics include mean (average), variance, skewness, and kurtosis. Inferential statistics include linear regression analysis, analysis of variance (ANOVA), logit/Probit models, and null hypothesis testing.

I.4.1. Descriptive Statistics

I.4.1.1. Mean

The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar. Calculating the mean is very simple. You just add up all of the values and divide by the number of observations in your dataset.

$$\overline{x} = \frac{x_1 + x_2 \ldots . + x_n}{n} \qquad (2)$$

The calculation of the mean incorporates all values in the data. If you change any value, the mean changes. However, the mean doesn't always locate the center of the data accurately.

I.4.1.2. Median

The median is the middle value. It is the value that splits the dataset in half, making it a natural measure of central tendency.To find the median, order your data from smallest to largest, and then find the data point that has an equal number of values above it and below it. The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values.

I.4.1.3. Mode

The mode is the value that occurs the most frequently in your data set, making it a different type of measure of central tendency than the mean or median.To find the mode, sort the values in your dataset by numeric values or by categories.Then identify the value that occurs most often (Frost, 2018).

Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured. It includes metrics such as range, variance, and standard deviation (Chappelow, 2024).

I.4.2. Inferential Statistics

I.4.2.1. Standard deviation

Standard deviation is calculated by taking the square root of a value derived from comparing data points to a collective mean of a population (Hargrave, 2024).

The formula is:

$$\textit{Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}, \qquad (3)$$

Where:

x_i=Value of the point in the data set.

$\bar{x}$=The mean value of the data set.

n=The number of data points in the data set.

I.4.2.2 Variance

Variance is a statistical measurement of the spread between numbers in a data set. It measures how far each number in the set is from the mean (Hayes, 2024).

The formula is:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{N} \qquad (4)$$

Where:

x_i=Each value in the data set.

$\bar{x}$=Mean of all values in the data set.

N=Number of values in the data set.

The distribution refers to the overall “shape” of the data, which can be depicted on a chart such as a histogram or a dot plot, and includes properties such as the probability

distribution function, skewness, and kurtosis. Descriptive statistics can also describe differences between observed characteristics of the elements of a data set. They can help us understand the collective properties of the elements of a data sample and form the basis for testing hypotheses and making predictions using inferential statistics (Chappelow, 2024).

I.5. Principal Component Analysis

Principal component analysis **(abbreviated as PCA)** is a versatile statistical method for reducing a cases-by-variables data table to its essential features, called principal components. Principal components are a few linear combinations of the original variables that maximally explain the variance of all the variables. In the process, the method provides an approximation of the original data table using only these few major components (Greenacre et al, 2022).

I.5.1. What Are Principal Components?

Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components. So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on (Jaadi, 2024).

I.5.1.1. First Principal Component (Z^1)

The first principal component is a linear combination of original predictor variables that captures the data set's maximum variance. It determines the direction of highest variability in the data. Larger the variability captured in the first component, larger the information captured by component. No other component can have variability higher than first principal component.

The first principal component results in a line that is closest to the data.

I.5.1.2. Second Principal Component (Z^2)

The second principal component is also a linear combination of original predictors, which captures the remaining variance in the data set and is uncorrelated with Z^1. In other words, the correlation between first and second components should be zero (Singh, 2025).

It can be represented as:

$$Z^2 = \phi^{12}X^1 + \phi^{22}X^2 + \phi^{32}X^3 + \cdots + \phi^{p2}X^p \qquad (5)$$

I.5.2. Step-by-step Explanation of PCA

Step 1: Standardization

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (for example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.

Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation} \qquad (6)$$

Once the standardization is done, all the variables will be transformed to the same scale.

Step 2: Covariance Matrix Computation

The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix**.**

The covariance matrix is a $p \times p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables. For

example, for a 3-dimensional data set with 3 variables *x*, *y*, and *z*, the covariance matrix is a 3×3 data matrix of this from:

$$\begin{bmatrix} \boldsymbol{cov(x,x)} & \boldsymbol{cov(x,y)} & \boldsymbol{cov(x,z)} \\ \boldsymbol{cov(y,x)} & \boldsymbol{cov(y,y)} & \boldsymbol{cov(y,z)} \\ \boldsymbol{cov(z,x)} & \boldsymbol{cov(z,y)} & \boldsymbol{cov(z,z)} \end{bmatrix} \quad \textbf{(7)}$$

Since the covariance of a variable with itself is its variance ***(Cov (a, a) =Var(a))***, in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable. And since the covariance is commutative ***(Cov (a, b) =Cov(b, a))***, the entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.

What do the covariances that we have as entries of the matrix tell us about the correlations between the variables?

It's actually the sign of the covariance that matters:

- If positive then: the two variables increase or decrease together (correlated)
- If negative then: one increases when the other decreases (Inversely correlated)

Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

What you first need to know about eigenvectors and eigenvalues is that they always come in pairs, so that every eigenvector has an eigenvalue. Also, their number is equal to the number of dimensions of the data. For example, for a 3-dimensional data set, there are 3 variables, therefore there are 3 eigenvectors with 3 corresponding eigenvalues.

It is eigenvectors and eigenvalues who are behind all the magic of principal components because the eigenvectors of the Covariance matrix are actually the directions of the axes where there is the most variance (most information) and that we call Principal Components. And eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.

By ranking your eigenvectors in order of their eigenvalues, highest to lowest, you get the principal components in order of significance.

Step 4: Create a Feature Vector

As we saw in the previous step, computing the eigenvectors and ordering them by their eigenvalues in descending order, allow us to find the principal components in order of significance. In this step, what we do is, to choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call *Feature vector*.

So, the feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only ***p*** eigenvectors (components) out of ***n***, the final data set will have only ***p*** dimensions.

Step 5: Recast the Data Along the Principal Components Axes

In the previous steps, apart from standardization, you do not make any changes on the data, you just select the principal components and form the feature vector, but the input data set remains always in terms of the original axes (i.e, in terms of the initial variables).

In this step, which is the last one, the aim is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components (hence the name Principal Components Analysis). This can be done by multiplying the transpose of the original data set by the transpose of the feature vector (Jaadi, 2024).

Final Data Set = Feature Vector* Standardized Original Data Set

I.6. Regression

I.6.1. What Is Regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between a dependent variable and one or more independent variables.Linear regression is the most common form of this technique. Also called simple regression or ordinary least squares (OLS), linear regression establishes the linear relationship between two variables.

Linear regression is graphically depicted using a straight line of best fit with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of the dependent variable when the value of the independent variable is zero. Regression captures the correlation between variables observed in a data set and quantifies whether those correlations are statistically significant or not.

The two basic types of regression are simple linear regression and multiple linear regression, although there are nonlinear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome (Beers, 2024).

I.6.2. Simple linear regression

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

I.6.2.1. Assumptions of simple linear regression

Simple linear regression is a parametric test, meaning that it makes certain assumptions about the data. These assumptions are:

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. Normality: The data follows a normal distribution.

Linear regression makes one additional assumption:

4. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

I.6.2.2. Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 x + \epsilon \qquad (8)$$

Where:

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- β_0 is the intercept, the predicted value of **y** when the **x** is 0.
- β_1 is the regression coefficient – how much we expect **y** to change as x increases.
- X is the independent variable (the variable we expect is influencing y).
- ϵ is the error of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B_1) that minimizes the total error (e) of the model (Bevans, 2020).

I.6.3. Multiple linear regression

I.6.3.1. What Is Multiple Linear Regression (MLR)?

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of MLR is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

I.6.3.2. Formula and Calculation of Multiple Linear Regression (MLR)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon \qquad (9)$$

where, for $i=n$ observations:

y_i=dependent variable.

x_i=explanatory variables.

β_0=y-intercept (constant term).

β_p=slope coefficients for each explanatory variable.

ϵ=the model's error term (also known as the residuals).

I.6.3.3. Assumptions of multiple linear regression

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.
- y_i observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of **0** and variance σ.

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.

R2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables (Hayes, 2024).

I.6.4. Geostatistical Analysis

I.6.4.1. Historical Look at Geostatistics

The field of geostatistics boasts a fascinating history, evolving from a specific need in one industry to a widely used tool across various disciplines. Here's a glimpseintoitspast :

- Early Beginnings (1950s): The roots of geostatistics can be traced back to the 1950s in the mining industry. D.G. Krige, a South African mining engineer, and H.S. Sichel, a statistician, developed a new method called kriging to estimate ore reserves more

accurately. This method took into account the spatial distribution of the ore, which was a significant improvement upon traditional methods that ignored spatial patterns.

- Formalization and Expansion (1960s-1970s): French engineer Georges Matheron is considered the father of geostatistics. He further developed and formalized Krige's concept of kriging, coining the term itself and establishing a theoretical framework for geostatistics. During this time, geostatistics began to see applications beyond mining, venturing into fields like forestry and meteorology.
- Rise of Computing and Wider Adoption (1980s-present): The advent of high-speed computers in the 1970s significantly accelerated the adoption of geostatistics. Complex calculations required for geostatistical modeling became feasible, leading to its wider use in various scientific disciplines. The oil and gas industry adopted geostatistics in the late 1980s, further solidifying its place as a valuable tool for understanding and managing natural resources.
- Application in Medical Geography and Environmental Health (2000s-present): As recognition of links between human health and environment burgeoned, geostatistics has emerged as a foundational technique in medical geography. Here, geostatistics is used to estimate underlying disease risk, detect areas with significantly higher risk, and analyze relationships with putative risk factors.

I.6.4.2. What is geostatistics?

Geostatistics is a branch of statistics that specifically deals with spatial data, meaning data that has a location associated with it. Unlike traditional statistics, which focus on analyzing data without considering its spatial context, geostatistics takes into account the fact that things closer together are often more similar than things farther apart. This principle, known as Tobler's First Law of Geography, forms the foundation of geostatistical analysis. Geostatistics was originally developed for the mining industry to predict ore grades. However, it's now used in a wide range of fields, including environmental science, hydrology, meteorology, agriculture, and even epidemiology.

I.6.4.2.1. Techniques

Geostatisticians use various techniques and tools to analyze and model spatial data. These techniques include:

- **Variograms:** These graphs depict the spatial dependence of data points, helping to identify patterns and trends.
- **Kriging:** This is a method for interpolating values at unsampled locations, taking into account the spatial autocorrelation of the data. (Jacquez, 2024).

I.6.4.2.1.1. Kriging

Kriging is a geostatistical technique that provides the Best Linear Unbiased Estimate (BLUE) of a variable at an unobserved location based on the spatial correlation of data. Here's how Kriging Works:

1. **Spatial Correlation:** Kriging is based on the 1st law of geography, which states that nearby data points are more correlated than distant ones. It quantifies this spatial correlation using a semivariogram or covariance function.
2. **Weighted Averages:** To estimate a value at an unobserved location, Kriging computes a weighted average of nearby sample points, where the weights are determined by the spatial correlation and data geometry (e.g. clustered data receive less weight than isolated observations because of their spatial redundancy).
3. **Optimality:** Kriging aims to minimize the estimation error variance while ensuring that the estimates are unbiased.

Kriging is versatile and effective for mapping both continuous and categorical variables**.** (Goovaerts, 2023).

I.6.4.2.1.1.2. Types of kriging

I.6.4.2.1.1.2.1. Ordinary Kriging

Ordinary kriging is the most widely used kriging method. It serves to estimate a value at a point of a region for which a variogram is known, using data in the neighborhood of the estimation location. Ordinary kriging can also be used to estimate a block value. With local second-order stationarity, ordinary kriging implicitly evaluates the mean in a moving neighborhood. To see this, first a kriging estimates of the local mean is set up, then a simple kriging estimator using this kriged mean is examined.

I.6.4.2.1.1.2.2. Ordinary kriging problem

We wish to estimate a value at Xo using the data values:

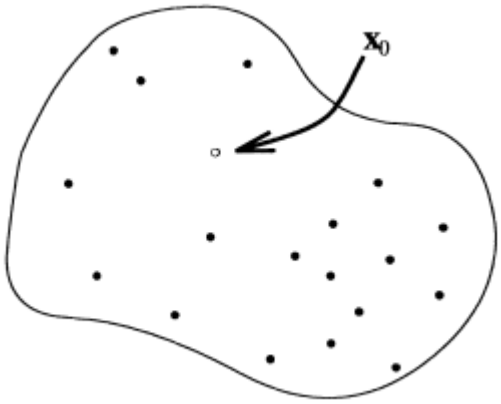

Figure 1 : Linear regression and simple kriging (Paláncz et al, 2023).

$$Z^*_{OK}(x_0) = \sum_{\alpha=1}^{n} Z(x_\alpha). \qquad (10)$$

Obviously, we have to constrain the weights to sum up to one because in the particular case when all data values are equal to a constant, the estimated value should also be equal to this constant.

We assume that the data are part of a realization of an intrinsic random function ***Z(x)***with a variogram gamma***(h)***. The unbiasedness is warranted with unit sum weights.

$$E[Z^*(x_0) - Z(x_0)] = \mathrm{E}[\sum_{\alpha=1}^{n} \omega_\alpha Z(x_\alpha) - Z(x_0) * \sum_{\alpha=1}^{n} \omega_\alpha] \qquad \Rightarrow \sum_{\alpha=1}^{n} \omega_\alpha = 1$$

$$= \sum_{\alpha=1}^{n} \omega_\alpha E[Z(x_\alpha) - Z(x_0)] = 0 \qquad (11)$$

because the expectations ofthe incrementsare zero. The estimation variance equals var$(Z^*(x_0) - Z(x_0))$ is the variance of the linear combination.

$$Z^*(x_0) - Z(x_0) = \sum_{\alpha=1}^{n} \omega_\alpha Z(x_\alpha) - 1 * Z(x_0) = \sum_{\alpha=0}^{n} \omega_\alpha Z(x_\alpha) \qquad (12)$$

with a weight Wo equal to -1 and:

$$\sum_{\alpha=0}^{n} \omega_\alpha = 0. \qquad (13)$$

Thus, the condition that the weights numbered from 1 to n sum up to one also implies that the use of the variogram is authorized in the computation of the variance of the estimation error.

The estimation variance is:

$$\sigma_E^2 = E[(Z^*(x_0) - Z(x_0))^2]$$

$$= -\gamma(x_0 - x_0) - \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \omega_\alpha \omega_\beta \gamma(x_\alpha - x_\beta) + 2\sum_{\alpha=1}^{n} \omega_\alpha \gamma(x_\alpha - x_0) \quad (14)$$

By minimizing the estimation variance with the constraint on the weights, we obtain the ordinary kriging system (OK) (Wackernagel, 2003).

$$\begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \omega_1^{OK} \\ \vdots \\ \omega_n^{OK} \\ \mu_{OK} \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix} (15)$$

I.6.4.2.1.1.2.2. Co-Kriging

Co-Kriging is an extension of univariate Kriging that allows for incorporating secondary (auxiliary) variables, in addition to the primary variable of interest. These secondary variables can help improve predictions by providing additional information about the spatial variability. Co-Kriging predicts the primary variable while considering the cross-covariances between the primary and secondary variables. It shares the same characteristics of minimizing the variance of prediction errors, unbiasedness, and calculating estimates as weighted averages of surrounding data. However, it's more tedious to implement than kriging as it requires the joint modeling of multiple direct and cross-semivariograms. (Goovaerts, 2023).

I.6.4.2.1.1.2.2.1. Co-kriging formula

In geostatistics, an observed value at a spatial location u_i is modelled as a realisation of a random variable $Z(u_i)$. For example, Z may represent a spatial variable, such as temperature, and $u_i = \{x_i, v_i\}$ is a spatial location on the plane; thus $Z(u_i)$ models the temperature at that location. The set of all random variables $Z(u_i)$ in a region χ of the space $u_i \in \chi$ is a random function, or a random field, $Z(u)$. For $\chi \subset Rd$ and $d=2$, the problem is two-dimensional, which is the case in the work presented here.

It is assumed that the random function ***Z(u)*** is second-order stationary, that is, with constant spatial mean and covariance function ***CZ(u,u+h)*** that depends only on the vector h

$$E\{Z(u)\} = mZ, \quad (16)$$

$$CZ(h) = E\{Z(u)Z(u+h)\} - mZ2, \quad (17)$$

$$CZ(0) = \sigma Z2, \quad (18)$$

where ***mZ***,***σZ2*** and ***CZ(h)*** are the mean, variance and covariance, respectively, of the random function ***Z(u)*** and E***{Z(u)}*** is the mathematical expectation operator. Although the variogram is most often used in geostatistics, only covariances are used in the work presented here.

Co-kriging is a linear estimator that can be written as (e.g., Journel and Huijbregts 1978; Ver Hoef and Cressie 1993)

$$\boldsymbol{Z^*(u_0) = \sum_{i=1}^{n} \lambda_i^0 Z(u_i) + \sum_{i=1}^{m} \beta_j^0 Y(u_j)} \quad \textbf{(19)}$$

where$\boldsymbol{Z^*(u_0)}$ is the estimated value of the primary variable at the spatial location $\boldsymbol{u_0}$= ***{***$\boldsymbol{x_0}$,$\boldsymbol{y_0}$ ***}***, that is, $\boldsymbol{Z(u_0)}$, where $\boldsymbol{u_0}$ is a point on the plane with coordinates x0 as easting and y0 as northing. ***{***$\boldsymbol{Z(u_i)}$***;i=1,…,n}*** is the set of anexperimental data of the primary variable in last Eq, and ***{***$\boldsymbol{Y(u_j)}$***;j=1,…,m}*** is the set of m experimental data of the auxiliary variable used in Eq. $\boldsymbol{\lambda_i^0}$ is the weight applied to the primary variable$\boldsymbol{Z(u_i)}$ in the estimation of$\boldsymbol{Z(u_0)}$, and $\boldsymbol{\beta_j^0}$ is the weight applied to the auxiliary variable$\boldsymbol{Z(u_j)}$ in the estimation of$\boldsymbol{Z(u_0)}$.

The set of optimal weights ***{***$\boldsymbol{\lambda_i^0}$***; i=1…, n}*** and ***{***$\boldsymbol{\beta_j^0}$***; i=1…, n}*** are obtained by minimising the variance of the estimation error.

$$\boldsymbol{Var\{Z^*(u_0) - Z(u_0)\}} \quad \textbf{(20)}$$

subject to the unbiasedness condition

$$\boldsymbol{E\{Z^*(u_0) - Z(u_0)\} = 0} \quad \textbf{(21)}$$

The position of $\boldsymbol{u_0}$ can vary in order to define a grid (raster image), a polygon, lineation, and so on.

I.6.4.2.1.1.2.3. Ordinary Co-kriging

In the simplest case, a spatial variable of interest ***Z(u)***, or primary variable, is to be estimated at a location $\boldsymbol{u_0}$ at which the variable was not sampled. The variable is to be estimated by

using the experimental values of the primary variable $\{Z(u_i); i=1\ldots, n\}$ and the experimental values of an auxiliary variable $\{Y(u_j); j=1\ldots, m\}$. The ordinary co-kriging estimator is given in 19thEq, and the optimal weights are obtained by minimising the variance of the estimation error given in 20thEqsubject to the unbiasedness condition in 21thEq. The unbiasedness condition of the co-kriging in 21thEqimplies that the following conditions must be satisfied.

$$\sum_{i=1}^{n} \lambda_i^0 = 1 \qquad (22)$$

$$\sum_{j=1}^{m} \beta_j^0 = 0 \qquad (23)$$

Isaaks and Srivastava (1989) showed that a single unbiased condition in co-kriging

$$\sum_{i=1}^{n} \lambda_i^0 + \sum_{j=1}^{m} \beta_j^0 = 1 \qquad (24)$$

The variance of the estimation error, or estimation variance, can be written as (Myers 1982, 1983, 1991; Isaaks and Srivastava 1989; Wackernagel 2003).

$$Var\{Z^*(u_0) - Z(u_0)\} = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^0 \lambda_j^0 C_{YZ}(h_{ij}) + \sum_{i=1}^{m}\sum_{j=1}^{m} \beta_i^0 \beta_j^0 C_Y(h_{ij}) + \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_i^0 \beta_j^0 C_{ZY}(h_{ij}) + \sum_{j=1}^{m}\sum_{i=1}^{n} \beta_j^0 \lambda_i^0 C_{YZ}(h_{ji}) - 2\sum_{i=1}^{n} \lambda_i^0 C_Z(h_{i0}) - 2\sum_{j=1}^{m} \lambda_j^0 C_Y(h_{j0}) + C_Z(h_{00}) \qquad (25)$$

where $C_Z(h_{ij})$ is the covariance between the random variables $Z(u_i)$ and$Z(u_j)$ for which the spatial distance is equal to the vector (h_{ij})=**uj**−**ui**. In a similar way, $C_{ZY}(h_{ij})$is defined as the cross-covariance between the random variables$Z(u_i)$ and$Y(u_j)$

$$C_{ZY}(h_{ij}) = E\{Z(u_i)Y(u_j)\} - m_Z m_Y \qquad (26)$$

Similarly, $C_{YZ}(h_{ji})$is the cross-covariance between the random variables $Y(u_j)$ and$Z(u_i)$(Paláncz& all, 2023).

I.7. Variogram

I.7.1. What is variogram?

The variogram is the cornerstone of many geostatistical applications. The experimental variogram and any model fitted to it should be accurate. Only then can the model describe the variation reliably. Kriging requires a variogram, and it is by ensuring its accuracy that you will eventually obtain minimum-variance predictions by kriging. If the variogram describes the variation poorly then the kriged predictions are likely to be poor also, and they might have little or no validity no matter how 'pretty' the map. The term 'cartographic pornography' has been used by those who realize that no confidence can be placed in many of the beautiful smooth maps that exist because of sparsity of the data that underlies them (Oliver and Webster, 2015).

I.7.2. Importance of the Variogram in Geostatistics

The variogram is used by most geostatistical mapping and modeling algorithms. Object-based facies models and certain iterative algorithms, such as simulated annealing, do not use variograms. Not only is the variogram used extensively, it has a great effect on predictions. Occasionally there are enough data to control the appearance and behavior of the numerical models; however, these cases are infrequent and of lesser importance than the common case of sparse data control. The available data are too widely spaced to provide effective control on the numerical model. The variogram provides the only effective control on the resulting numerical models. The lack of data, which makes the variogram important, also makes it difficult to calculate, interpret, and model a reliable variogram (Cressie and Hawkins, 1980;Genton, 1998). Practitioners have been aware of this problem for some time with no satisfactory solution. Variogram modeling is important and the "details" often have a crucial impact on prediction. In particular, the treatment of zonal anisotropy and systematic vertical or horizontal trends is critical (Gringarten and Deutsch, 2000).

I.7.3. Variogram Models

Most geostatistical estimation or simulation algorithms require an analytical variogram model, which they will reproduce with statistical fluctuation. Variogram modeling is not an easy or straightforward task. The development of an appropriate variogram model for a data set requires the understanding and application of advanced statistical concepts and tools. Variogram models cannot be just expressed by any function. They must satisfy a positive

definite condition. One way to satisfy the positive definite condition is to use only a few functions that are known to be positive definite.

I.7.3.1. Spherical model

The spherical model is the most common type of variogram with curves that increase as distance increases up to the range (limit of influence). Beyond the range, mean squared differences do not change and the curve flattens out. Spherical variograms can characteristics of data with well developed areas of influences and good continuities. Spherical variograms have been used to describe such diverse deposits as iron ores, porphyry copper, bauxite, gold, uranium, coal, and phosphate. This model occurs naturally in all deposits where grades become independent of each other once a given distance, a, is reached. It is the common rule in most sedimentary deposits and also in porphyry gold deposits. Its equation is given by:

$$\gamma(|\boldsymbol{h}|) = \boldsymbol{C_0} + \boldsymbol{C}\left[\boldsymbol{1.5}\left(\frac{|\boldsymbol{h}|}{\boldsymbol{a}} - \boldsymbol{0.5}\left(\frac{|\boldsymbol{h}|}{\boldsymbol{a}}\right)^{\boldsymbol{2}}\right), \boldsymbol{for}\ \boldsymbol{0} < |\boldsymbol{h}| < a;\right]$$

$$= \boldsymbol{C_0} + \boldsymbol{C}, \boldsymbol{if}\ |\boldsymbol{h}| > a$$

$$= \boldsymbol{C_0}\,, \boldsymbol{if}\ |\boldsymbol{h}| = \boldsymbol{0} \qquad (\boldsymbol{27})$$

Where $\boldsymbol{a}$ is the range, $\boldsymbol{C_0}$ is the nugget effect and $\boldsymbol{C}$ is the sill value. This model has a linear behavior at the origin and reaches the sill at distance ***(a)***. In fitting this model, it is useful to remember that the tangent at the origin cuts the sill at 2/3 of the range (Rashad et al, 2007).

I.7.3.2. Exponential model

The exponential model is commonly used; its equation it given by:

$$\boldsymbol{\gamma}(|\boldsymbol{h}|) = \boldsymbol{C_0} + \boldsymbol{C}\left(\boldsymbol{1} - \boldsymbol{e}^{\frac{-\boldsymbol{3}|\boldsymbol{h}|}{\boldsymbol{a}}}\right) \qquad (\boldsymbol{28})$$

Where$\boldsymbol{C_0}$is the nugget effect value, $\boldsymbol{C}$ is the sill value and a is the practical range.

This model reaches its sill asymptotically and has a linear behavior at the origin. In fitting this model, it is useful to remember that the tangent at the origin cuts the sill at 1/3 of the practical range.

I.7.3.3. Gaussian model

The Gaussian model is a model used for extremely continuous phenomena. Its equation is given by:

$$\boldsymbol{\gamma(|h|) = C_0 + C\left(1 - e^{\frac{-|3h^2|}{a^2}}\right)} \textbf{(29)}$$

Where $\boldsymbol{C}$ is the positive variance or sill value and a is the practical range, namely, the distance at which the variogram value is 95 % of the sill. It is the only basic model whose shape has inflection point.

I.7.3.4. Power model(linear)

The simplest model that one tries to use as much as possible for obvious reasons is just a linear one. The power model is not a transition model. It does not reach a sill but increases with the magnitude of distance $\boldsymbol{|h|}$. Its equation is given by:

$$\boldsymbol{(|h|) = C_0 + C|h|^2} \textbf{(30)}$$

Where *C* is the positive variance contribution (slope) and a is a power between 0 and 2.

I.7.3.5. Nested model

Some variograms cannot be described by one of the previous equations while, however, they can be adequately represented by the sum of two or more such equations. The overall variogram $\boldsymbol{\gamma(|h|)}$ can be written as:

$$\boldsymbol{\gamma(|h|) = \gamma_1(|h|) + \gamma_2(|h|) + \cdots + \gamma_n(|h|)} \qquad \textbf{(31)}$$

The way to obtain the parameters of this model is by trial and error, knowing that the total variance is$\boldsymbol{C_1 + C_2 + \cdots + C_0}$, much easier way is to use least squares fitting program (software).

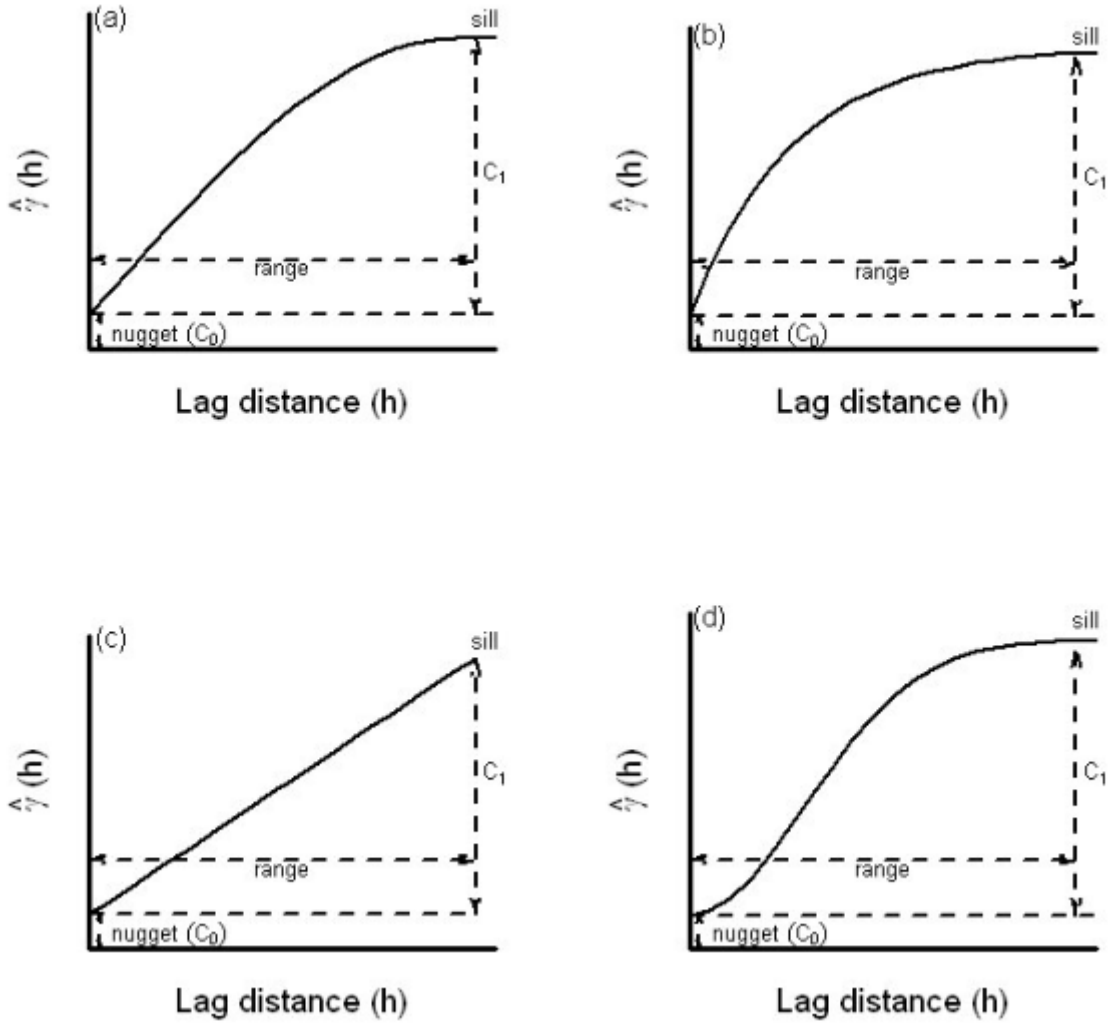

Figure 2: The variogram models. (a) spherical; (b) exponential; (c) linear; and (d) Gaussian. (Rashad et al, 2007).

I.8. The Experimental Variogram

The first task in turning theory into practice is to estimate the variogram from sample data, say $\boldsymbol{z(x_1)}$, $\boldsymbol{z(x_2)}$ …, where $\boldsymbol{x_1}$,$\boldsymbol{x_2}$… denote the positions of the sample in two-dimensional space. We assume that those positions have been selected without bias. They need not be random, as in design-based estimation, because we treat the variables as the outcomes of random processes. Therefore, we can take a relaxed attitude to the sampling design, which may be systematic, random, nested or some combination. The usual equation to compute the variogram is Matheron's method of moments (MoM) estimator:

$$\hat{\gamma}(\boldsymbol{h}) = \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{m}(\boldsymbol{h})} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}(\boldsymbol{h})} \{\boldsymbol{z}(\boldsymbol{x}_i) - \boldsymbol{z}(\boldsymbol{x}_i + \boldsymbol{h})\}^{\mathbf{2}} \qquad (\mathbf{32})$$

where$\boldsymbol{z}(\boldsymbol{x}_i)$ and $\boldsymbol{z}(\boldsymbol{x}_i + \boldsymbol{h})$ are the observed values of z at places xi and $\boldsymbol{x}_i + \boldsymbol{h}$, and $\boldsymbol{m}(\boldsymbol{h})$ is the number of paired comparisons at lag $\boldsymbol{h}$. By changing $\boldsymbol{h}$, we obtain an ordered set of semivariances; these constitute the experimental or sample variogram. The way that equation is implemented as an algorithm depends on whether the data are regularly spaced in one dimension, are on a regular grid or are irregularly distributed in two dimensions (Oliver and Webster, 2015).

I.9. Methodology for variogram interpretation and modeling

The methodology advocated in this paper is classical in that it assumes the regionalized variable is made up of a sum of independent random variables. Each constituent random variable has its own variogram structure. The component variogram structures may be added arithmetically to create a complete 3D variogram model to be used for geostatistical algorithms. Since each variogram structure corresponds to a specific underlying geological phenomenon, the actual modeling phase (traditional curve fitting exercise) is preceded by a necessary interpretation stage.

The steps of the methodology are as follows:

1. Compute and plot experimental variograms in what are believed to be the principal directions of continuity based on a priori geological knowledge. If geological information is ambiguous, one can use 2D variogram maps to determine major horizontal directions of continuity.
2. Place a horizontal line representing the theoretical sill. Use the value of the experimental (stationary) variance for continuous variables, 1 if the data has been transformed to normal score, and $\boldsymbol{p}\ \boldsymbol{(1-p)}$ for categorical variables where p is the global proportion of the category of interest. A feature of our proposed methodology is that the variograms are systematically fitted to the theoretical sill and the whole variance below the sill must be explained in the following steps.
3. If the experimental variogram clearly rises above the theoretical sill, then it is very likely that there exists a trend in the data. The trend should be removed as detailed in the above section Removing the Trend, before proceeding to interpretation of the experimental variogram.

4. Interpretation

- *Short-scale variance:* The Nugget effect is a discontinuity in the variogram at the origin corresponding to short scale variability. On the experimental variogram, it can be due to measurement errors or geological structures with correlation ranges shorter than the sampling resolution. It must be chosen as to be equal in all directions. It is picked from the directional experimental variogram exhibiting the smallest nugget. It is the interpreter's decision to possibly lower it or even set it to 0.0.
- *Intermediate-scale variance:* Geometric anisotropy corresponds to a phenomenon with different correlation ranges in different directions. Each direction encounters the total variability of the structure. There may exist more than one such variance structure.
- *Large-scale variance:* ***(1) Zonal anisotropy*** is characterized by directional variograms reaching a plateau at a variance lower than the theoretical sill, i.e., the whole variability of the phenomenon is not visible in those directions. ***(2) hole effect*** is representative of a "periodic" phenomenon (cyclicity) and characterized by undulations on the variogram. The hole effect does not actually contribute to the total variance of the phenomena; however, its amplitude and frequency must be identified during the interpretation procedure, also, it can only exist in one direction.

5. Modeling Once all the variance regions have been explained and each structure has been related to a known geological process, one may proceed to variogram modeling by selecting a licit model type (spherical, exponential, Gaussian, etc.) and correlation ranges for each structure. This step can be referred to as the parameter estimation part of variogram analysis. Constraining the variogram model by a prior interpretation step with identification of structure types can help fit the experimental variograms (Genton, 1998) (Gringarten and Deutsch, 2000).

Chapter II: Review on Artificial Intelligence

Chapter II: Review on Artificial Intelligence

II.1. Birth of AI and the golden age

The period before the year 1956 is regarded as the incubation period of AI. Until that time, scientists and engineers tried to replace part of their mental work with machines. In 1936, mathematician Alan Turing proposed a mathematical model of an ideal computer, which laid the theoretical foundation for later to come electronic computers. Neurophysiologists W. McCulloch and W. Pitts built the first neural network model (M-P model) in 1943. The M-P model is the first mathematical model constructed to mimic the structure and the working principle of biological neurons. It can be regarded as the earliest artificial neural network. Back in 1949, Hebb proposed the learning mechanism based on neuropsychology. “Hebb learning rule” is an unsupervised learning rule, which can extract statistical features of the training sets and classify data according to data similarity. This is the earliest idea of machine learning (ML), and it is very close to the process of human cognition. In 1952, IBM scientist Arthur Samuel developed a checkers program that could learn implicit models from the current position and instruct subsequent moves. In such context, chess programs were among the earliest work of evolutionary computing. The algorithm compares a modified copy to the best version, and the winner becomes the new standard. It was John McCarthy who originally coined the term AI at the 1956 Dartmouth Summer Research Project on Artificial Intelligence. He is therefore considered to be the father of artificial intelligence. Since this event, the research into AI has rendered many remarkable achievements, including machine learning, theorem proving, pattern recognition, problem-solving, expert systems, and natural language processing.

In 1957, American psychologist Frank Rosenblatt introduced the “Perceptron” model. It can be used to build a system that uses “neurons” for recognition—the capability of learning profoundly impacted the subsequent design of neural networks and the connection mechanisms. The pioneering work on the perceptron is still a trendy topic in the introductory courses for AI today. In 1960, the perceptron algorithm was transmitted to a physical hardware implementation called “Mark 1 Perceptron” as the artificial brain, consisting of an array of perceptions, connected to a camera. At that time, it was magical and inspiring for the people to see that a machine could be trained to distinguish whether a person in a photograph is a male or a female with reasonable accuracy. In the same year, Newell et al. summarized

the thinking rules of humans through psychological experiments. They compiled a general problem-solving program, which could be used to solve 11 different types of problems. A few years later, E. A. Feigenbaum from Stanford University designed an expert system. The system could determine the molecular structure of compounds based on the experimental analysis of mass spectrometers. The study of AI during this period was groundbreaking, with outcomes showing excellent prospects. Meanwhile, the field of computing developed as a field of its own. Transistors and hardware architectures progressed every year, and the essential software and computer programs emerged to implement the AI theories and algorithms (Jiang et al, 2022).

II.2. Artificial Intelligence

II.2.1. What Is Artificial Intelligence?

Artificial intelligence refers to computer systems that are capable of performing tasks traditionally associated with human intelligence — such as making predictions, identifying objects, interpreting speech and generating natural language. AI systems learn how to do so by processing massive amounts of data and looking for patterns to model in their own decision-making. In many cases, humans will supervise an AI's learning process, reinforcing good decisions and discouraging bad ones, but some AI systems are designed to learn without supervision. Over time, AI systems improve on their performance of specific tasks, allowing them to adapt to new inputs and make decisions without being explicitly programmed to do so. In essence, artificial intelligence is about teaching machines to think and learn like humans, with the goal of automating work and solving problems more efficiently.

II.2.2. How Does AI Work?

Artificial intelligence systems work by using algorithms and data. First, a massive amount of data is collected and applied to mathematical models, or algorithms, which use the information to recognize patterns and make predictions in a process known as training. Once algorithms have been trained, they are deployed within various applications, where they continuously learn from and adapt to new data. This allows AI systems to perform complex tasks like image recognition, language processing and data analysis with greater accuracy and efficiency over time.

II.2.3. Types of Artificial Intelligence

Artificial intelligence can be classified in several different ways.

II.2.3.1. Strong AI vs. Weak AI

AI can be organized into two broad categories: weak AI and strong AI.

- Weak AI (or narrow AI) refers to AI that automates specific tasks. It typically outperforms humans, but it operates within a limited context and is applied to a narrowly defined problem. For now, all AI systems are examples of weak AI, ranging from email inbox spam filters to recommendation engines to chatbots.
- Strong AI, often referred to as artificial_general_intelligence_(AGI), is a hypothetical benchmark at which AI could possess human-like intelligence and adaptability, solving problems it's never been trained to work on. AGI does not actually exist yet, and it is unclear whether it ever will.

II.2.4. The 4 Kinds of AI

AI can then be further categorized into four main types: reactive machines, limited memory, theory of mind and self-awareness.

1. **Reactive machines** perceive the world in front of them and react. They can carry out specific commands and requests, but they cannot store memory or rely on past experiences to inform their decision making in real time. This makes reactive machines useful for completing a limited number of specialized duties. Examples include Netflix's recommendation engine and IBM's Deep Blue (used to play chess).
2. **Limited memory** AI has the ability to store previous data and predictions when gathering information and making decisions. Essentially, it looks into the past for clues to predict what may come next. Limited memory AI is created when a team continuously trains a model in how to analyze and utilize new data, or an AI environment is built so models can be automatically trained and renewed. Exemples include ChatGPT and self-Driving cars.
3. **Theory of mind** is a type of AI that does not actually exist yet, but it describes the idea of an AI system that can perceive and understand human emotions, and then use that information to predict future actions and make decisions on its own.

4. **Self-aware AI** refers to artificial intelligence that has self-awareness, or a sense of self. This type of AI does not currently exist. In theory, though, self-aware AI possesses human-like consciousness and understands its own existence in the world, as well as the emotional state of others (Glover, 2024).

II.2.5. Using Artificial Intelligence

Artificial intelligence can be applied to many sectors and industries, including the healthcare industry for suggesting drug dosages, identifying treatments, and aiding in surgical procedures in the operating room.Other examples of machines with artificial intelligence include computers that play chess and self-driving cars. AI has applications in the financial industry, where it detects and flags fraudulent banking activity. Applications for AI can help streamline and make trading easier.

In 2022, AI entered the mainstream with applications of the Generative Pre-Trained Transformer (GPT). The most popular applications are OpenAI's DALL-E text-to-image tool and ChatGPT. According to a 2024 survey by Deloitte, 79% of respondents who are leaders in the AI industry expect generative AI to transform their organizations by 2027(The Investopedia Team, 2025).

II.2.6. What is artificial general intelligence (AGI)?

Artificial general intelligence (AGI) refers to a theoretical state in which computer systems will be able to achieve or exceed human intelligence. In other words, AGI is "true" artificial intelligence, as depicted in countless science fiction novels, television shows, movies, and comics. As for the precise meaning of "AI" itself, researchers don't quite agree on how we would recognize "true" artificial general intelligence when it appears. However, the most famous approach to identifying whether a machine is intelligent or not is known as the Turing Test or Imitation Game, an experiment that was first outlined by influential mathematician, computer scientist, and cryptanalyst Alan Turing in a 1950 paper on computer intelligence. There, Turing described a three-player game in which a human "interrogator" is asked to communicate via text with another human and a machine and judge who composed each response. If the interrogator cannot reliably identify the human, then Turing says the machine can be said to be intelligent.

To complicate matters, researchers and philosophers also can't quite agree whether we're beginning to achieve AGI, if it's still far off, or just totally impossible. For example,

while a recent paper from Microsoft Research and OpenAI argues that Chat GPT-4 is an early form of AGI, many other researchers are skeptical of these claims and argue that they were just made for publicity.Regardless of how far we are from achieving AGI, you can assume that when someone uses the term artificial general intelligence, they're referring to the kind of sentient computer programs and machines that are commonly found in popular science fiction(Coursera Staff, 2024).

II.2.7. AI technology

In the early 21st century faster processing power and larger datasets ("big data") brought artificial intelligence out of computer science departments and into the wider world. Moore's law, the observation that computing power doubled roughly every 18 months, continued to hold true. The stock responses of the early chatbot Eliza fit comfortably within 50 kilobytes; the language model at the heart of ChatGPT was trained on 45terabytes of text (Copeland,2025).

II.2.8. How Companies Are Using AI Today

AI and machine learning can lead to a variety of automated tasks. The technology affects virtually every industry from IT security malware search, to weather forecasting, to stockbrokers looking for optimal trades. Machine learning in particular requires complex math and a lot of coding to achieve the desired functions and results. Machine learning also incorporates classical algorithms for various kinds of tasks such as clustering, regression or classification. We have to train these algorithms on large amounts of data. The more data you provide for your algorithm, the better your model and desired outcome gets (Oppermann, 2023).

II.3. Machine learning

II.3.1. What is Machine Learning?

Machine Learning, often abbreviated as ML, is a subset of artificial intelligence (AI) that focuses on the development of computer algorithms that improve automatically through experience and by the use of data. In simpler terms, machine learning enables computers to learn from data and make decisions or predictions without being explicitly programmed to do so.

At its core, machine learning is all about creating and implementing algorithms that facilitate these decisions and predictions. These algorithms are designed to improve their performance over time, becoming more accurate and effective as they process more data.In traditional programming, a computer follows a set of predefined instructions to perform a task. However, in machine learning, the computer is given a set of examples (data) and a task to perform, but it's up to the computer to figure out how to accomplish the task based on the examples it's given.

II.3.2. Types of Machine Learning

Machine learning can be broadly classified into three types based on the nature of the learning system and the data available: supervised learning, unsupervised learning, and reinforcement learning.

II.3.2.1. Supervised learning

Supervised learning is the most common type of machine learning. In this approach, the model is trained on a labeled dataset. In other words, the data is accompanied by a label that the model is trying to predict. This could be anything from a category label to a real-valued number. The model learns a mapping between the input (features) and the output (label) during the training process. Once trained, the model can predict the output for new, unseen data.

Common examples of supervised learning algorithms include **linear regression** for regression problems and logistic regression, **decision trees**, and support vector machines for classification problems.

II.3.2.2. Unsupervised learning

Unsupervised learning, on the other hand, involves training the model on an unlabeled dataset. The model is left to find patterns and relationships in the data on its own. This type of learning is often used for clustering and dimensionality reduction. Clustering involves grouping similar data points together, while dimensionality reduction involves reducing the number of random variables under consideration by obtaining a set of principal variables.

Common examples of unsupervised learning algorithms include k-means for clustering problems and Principal Component Analysis (PCA) for dimensionality reduction problems.

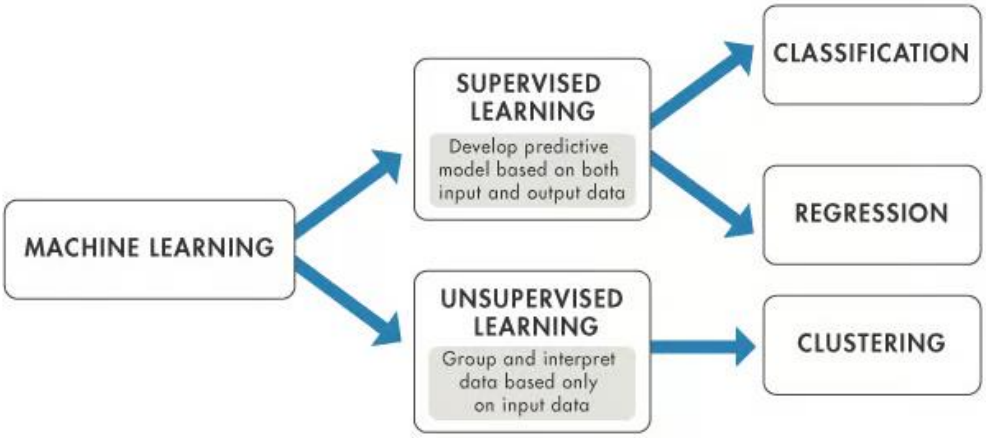

Figure 3 :Scientific Exploration of Conceptual and Algorithmic Terminologies of Machine Learning (Singh, 2022).

II.3.2.3. Reinforcement learning

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with its environment. The agent is rewarded or penalized (with points) for the actions it takes, and its goal is to maximize the total reward.Unlike supervised and unsupervised learning, reinforcement learning is particularly suited to problems where the data is sequential, and the decision made at each step can affect future outcomes (Crabtree, 2024).

II.3.3. How does supervised machine learning work?

Supervised learning supplies algorithms with labeled training data and defines which variables the algorithm should assess for correlations. Both the input and output of the algorithm are specified. Initially, most ML algorithms used supervised learning, but unsupervised approaches are gaining popularity.

Supervised learning algorithms are used for numerous tasks, including the following:

- **Binary classification.** This divides data into two categories.
- **Multiclass classification.** This chooses among more than two categories.

- **Ensemble modeling.** This combines the predictions of multiple ML models to produce a more accurate prediction.
- **Regression modeling.** This predicts continuous values based on relationships within data.

II.3.4. How does unsupervised machine learning work?

Unsupervised learning doesn't require labeled data. Instead, these algorithms analyze unlabeled data to identify patterns and group data points into subsets using techniques such as gradient descent. Most types of deep learning, including neural networks, are unsupervised algorithms.

Unsupervised learning is effective for various tasks, including the following:

- Splitting the data set into groups based on similarity using clustering algorithms.
- Identifying unusual data points in a data set using anomaly detection algorithms.
- Discovering sets of items in a data set that frequently occur together using association rule mining.
- Decreasing the number of variables in a data set using dimensionality reduction techniques.

II.3.5. How does reinforcement learning work?

Reinforcement learning involves programming an algorithm with a distinct goal and a set of rules to follow in achieving that goal. The algorithm seeks positive rewards for performing actions that move it closer to its goal and avoids punishments for performing actions that move it further from the goal.

Reinforcement learning is often used for tasks such as the following:

- Helping robots learn to perform tasks in the physical world.
- Teaching bots to play video games.
- Helping enterprises plan allocation of resources. (Lev, 2024).

II.3.6. How Does Machine Learning Work?

Understanding how machine learning works involves delving into a step-by-step process that transforms raw data into valuable insights. Let's break down this process:

Step 1: Data collection

The first step in the machine learning process is data collection. Data is the lifeblood of machine learning; the quality and quantity of your data can directly impact your model's performance. Data can be collected from various sources such as databases, text files, images, audio files, or even scraped from the web.Once collected, the data needs to be prepared for machine learning. This process involves organizing the data in a suitable format, such as a CSV file or a database, and ensuring that the data is relevant to the problem you're trying to solve.

Step 2: Data preprocessing

Data preprocessing is a crucial step in the machine learning process. It involves cleaning the data (removing duplicates, correcting errors), handling missing data (either by removing it or filling it in), and normalizing the data (scaling the data to a standard format).Preprocessing improves the quality of your data and ensures that your machine learning model can interpret it correctly. This step can significantly improve the accuracy of your model.

Step 3: Choosing the right model

Once the data is prepared, the next step is to choose a machine learning model. There are many types of models to choose from, including linear regression, decision trees, and neural networks. The choice of model depends on the nature of your data and the problem you're trying to solve.Factors to consider when choosing a model include the size and type of your data, the complexity of the problem, and the computational resources available.

Step 4: Training the model

After choosing a model, the next step is to train it using the prepared data. Training involves feeding the data into the model and allowing it to adjust its internal parameters to better predict the output.During training, it's important to avoid overfitting (where the model performs well on the training data but poorly on new data) and underfitting (where the model performs poorly on both the training data and new data).

Step 5: Evaluating the model

Once a model is trained, evaluating its performance on unseen data is essential before deployment. With MLOps, monitoring doesn’t stop at this initial stage; it involves

ongoing evaluation to detect model drift (when a model's performance declines due to changes in data patterns) and maintaining model quality over time. Continuous monitoring and retraining workflows help organizations ensure their models remain effective and reliable in production environments.

Step 6: Hyperparameter tuning and optimization

Beyond tuning for accuracy, hyperparameter optimization within an MLOps pipeline includes tools for automated hyperparameter searches, ensuring efficiency and reproducibility. Many teams employ MLOps platforms that support hyperparameter tuning, so experiments are repeatable and well-documented, allowing for consistent optimization over time.Techniques for hyperparameter tuning include grid search (where you try out different combinations of parameters) and cross validation (where you divide your data into subsets and train your model on each subset to ensure it performs well on different data).

Step 7: Predictions and deployment

Deploying a machine learning model involves integrating it into a production environment, where it can deliver real-time predictions or insights. MLOps (Machine Learning Operations) has emerged as a standard practice to streamline this process. It encompasses version control, monitoring, and automated testing to ensure models are reproducible, reliable, and robust. MLOps frameworks like MLflow or Kubeflow support these goals by providing seamless workflows for deployment, retraining, and model rollback if issues arise(Crabtree, 2024).

II.3.7. How businesses are using machine learning

Machine learning is the core of some companies' business models, like in the case of Netflix's suggestions algorithm or Google's search engine. Other companies are engaging deeply with machine learning, though it's not their main business proposition.Others are still trying to determine how to use machine learning in a beneficial way. "In my opinion, one of the hardest problems in machine learning is figuring out what problems I can solve with machine learning.Companies are already using machine learning in several ways, including:

Recommendation algorithms. The recommendation engines behind Netflix and YouTube suggestions, what information appears on your Facebook feed, and product

recommendations are fueled by machine learning. “[The algorithms] are trying to learn our preferences,” Madry said. “They want to learn, like on Twitter, what tweets we want them to show us, on Facebook, what ads to display, what posts or liked content to share with us.”

Image analysis and object detection. Machine learning can analyze images for different information, like learning to identify people and tell them apart — though **facial recognition algorithms are controversial.** Business uses for this vary. Shulman noted that hedge funds famously use machine learning to analyze the number of cars in parking lots, which helps them learn how companies are performing and make good bets.

Fraud detection. Machines can analyze patterns, like how someone normally spends or where they normally shop, to identify potentially fraudulent credit card transactions, log-in attempts, or spam emails.

Automatic helplines or chatbots. Many companies are deploying online chatbots, in which customers or clients don’t speak to humans, but instead interact with a machine. These algorithms use machine learning and natural language processing, with the bots learning from records of past conversations to come up with appropriate responses.

Self-driving cars. Much of the technology behind self-driving cars is based on machine learning, deep learning in particular.

Medical imaging and diagnostics. Machine learning programs can be trained to examine medical images or other information and look for certain markers of illness, like a tool that can predict cancer risk based on a mammogram (Brown, 2021).

II.4. Deep machine learning

II.4.1. What is Deep Learning?

Deep learning is a type of machine learning that teaches computers to perform tasks by learning from examples, much like humans do. Imagine teaching a computer to recognize cats: instead of telling it to look for whiskers, ears, and a tail, you show it thousands of pictures of cats. The computer finds the common patterns all by itself and learns how to identify a cat. This is the essence of deep learning. In technical terms, deep learning uses something called "neural networks," which are inspired by the human brain. These networks consist of layers of interconnected nodes that process information. The more

layers, the "deeper" the network, allowing it to learn more complex features and perform more sophisticated tasks.

II.4.2. Deep Learning Models

Let's learn about different types of deep learning models and how they work.

II.4.2.1. Supervised Learning

Supervised learning uses a labeled dataset to train models to either classify data or predict values. The dataset contains features and target labels, which allow the algorithm to learn over time by minimizing the loss between predicted and actual labels. Supervised learning can be divided into classification and regression problems.

II.4.2.2. Classification

The classification algorithm divides the dataset into various categories based on feature extractions. The popular deep learning models are **ResNet50** for image classification and **BERT (language model)**for text classification.

II.4.2.3. Regression

Instead of dividing the dataset into categories, the regression model learns the relationship between input and output variables to predict the outcome. Regression models are commonly used for predictive analysis, weather forecasting, and predicting stock market performance. **LSTM (Long short-term memory)** and **RNN (Recurrent neural networks)** are popular deep learning regression models.

II.4.2.4. Unsupervised Learning

Unsupervised learning algorithms learn the pattern within an unlabeled dataset and create clusters. Deep learning models can learn hidden patterns without human intervention and these models are often used in recommendation engines. Unsupervised learning is used for grouping various species, medical imaging, and market research. The most common deep learning model for clustering is the deep embedded clustering algorithm (Awan, 2023).

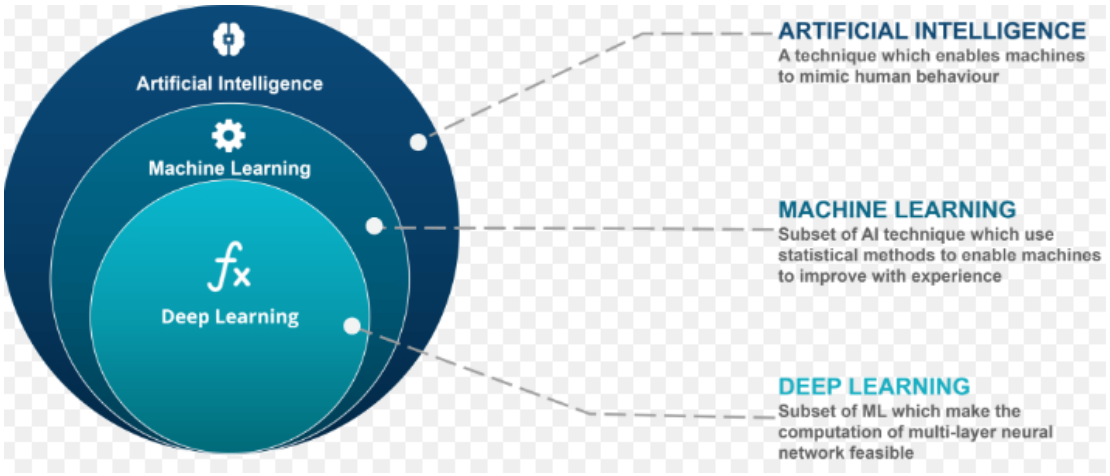

Figure 4: Artificial Intelligence Deep Learning Machine Learning (Singh, 2022).

II.5. Generative AI

Generative AI refers to deep-learning models that can take raw data — say, all of Wikipedia or the collected works of Rembrandt — and "learn" to generate statistically probable outputs when prompted. At a high level, generative models encode a simplified representation of their training data and draw from it to create a new work that's similar, but not identical, to the original data.Generative models have been used for years in statistics to analyze numerical data. The rise of deep learning, however, made it possible to extend them to images, speech, and other complex data types. Among the first class of models to achieve this cross-over feat were variational autoencoders, or VAEs, introduced in 2013. VAEs were the first deep-learning models to be widely used for generating realistic images and speech (Martineau, 2023).

II.6. Neural network

II.6.1. What Is a Neural Network?

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature.Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

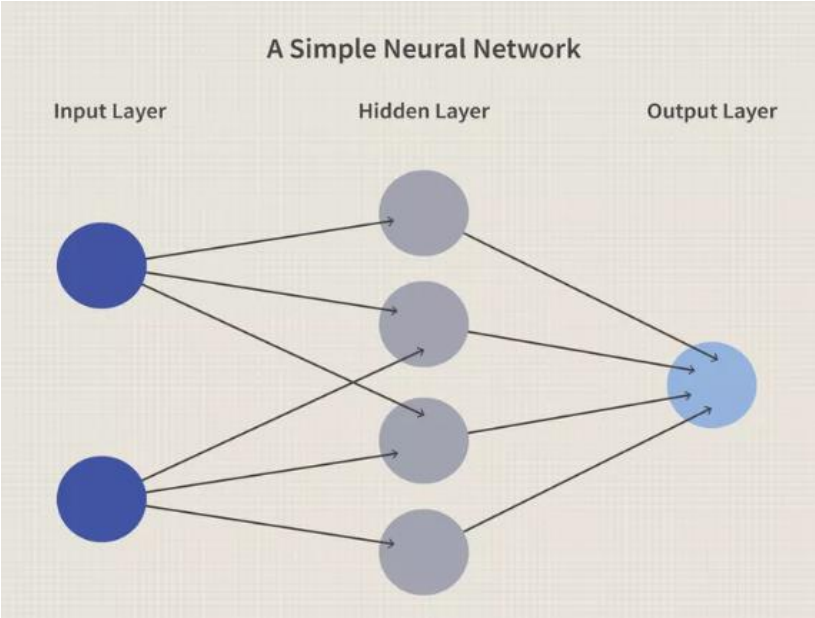

Figure 5 : Simple neural network (Chen, 2024).

II.6.2. Types of Neural Networks

II.6.2.1. Feed-Forward Neural Networks

Feed-forward neural networks are one of the simpler types of neural networks. It conveys information in one direction through input nodes; this information continues to be processed in this single direction until it reaches the output mode. Feed-forward neural networks may have hidden layers for functionality, and this type of most often used for facial recognition technologies.

II.6.2.2. Recurrent Neural Networks

A more complex type of neural network, recurrent neural networks take the output of a processing node and transmit the information back into the network. This results in theoretical "learning" and improvement of the network. Each node stores historical processes, and these historical processes are reused in the future during processing.This becomes especially critical for networks in which the prediction is incorrect; the system will attempt to learn why the correct outcome occurred and adjust accordingly. This type of neural network is often used in text-to-speech applications.

II.6.2.3. Convolutional Neural Networks

Convolutional neural networks, also called ConvNets or CNNs, have several layers in which data is sorted into categories. These networks have an input layer, an output layer,

and a hidden multitude of convolutional layers in between. The layers create feature maps that record areas of an image that are broken down further until they generate valuable outputs. These layers can be pooled or entirely connected, and these networks are especially beneficial for image recognition applications.

II.6.2.4. Deconvolutional Neural Networks

Deconvolutional neural networks simply work in reverse of convolutional neural networks. The application of the network is to detect items that might have been recognized as important under a convolutional neural network. These items would likely have been discarded during the convolutional neural network execution process. This type of neural network is also widely used for image analysis or processing.

II.6.2.5. Modular Neural Networks

Modular neural networks contain several networks that work independently from one another. These networks do not interact with each other during an analysis process. Instead, these processes are done to allow complex, elaborate computing processes to be done more efficiently. Similar to other modular industries such as modular real estate, the goal of the network independence is to have each module responsible for a particular part of an overall bigger picture (Chen, 2024).

II.7. Multilayer Feedforward NN (MLFFNN)

Feedforward NNs are the artificial NNs in which theconnections between units do notform a cycle.Feedforward NNs were the first type of artificial NNinvented. In addition,they are simpler than theircounterpart, the RNNs. They are called Feedforwardbecause information only travels forward in the network (no loops), first through the input nodes, then throughthe hidden nodes (if present), and in final, through theoutput nodes. The Feedforward NN divides into twomain types as follows.

II.8. Single-Layer Feedforward NN (SLFFNN)

The Single layer is referring to the output layer ofthe computation nodes (neurons). Theinput layer ofsource nodes is not counted because no computationis performed there. The SLFFNN is shown in Figure 6:

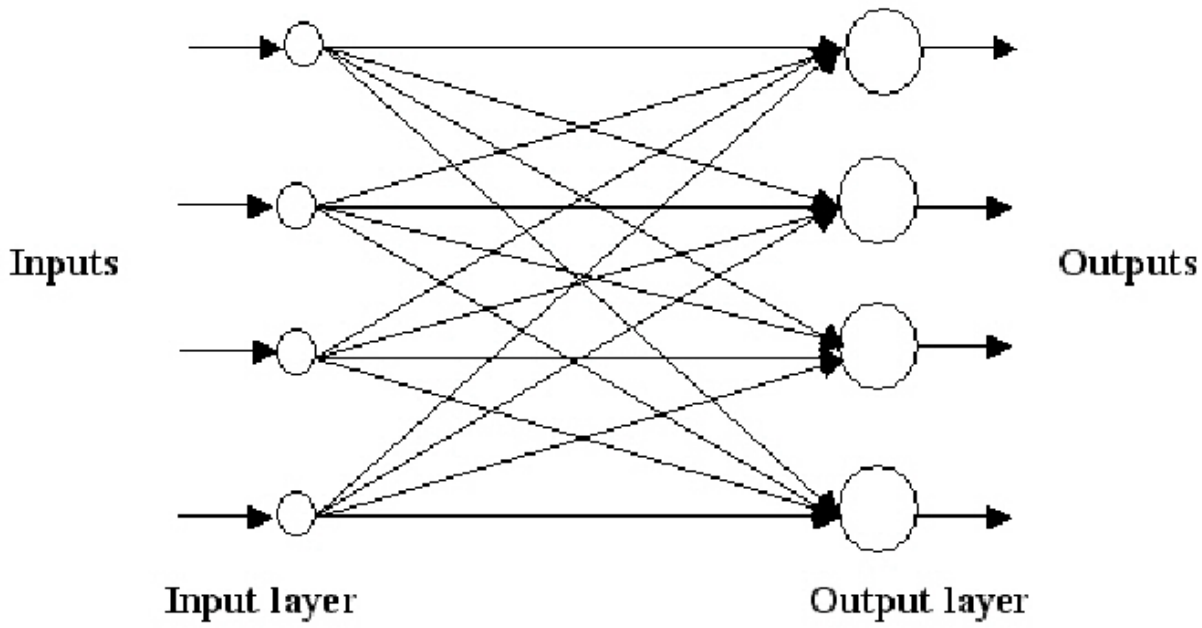

Figure 6: A single layer feed-forward neural network (Hüsnü SAZLI, 2006).

II.9. Multilayer feed-forward NN

The architecture of this NN has one or more hiddenlayers. Figure shows fully connected feedforwardnetwork with one hidden layer and one output layer (Sharkawy, 2020).

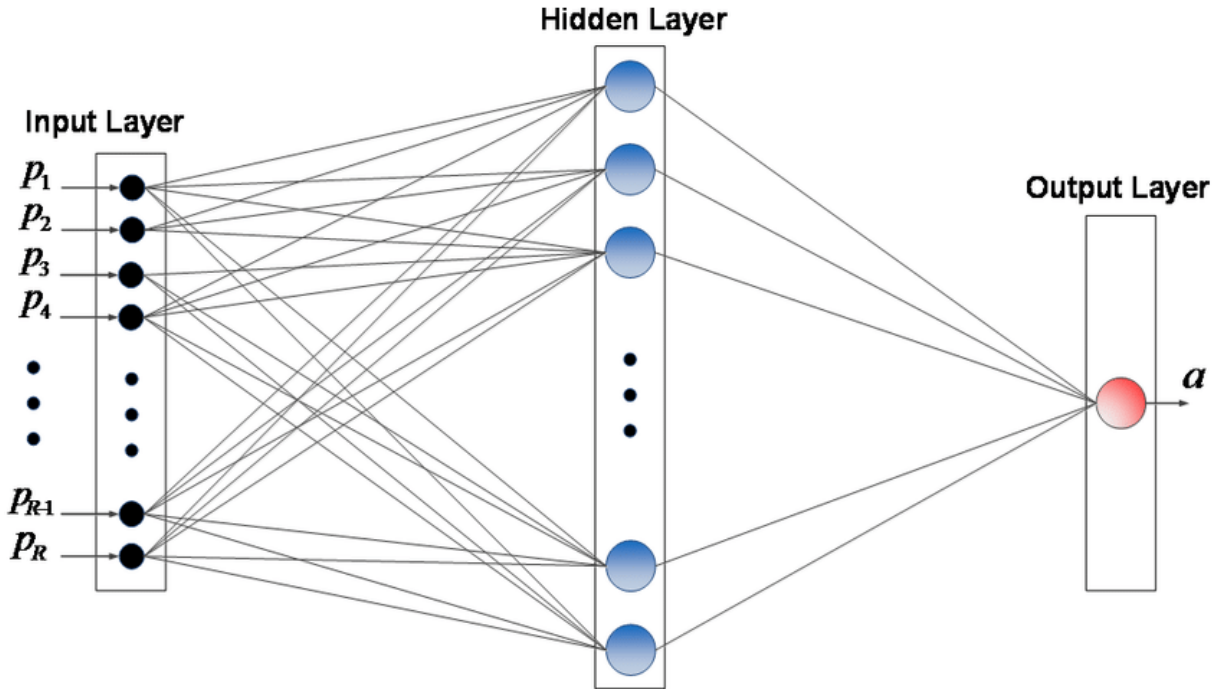

Figure 7: Structure of the multilayer feed-forward neural network. (Chen et al, 2017).

II.10. Statistical models in soils

II.10.1. Artificial Bee Colony model

II.10.1.1. What is Artificial Bee Colony model

Artificial Bee Colony Algorithm (ABC) is nature-inspired metaheuristic, which imitatesthe foraging behavior of bees. ABC as a stochastic technique is easy toimplement, has fewercontrolparameters, and could easily be modify and hybridized with other metaheuristic algorithms. Due to itssuccessful implementation, several researchers in the optimization and artificial intelligence domains haveadopted it to be the main focus of their research work. Since 2005, several related works have appeared toenhance the performance of the standard ABC in the literature, to meet up with challenges ofrecent research problems being encountered. Interestingly, ABC has been tailored successfully, to solvea wide variety of discrete and continuous optimization problems. Some other works have modified andhybridized ABC to other algorithms, to further enhance the structure of its framework.

II.10.1.2. Growth of ABC algorithm in the literature

In 2005, ABC was developed and evaluated, using multidimensional and multivariable optimization problems. Later on in 2006, the performance of ABC was compared to Genetic Algorithm (GA) on numeric functional optimization and was later used in 2007 for the training of artificial neural networks.Subsequently in 2007, its performance on different problems was extensively studied and the results compared with other well known, successful algorithms such as GA, Particle swarm optimization (PSO), particle swarm inspired evolutionary (PS-EA), differential evolution (DE), back propagation (BP) algorithms.Similarly in 2008, the performance of ABC was compared with DE, PSO and EA with regards to multidimensional numeric problems. It is important to note that most work carried out on ABC from year 2005 to 2008 were by Karaboga and his colleagues.

Year 2009 witnessed major studies on ABC across different disciplines, where different modifications and hybridizations were attempted by various researchers to tackle various problems in engineering, digital image processing and pattern recognition, protein structure predictions, numerical, real parameter and complex optimization, information technology, to name the major ones.

By 2010, the diversification of ABC into other disciplines was also reported. Prominent among these include scheduling, real parameter and other optimization problems, engineering design and applications, information and applied technology, and protein structure prediction. Modifications based on best global algorithm for numerical function were also reported in the same year.

In the year 2011, tremendous increase in the number of ABC publications was witnessed, where series of applications, modifications, parameters tuning and hybridization with different optimization algorithms were used to enhance the performance ABC across various fields.

II.10.1.3. Advantages of ABC

The major advantages which ABC holds over other optimization algorithms include its:

• Simplicity, flexibility and robustness

• Use of fewer control parameters compared to many other search techniques

• Ease of hybridization with other optimization algorithms

• Ability to handle the objective cost with stochastic nature

• Ease of implementation with basic mathematical and logical operations.

II.10.2. Fundamentals to the ABC

II.10.2.2. Artificial Bee Colony: Analogy

The ABC consists of three groups of artificial bees: employed foragers, onlookers and scouts. The employed bees comprise the first half of the colony whereas the second half consists of the onlookers. The employed bees are linked to particular food sources. In other words, the number of employed bees is equal to the number of food sources for the hive. The onlookers observe the dance of the employed bees within the hive, to select a food source, whereas scouts search randomly for new food sources. Analogously in the optimization context, the number of food sources (that is the employed or onlooker bees) in ABC algorithm, is equivalent to the number of solutions in the population. Furthermore, the position of a food source signifies the position of a promising solution to the optimization problem, whereas the quality of nectar of a food source represents the fitness cost (quality) of the associated solution.

The search cycle of ABC consists of three rules:

1. sending the employed bees to a food source and evaluating the nectar quality;
2. onlookers choosing the food sources after obtaining information from employed bees and calculating the nectar quality;
3. determining the scout bees and sending them onto possible food sources.

The positions of the food sources are randomly selected by the bees at the initialization stage and their nectar qualities are measured. The employed bees then share the nectar information of the sources with the bees waiting at the dance area within the hive. After sharing this information, every employed bee returns to the food source visited during the previous cycle, since the position of the food source had been memorized and then selects another food source using its visual information int he neighborhood of the present one. At the last stage, an onlooker uses the information obtained from the employed bees at the dance area to select a food source. The probability for the food sources to be selected increases with increase in its nectar quality. Therefore, the employed bee with information of a food source with the highest nectar quality recruits the onlookers to that source. It subsequently chooses another food source in the neighborhood of the one currently in her memory based on visual information (i.e. comparison of food source positions). A new food source is randomly generated by a scout bee to replace the one abandoned by the onlooker bees. (La'aro Bolaji et al, 2013).

II.10.2.3. Artificial Bee Colony: Procedure

The ABC consists of four main phases:

1. **Initialization Phase:** The food sources, whose population size is SN, are randomly generated by scout bees. Each food source, represented by $\boldsymbol{x_m}$ is an input vector to the optimization problem, $\boldsymbol{x_m}$has D variables and D is the dimension of searching space of the objective function to be optimized. The initial food sources are randomly produced via the expression **(33)**

$$\boldsymbol{x_m = l_i + rand(0,1) * (u_i - l_i)} \qquad \boldsymbol{(33)}$$

Where $\boldsymbol{u_i}$and $\boldsymbol{l_i}$are the upper and lower bound of the solution space of objective function, rand (0, 1) is a random number within the range [0, 1].

2. **Employed Bee Phase:** Employed bee flies to a food source and finds a new food source within the neighborhood of the food source. The higher quantity food source is memorized by the employed bees. The food source information stored by employed bee will be shared with onlooker bees. A neighbor food source $\boldsymbol{v_{mi}}$ is determined and calculated by the following equation **(34)**

$$\boldsymbol{v_{mi} = x_{mi} + \phi_{mi}(x_{mi} - x_{ki})} \qquad (34)$$

Where $\boldsymbol{i}$is a randomly selected parameter index, $\boldsymbol{x_k}$is a randomly selected food source,ϕ_m is a random number within the range [-1, 1]. The range of this parameter can make an appropriate adjustment on specific issues. The fitness of food sources is essential in order to find the global optimal. The fitness is calculated by the following formula **(35)**, after that a greedy selection is applied between$\boldsymbol{x_m}$ and $\boldsymbol{v_m}$.

$$\boldsymbol{fit_m(x_m)} = \begin{pmatrix} \frac{1}{1+f_m(x_m)}, f_m(x_m)>0 \\ 1+|f_m(x_m)|, f_m(x_m)<0 \end{pmatrix} \qquad \textbf{(35)}$$

Where $\boldsymbol{f_m(x_m)}$ is the objective function value of $\boldsymbol{x_m}$.

3. **Onlooker Bee Phase:** Onlooker bees calculate the profitability of food sources by observing the waggle dance in the dance area and then select a higher food source randomly. After that onlooker bees carry out randomly search in the neighborhood of food source. The quantity of a food source is evaluated by its profitability and the profitability of all food sources.$\boldsymbol{p_m}$is determined by the formula:

$$\boldsymbol{p_m} = \frac{\boldsymbol{fit_m(x_m)}}{\sum_m^{SN} \boldsymbol{fit_m(x_m)}} \qquad (36)$$

Where $\boldsymbol{fit_m(x_m)}$is the fitness of$\boldsymbol{x_m}$. Onlooker bees search the neighborhood of food source according to the expression **(37)**:

$$\boldsymbol{v_{mi} = x_{mi} + \phi_{mi}(x_{mi} - x_{ki})} \qquad (37)$$

4. **Scout Phase:** If the profitability of food source cannot be improved and the times of unchanged greater than the predetermined number of trials, which called "limit", the solutions will be abandoned by scout bees. Then, the new solutions are

randomly searched by the scout bees. The new solution $\boldsymbol{x_{mi}}$ will be discovered by the scout by using the expression **(38)** (Balwant and Dharmender, 2013).

$$\boldsymbol{x_m = l_i + rand(0,1) * (u_i - l_i)} \qquad (\mathbf{38})$$

Chapter III: Geotechnical Analysis and Soil Mapping

Chapter III: Geotechnical Analysis and Soil Mapping

III. Introduction

the purpose of the chapter is to identify the study area (Tebessa) from the geographic, geological, hydrogeological and climatological data (precipitation, temperature and humidity), and also the soil classification according to geotechnical parameters according to GTR (version 2023**).** Geotechnical maps were generated using Surfer software to illustrate the spatial distribution of soil types and key parameters such as bearing capacity and consistency. These results provide a valuable foundation for engineering decision-making, construction planning, and geotechnical risk assessment in the Tebessa area.

III.1. Geography of Tebessa

Tebessa is located in the far east of Algeria, at the “gates” of the Desert, about 230 km south of Annaba on the Mediterranean coast. The region is bound to the south by the province of El Oued, to the west by Constantine and to the east by Tunisia (Sedrati and Djabri, 2014).

The province covers an area of 13,878 km² and is bordered by:

- Souk-Ahras to the north.
- Oum-El-Bouaghi and Khenchela to the northwest.
- Tunisia to the east (a 300-km border).
- El-Oued to the south (Achou, 2024).

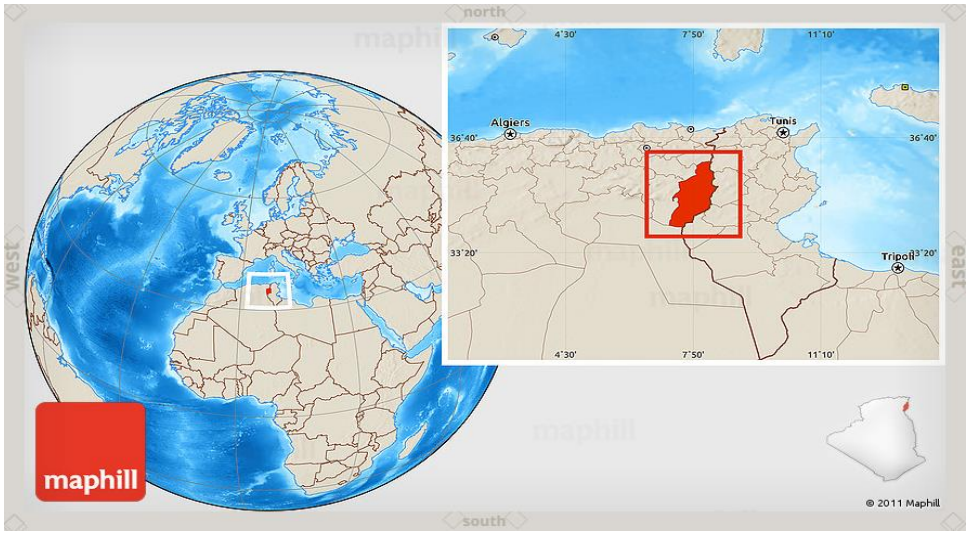

Figure 8: The geography of the region of Tebessa (map hill site, 2025).

Figure 9: The boundaries of the region of Tebessa (map hill site, 2025).

III.2. Geology of Tebessa

The geology of Tebessa is characterized by emergent formations of the Quaternary at the center of the plain; the latter consists of deposits. These deposits are distributed within the lower parts of reliefs and cover large surface areas (plains and present valleys). They are formed of limestone crusts, silt debris, gravel and gypsum conglomerates. This material is altered and transported by wind and water action. The Quaternary aquifer of the continental origin has a thickness of 10 to 30 m. Also note the presence of the Triassic formations, particularly those of Jebel Djebissa. This suggests that sheets of water are subject to salinization caused by these formations. The Turonian, Maastrichtian and Campanian are represented by limestone reliefs constituting the northeastern and southern boundaries of the plain, which also form important groundwater aquifers (Sedrati and Djabri, 2014).

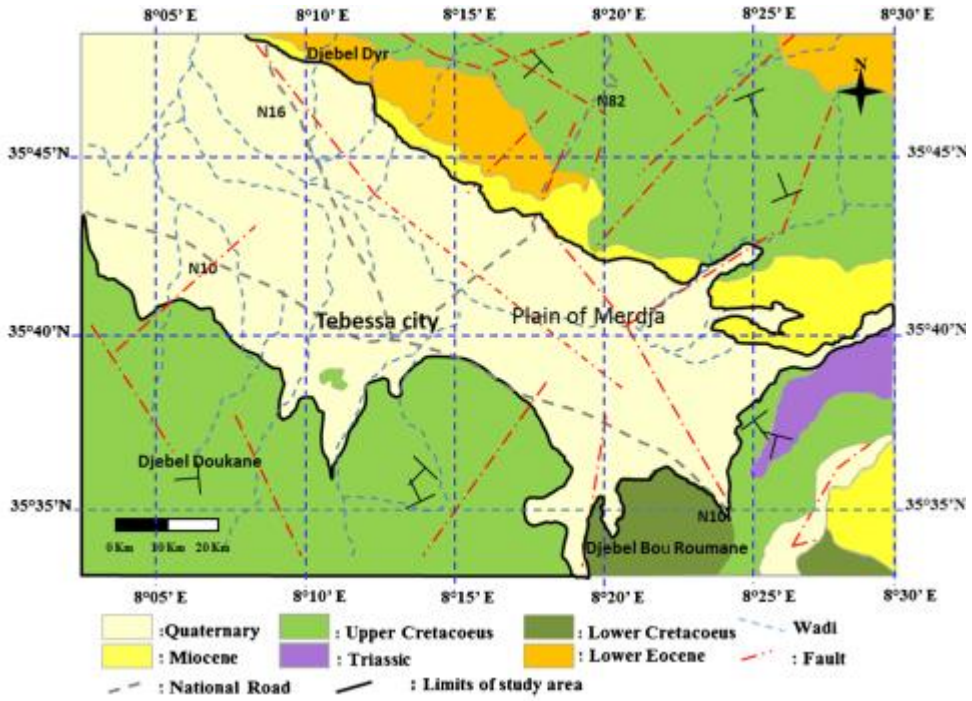

Figure 10: Simplified geological map of the study area (Djalali et al, 2022).

III.2.1. Geological characterization

III.2.1.1. Triassic

The Triassic outcrops in many places thanks to diapiric outcrops, which were first observed in the Eocene. It constitutes the majority of Djebissa at the eastern end of the plain. It is found south of Djebissa, near Djebissa North, and finally north of Morsott, where the outcrop extends widely towards El Aouinet and Souk Ahras.It is characterized by large masses of bluish-gray and sometimes reddish gypsum, as well as green and variegated clays. This facies indicates the existence of large lagoons where large quantities of marl, gypsum, and probably salt was deposited (F. Morel, 1955; Buisson, 1974).

III.2.1.2. Jurassic

The absence of Jurassic deposits has been noted in the region (Dubourdieu, 1956; Chevenine et al., 1989), although some oil drilling has encountered them outside the diapiric zones.

III.2.1.3. Barremian

The Barremian formations consist mainly of gray or yellow, fairly clayey marls containing pyritic ammonites. Toward the top, these marls are intercalated with thin layers of light gray nodular limestone with an ochre patina (Dubourdieu, 1956). These formations do not exceed 250 meters in thickness. They were deposited on a shoal relatively close to the surface (Dubourdieu, 1956), reflecting the shallow-sea sedimentation conditions that persisted until the early Aptian.

III.2.1.4. Aptian

Over vast areas, the Aptian is of great importance in the "diapir zone" of the Tebessa region due to the nature of its deposits and their substantial development (300 to 600 meters thick). It is characterized by neritic carbonate rocks that formed in warm, shallow waters. These conditions favored the development of organ genic sedimentation processes, leading to the formation of various structures. During the Aptian, the sea was more extensive than during the Barremian. The most remarkable phenomenon of this period is the formation of reefs on elongated ridges.

III.2.1.5. Albian

Generally, the Albian Formation begins with a carbonate facies described as "reef" in its lower part, then becomes marly and marly-limestone in its middle to upper part. In the Tebessa region, it is characterized by benthic fauna in shallow areas. In the Late Albian, the Albo-Aptian cover was interrupted by the Triassic, leading to the straightening of the layers and the complexification of the structures (flared folds, mushroom folds, etc.) generally sealed by the Vraconian (Bouzenoune, 1993; Othmanine, 1987; Nedjari-Belhacène and Nedjari, 1984). It is worth noting a monotonous series of dark gray marls to clays, with intercalations of limestones to black marls. Its thickness varies from 20 to 150 meters.

III.2.1.6. Vraconian

It is generally composed of marls with intercalations of argillaceous limestones and argillites with a thickness of 500 to 600 meters (Dubourdieu, 1956). In some regions, it generally transgresses into the Triassic (Thibiéroz and Madre, 1976) and the Aptian (Othmanine, 1987). With its marl-clay sedimentation, the Vraconian forms an immediate screen and exerts a primary control over polymetallic mineralization.

III.2.1.7. Cenomanian

In the Tebessa region, a monotonous series of greenish clayey marls was deposited in the Lower Cenomanian. In the Middle Cenomanian, these marls exhibited various poorly developed limestone intercalations, with more or less abundant fauna and fibrous calcite veinlets.In the Upper Cenomanian, the sedimentation became carbonate. The total thickness of the Cenomanian formation is estimated at between 750 and 1,100 meters. The Cenomanian sedimentation was deposited in abyssal (deep) conditions. This period is characterized by strike-slip faults reflecting NE-SW shortening (Othmanine, 1987). As the Turonian approaches, the sedimentation changes rapidly, and fossils disappear. The clayey marls are replaced by layers rich in calcium carbonate (Dubourdieu, 1956). The upper limit of the Cenomanian is difficult to distinguish (Dubourdieu, 1956).

III.2.1.8. Turonian

It is distinguished by its rapid change in sedimentation towards compact limestones, giving rise to marked reliefs. The Turonian forms the flanks of major anticlines and synclines, and its formations also outcrop outside the diapiric zones. The lithological analysis carried out by Salmi-Laouar (2004) reveals that the lower part of the Turonian in the Essouabaa massif, approximately 200 meters thick, consists of a series of strato-crescent alternations of marls, marly limestones, and limestones. The marls, sometimes clayey and grayish in color, vary in thickness from meters to several meters. The marly limestone beds and fine micritic limestones also have thicknesses ranging from meters to several meters.

III.2.1.9. Campanian-Santonian

These periods are not subdivided due to the lack of dating evidence. They are characterized by greenish-gray and yellow-gray clayey marls containing fibrous calcite plates, with intercalations of lumachelle marls. Their thickness varies from 200 to 600 meters.

III.2.1.10. Maastrichtian

This period is marked by well-bedded white limestones, approximately 60 meters thick,overlain by a strong accumulation of gray to black clayey marls (150 meters).These latter have some limestone intercalations at their base (Dubourdieu, 1956). The sedimentary regime of the Upper Campanian and Maastrichtian periods still indicates the persistence of the same depositional conditions in a shallow environment and a warm sea,with a predominance of chalky limestones, rarely reef-forming (Chevenine et al., 1989).

III.2.1.11. Paleocene

Its base is made up of marls similar to those of the Upper Maastrichtian, intercalated with phosphate layers towards the upper levels.

III.2.1.12. Eocene

The Lower and Middle Eocene is characterized by flint-bearing limestones and others containing Nummulites, particularly near the borders of the Tebessa region. Their thickness is approximately200 meters.

III.2.1.13. Miocene

The Lower and Middle Miocene deposits lie transgressively on older formations (Albian-Senonian and even Triassic). They consist of a significant accumulation of marls and sandstones, reaching a thickness of 1,000 meters in the Ouled Soukiès basin (northwest of Ouenza) (Dubourdieu, 1956; Kowalski and Hamimed, 2000). At their base are conglomerates containing various limestone elements, gray flint, ferruginous pebbles, and elements borrowed from the Triassic, indicating diapiric activity (Bouzenoune, 1993). The flint reworking, considered to be of Ypresian age to the base of the Miocene, indicates the existence of an Eocene Sea where marine sedimentation was deposited during the Eocene and early to middle Miocene. Sedimentation at the end of the Miocene marks the beginning of a regression phase. The average thickness of the Miocene in the study area is approximately 150 meters (Dubourdieu, 1956).

III.2.1.14. Quaternary

Quaternary deposits are of continental origin and are found in the lower parts of reliefs as well as over vast areas such as present-day plains and valleys. They are composed of limestone crusts, scree silts, pebbles, and puddingstones. The thickness of the Quaternary varies between 10 and 30 meters (Dubourdieu, 1956). The lithological formations that can be distinguished are current formations, ancient formations, and the PlioQuaternary(Cheikhne Cheikh El Mehdi, 2024).

III.3. Climatology of Tebessa

Tebessa’s region allows to distinguish that the climate is typically semi-arid (cold winter and hot summer). The average annual temperature (16.14 ° C) and annual average rainfall over aperiod of 25 years (1990-2014) is estimated at 296.46 mm.From the plain of

Hamamet to the plateau of Tazbent the passage is made by Oued Bouakkous whose major bed is dug between very steep slopes.The plateau of Tazbent has a flat topography, with wadis little printed, bumped of some hills with rounded forms. The altitudes range from 1230m to 1470m. The edge of the plateau at the northern borderextends from 1230m to 1320m, which corresponds to the average altitude of the surface (Legrioui et al, 2017).

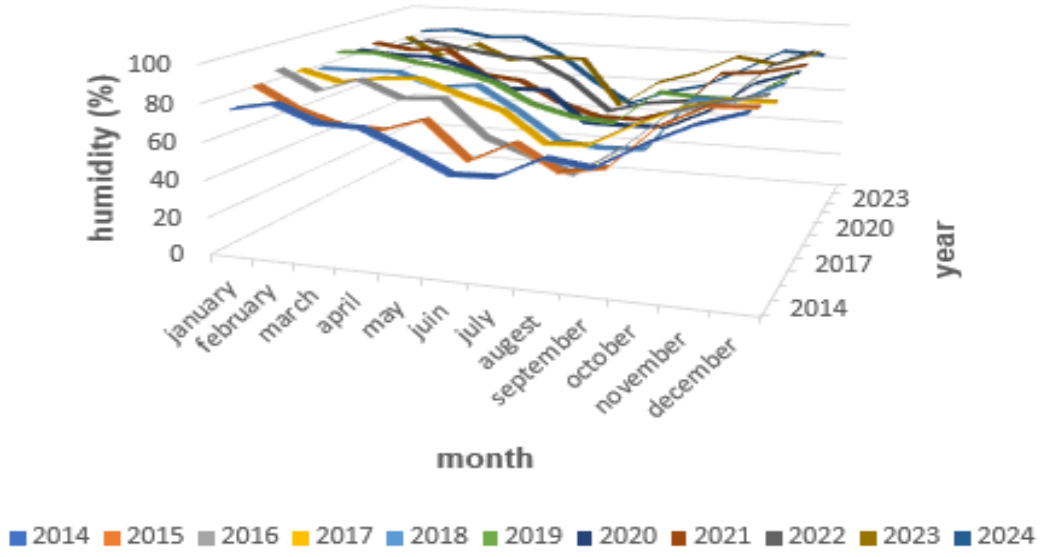

Figure 11: Distribution of the average annual temperature in the city of Tebessa.

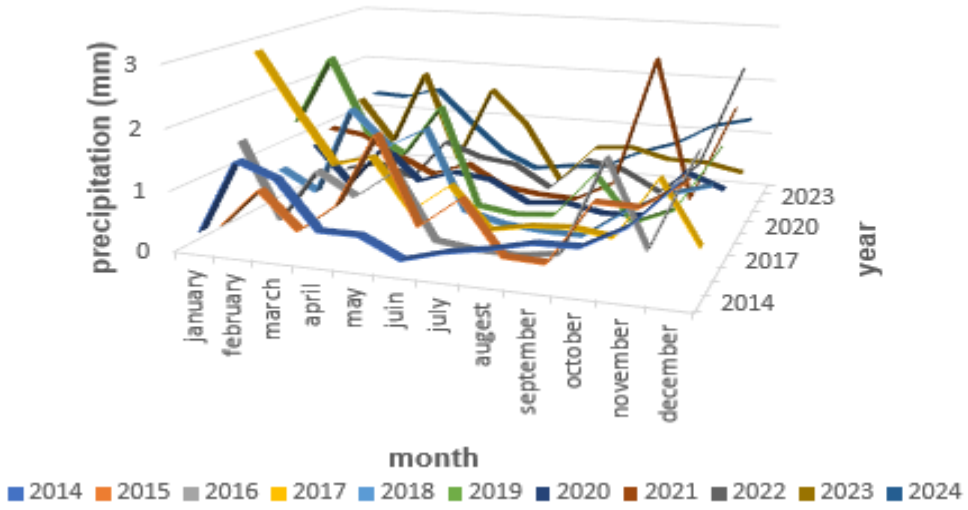

Figure 12: Distribution of annual precipitation in the city of Tebessa.

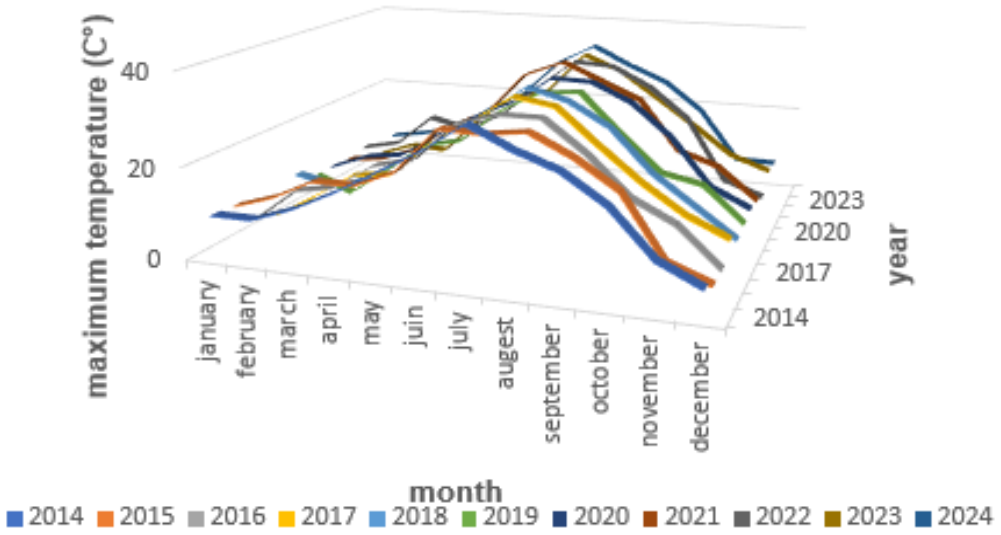

Figure 13: Distribution of average annual humidity in the city of Tebessa.

III.4. Hydrology of Tebessa

A network of watercourses, whether natural or artificial, permanent or temporary, that serve to drain runoff or groundwater discharge, either through springs or along wadi beds. This network is influenced by four main factors: geology (including susceptibility to erosion and the presence of structures that affect the flow direction), climate (for example, a dense network in mountainous and humid regions), and terrain topography (which can determine whether the network is in an erosive or sedimentary phase), and human impact (such as modifications to the original network route through drainage, embankments, dam construction, etc.). It is often characterized by its hierarchy, longitudinal profile, and development, which includes the number of watercourses and their lengths.

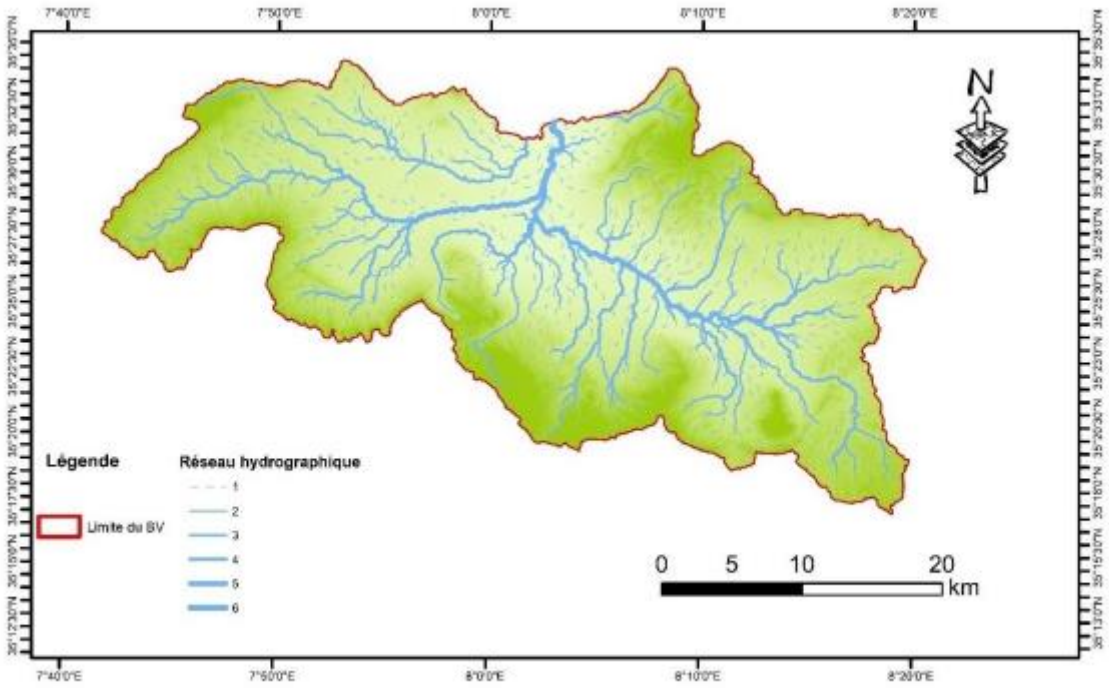

Figure 14: Hydrographic network of the Tebessa basin (Cheikhne Cheikh El Mehdi, 2024).

Transposing the above-mentioned information to the basin under study reveals that, despite its location in a semi-arid region, the number of rivers is limited. However, the network remains dense, characterized by numerous secondary tributaries, meandering routes, intermittence, and moderate slopes. The main rivers that collect runoff and direct it to the outlet are:

- **El Kebir valley:** originating in the east and following an east-west direction for approximately 32.22 km to its confluence with the Chabro valley. It is fed by the tributaries Oued Djebissa and Oued Hemadja, and receives the Raffana Wadi just before meeting the Chabro valley.
- **Serdies valley:** whose source is in the west (Dj. Serdies) and which runs for approximately 24.92 km before receiving the Boudiss valley and running along the western part of the plain.

It collects runoff from the west and southwest of the area.

- **Chabro valley:** formed by the confluence of these rivers and becoming the main river of the basin, located approximately in the center of the plain. The flow of this

interconnection system flows primarily toward the center from the east, south, and west, and then toward the north (Cheikhne Cheikh El Mehdi, 2024).

III.5. Soil classification according to GTR 2023 (NF EN 16907-2)

the GTR road paving guide was first published in 1992, revised in 2000, this guide updated in 2023 includes the European classification and defines the conditions of use, of processing and compaction of materials in fill and form layers.it is used to classify and test samples for their physico-mechanical and chemical properties which are: particle fraction lower than 0.063 mm, water content, dry density, Atterberg limits, methylene blue values, carbonate content and swelling pressure.

Table 1: Statistical data of the physico-mechanical and chemical properties of the tested samples.

Proprieties	Number	Maximum	Minimum	Medium
Particles<0,080	120	83	21	77;55
Water content (%)	120	24,746	14,467	19,668
Dry unit weight γ_d (KN/m^3)	120	2,21	1,39	1,695
Liquid limit (%)	120	77	22	55,125
Plastic limit (%)	120	51	5	24,575
Plasticity Index (%)	120	41	10	30,051
Methylene blue values (cm^3/g)	120	7,340	2,126	4,241
Carbonate content (%)	120	76,433	15,952	45,910
Swelling pressure (KN/m^2)	120	438,447	8,841	160,020

III.5.1. According to the plasticity index PI (NF P 94-051)

The plasticity index (PI) is related to the amount of clay in the soil and is a good indicator of compressibility. Its interpretation is all the more reliable as the proportion by weight of the fraction 0/400 µm (test fraction) contained in the soil studied is important, and the clay content of this fraction is high, PI values range from 10 to 41%which indicates that these soils are medium plastic silty clay and high plastic clay.

N.B: All the maps are generated by SURFER program version 17.

Table 2: Plasticity index classification (Roy and Bhalla, 2017).

Plasticity index (%)	Soil type	Degree of plasticity
0	Sand	Non-plastic
< 7	Silt	Low plastic
7-17	Silty clay	Medium plastic
> 17	Clay	High plastic

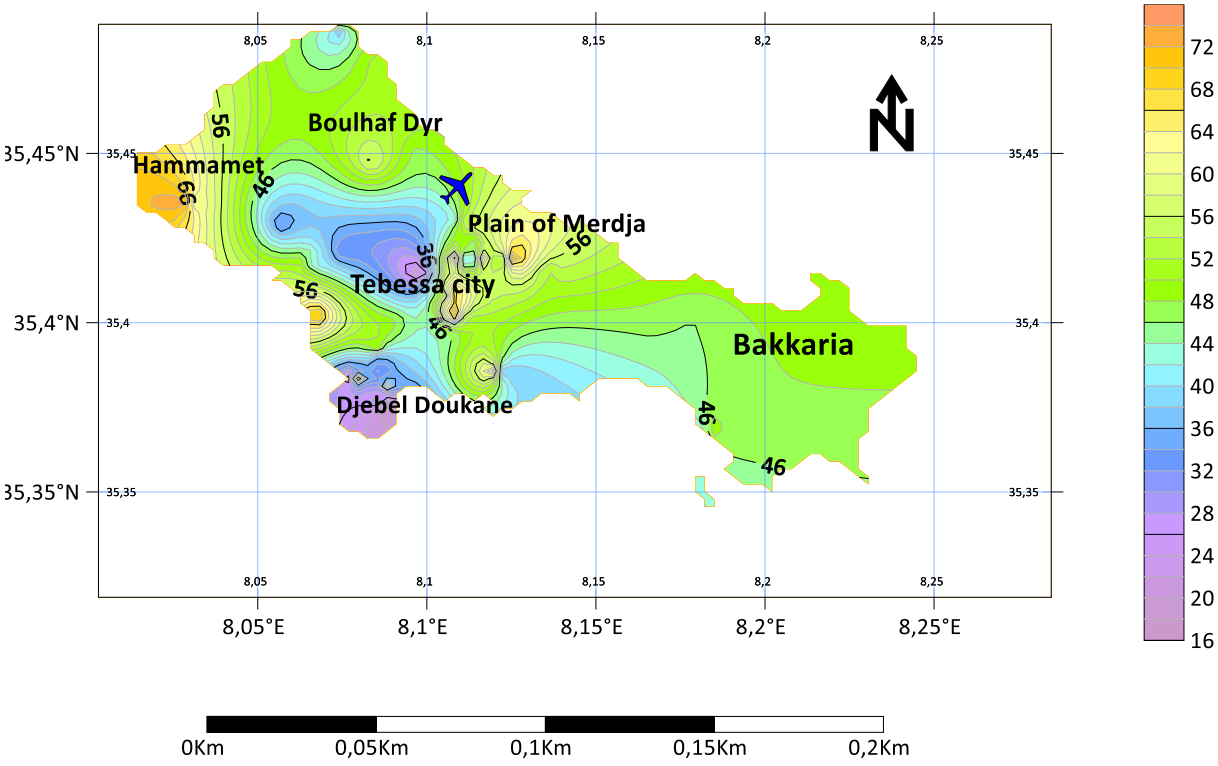

Figure 15: Distribution of the plasticity index in the supporting soil.

III.5.2. According to methylene blue values VBS (NF P 94-068)

Methylene blue (VBS) is an essential indicator for assessing soil quality, measuring the adsorption of methylene blue by clay particles. This test reveals the adsorption capacity of materials, the cleanliness and water sensitivity of a soil, as well as its ability to retain water, the VBS values range from 2,126 to 7,340 cm^3/g, it means that the soils are low-plastic sandy clay and high-plastic clays.

Table 3: VBS classification.

Methylene blue values	Soil categories
VBS< 0,1	Water-insensitive soil
0,2≤ VBS< 1,5	Sandy, silty soil, sensitive to water
1,5≤ VBS < 2,5	Sandy clay soil, low-plastics
2,5≤ VBS < 6	Medium-plasticity loamy soil
6≤ VBS < 8	Clay soil, high plastic
VBS > 8	Too much clay soil

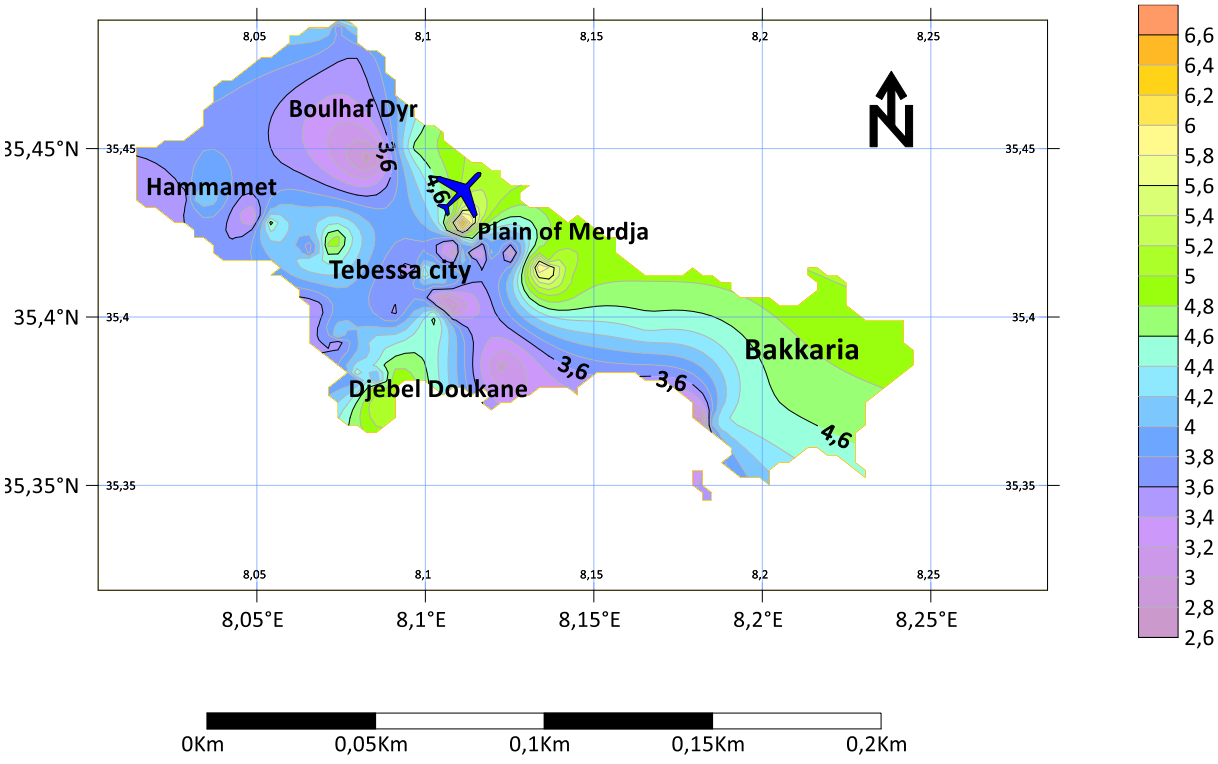

Figure 16: Distribution of VBS in the supporting soil.

III.5.3. According to carbonate content

The method consists in adding hydrochloric acid to a sample of soil to break down all carbonates present. The volume of carbon dioxide emitted is measured at Using a Scheibler apparatus and is compared with the volume of carbon dioxide produced by carbonate of pure calcium.

Carbonate values range from 16 to 76 %, that means that the soils are marls and marly limestones.

. Table 4: Carbonates classification (ISO, 1995).

Carbonate values	Classification	Soil categories
< 10	Non lime stone	Clay
10 to 29	Low-limestone	Marl
30 to 69	Moderately limestone	Marl
70 to 89	strongly limestone	Marly lime stone
≥ 90	Limestone	Limestone

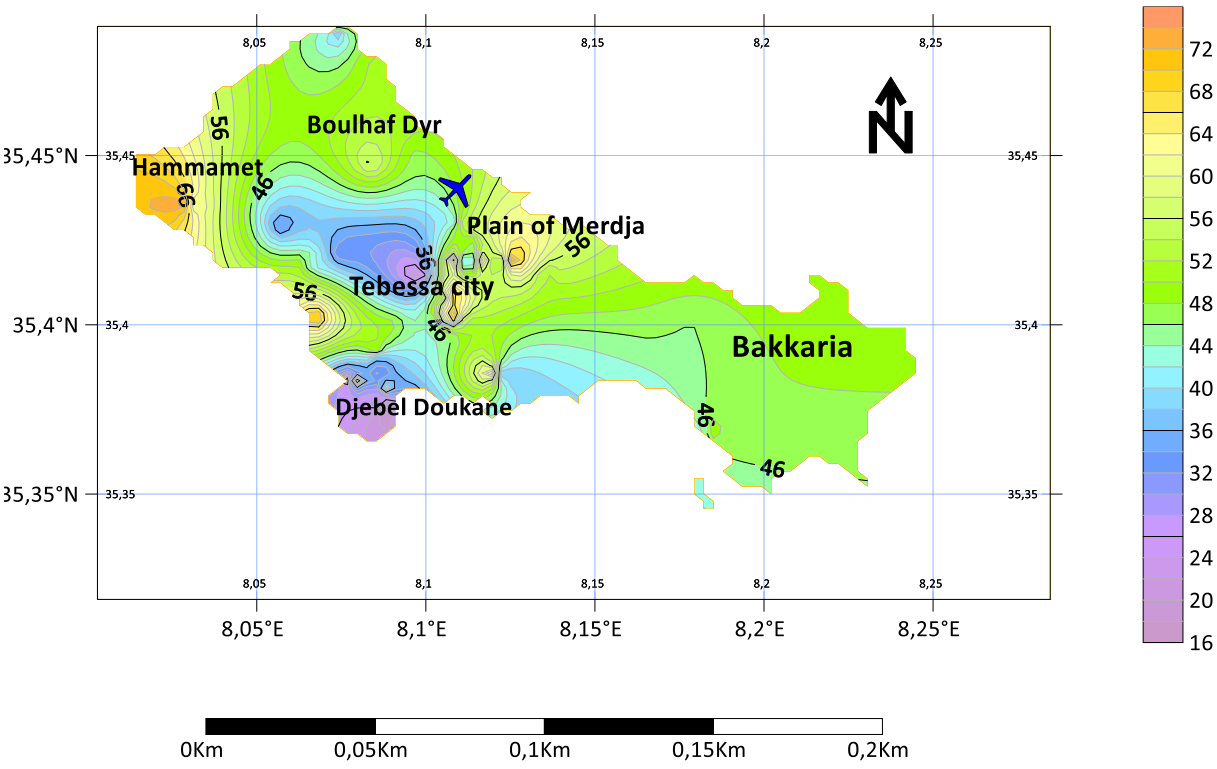

Figure 17: Distribution of carbonates content in the supporting soil.

III.5.4. According to water content

The water content of soil gives the basic idea of the moisture held by the soil which gives information about various engineering properties of the soil. The water content of soil is

defined as the ratio of the weight of water to the weight of soil solids present in the given soil mass. Its values range from 14 to 25% that means that the soils are slightly saturated.

Table 5: Soil states based on water content (Costet and Sanglerat, 1983).

Water content	Soil state
1-25	slightly saturated
25-50	Humid
50-75	Wet
75-99	Disabused
100	Saturated

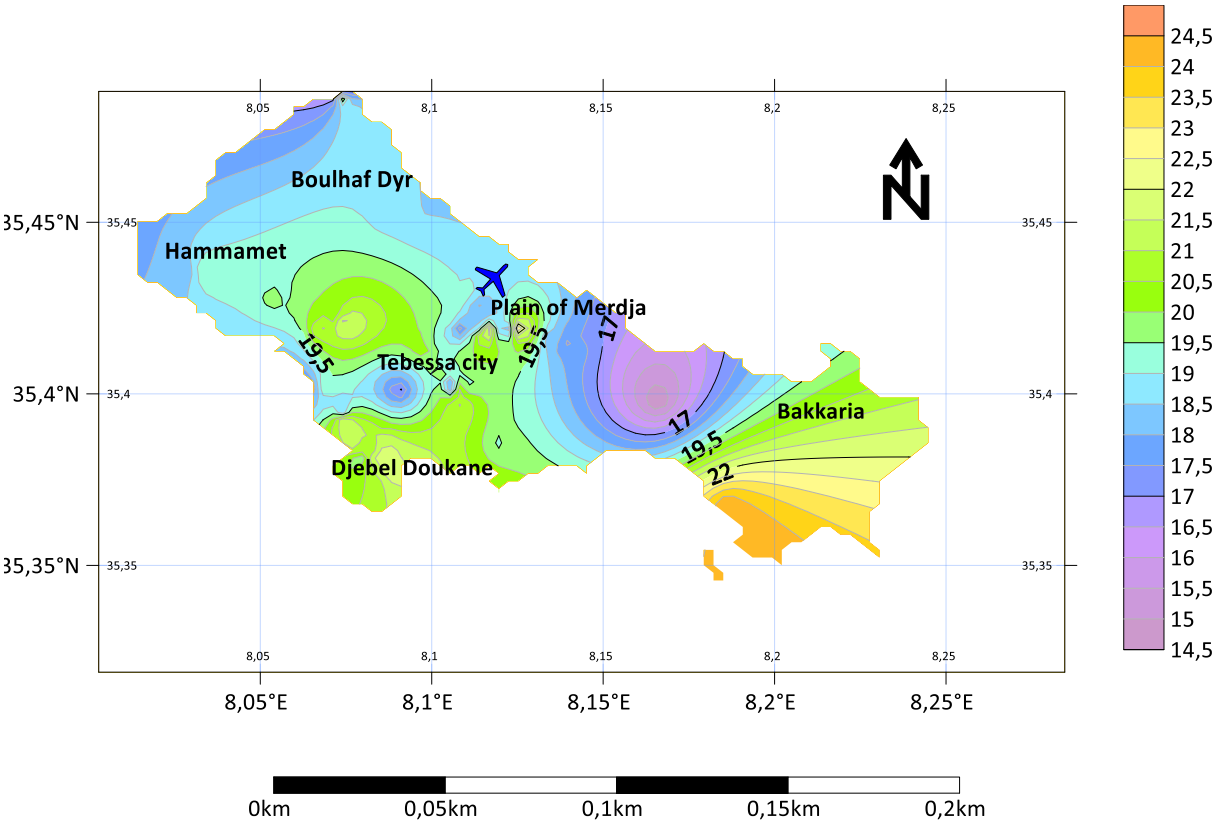

Figure 18: Distribution of water content in the supporting soil.

III.5.5. According to swelling pressure

The general concept of a swell test is to take an initial reservoir fluid and mix an injection fluid (typically a relevant injection gas) to estimate the miscibility effects of the given injection fluid with the reservoir fluid, The pressure varies from 9 to 438kPa that means that support soil has low to high swelling pressure.

Table 6: Direct measurement of swelling (Costet and Sanglerat, 1983).

Swelling pressure (Kpa)	Swelling categories
> 1000	Very high
250-500	High
150-250	Medium
50	Low

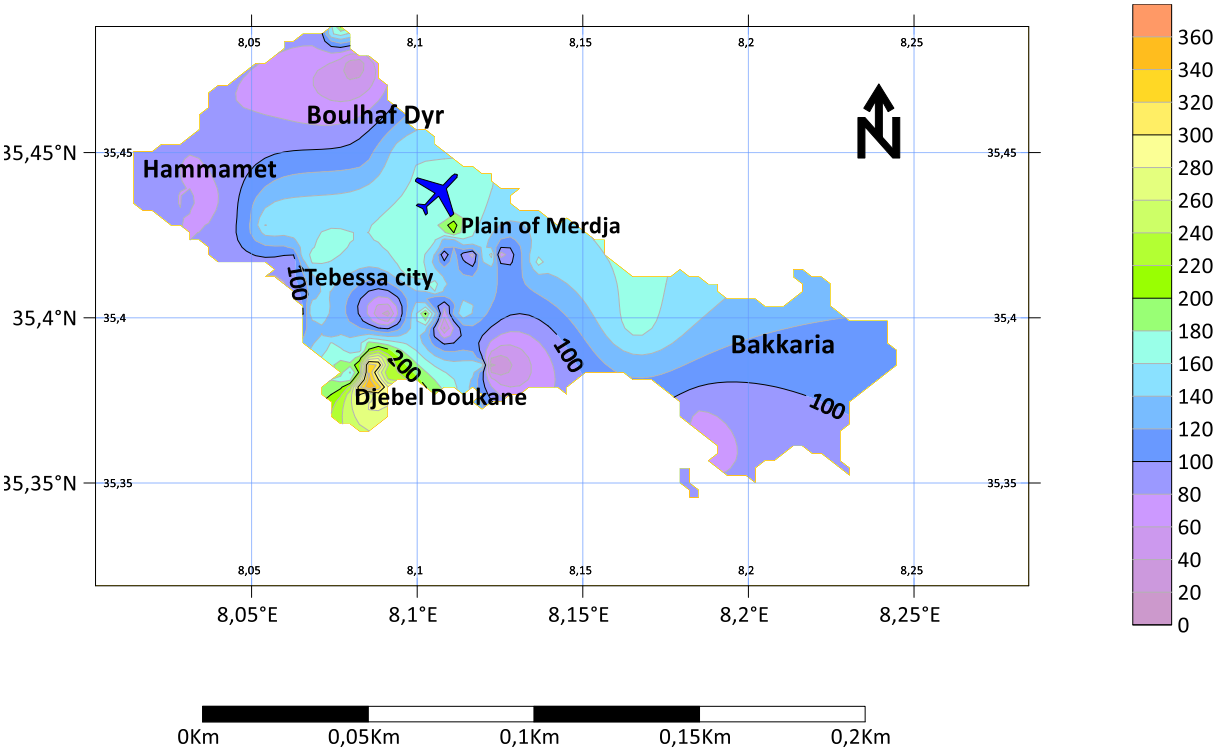

Figure 19: Distribution of swelling pressure in the supporting soil.

III.5.6. According to the total specific area

The total specific area is the surface of the solid grains on which the methylene blue binds, given by:

$$S_{st} = \left(\frac{VBS}{100}\right) * \left(\frac{N}{373}\right) * 130 * 10^{-20} \rightarrow S_{st} = 21VBS\ [m^2/g] \qquad (39)$$

Where:

$\frac{VBS}{100}$: blue fraction (< 2µ).

N: Avogadro number=6.023 10^{+23}

373: molecular weight of methylene blue in grams.

$\mathbf{130 * 10^{-20}}$**:** surface in cm^2 of a water molecule.

Total specific area values range from 44 to 156 m^2/g, according to the following table, the soils are composed of illites.

Table 7: Specific surface and CEC of some clay minerals (according to Morel, 1996.

Mineral	Internal surface (m^2/g)	Outer surface (m^2/g)	Total surface (m^2/g)	CEC (milliequivalent/100g)
Kaolinite	0	10-30	10-30	5-15
Illite	20-55	80-120	100-175	10-40
Montmrillonite	600-700	80	700-800	80-150
Vermiculite	700	40-70	760	100-150
Chlorite	-	100-175	100-175	10-40

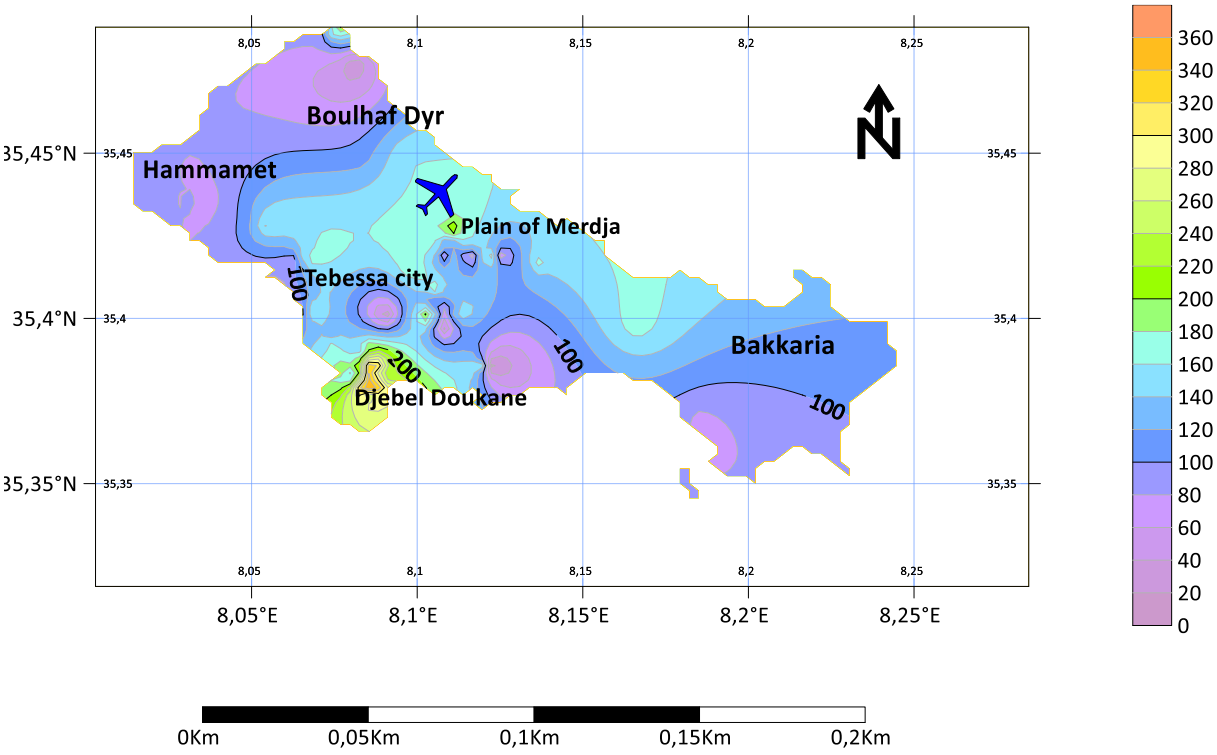

Figure 20: Distribution of the total specific area in the soil supporting.

III.5.7. According to CBR values

The Californian Bearing Ratio CBR test, developed in the 1930s in California, has become a global standard for assessing the mechanical strength of soils and granular materials. The

CBR test measures the bearing capacity of a soil by comparing its resistance to penetration with that of a reference material. The result, expressed as a percentage, indicates the quality of the tested soil compared to a standard material. The CBR values range from 2 to 40, which means that the field is composed of clays, hard clays and silt.

Table 8: Usual values soil / CBR.

Soil categories	CBR values
soft to very soft soil	< 2
Clays	2-10
Hard clays and silt	8-40
Sand	8-30
Gravels	15-80
Crushed	80-100

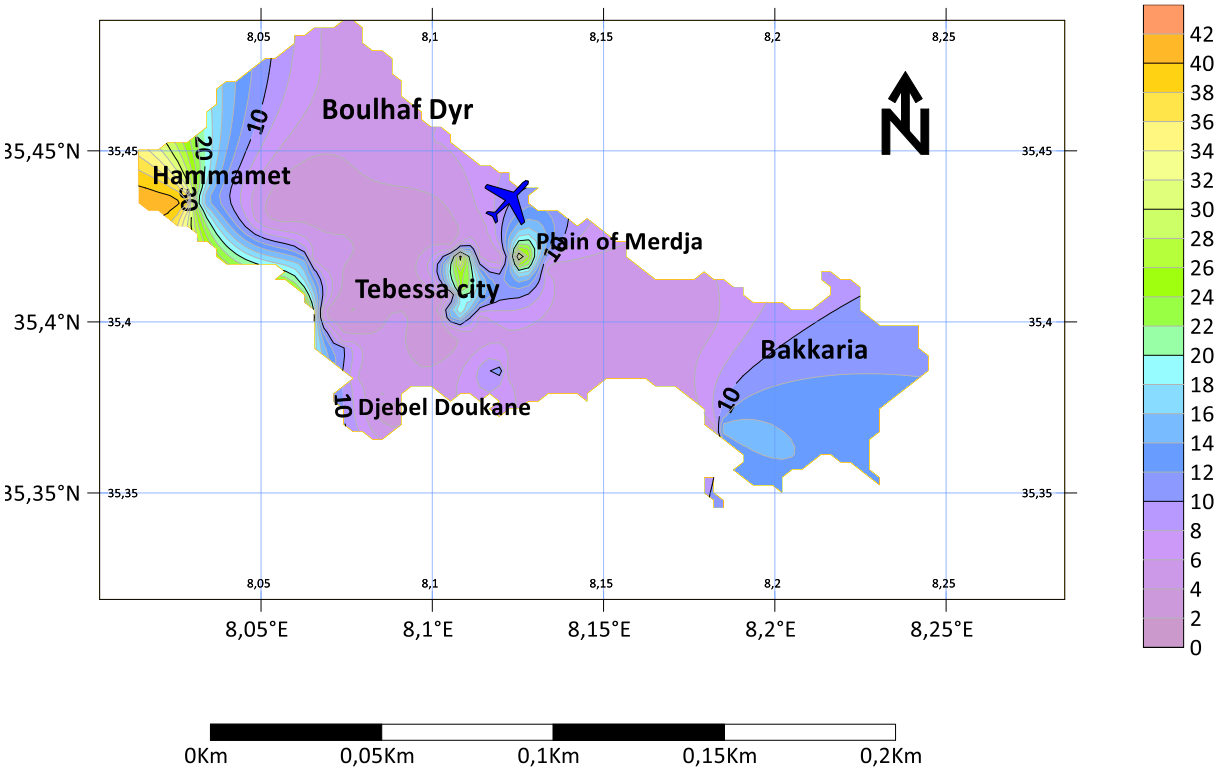

Figure 21: Distribution of CBR values in the supporting soil.

III.5.8. According to Casagrande plasticity chart

The liquid limit LL and the plasticity index PI are used not only to determine the plastic nature of soils, but also to give an idea about the potential for swelling, through the Casagrande chart, that means that the field is composed of high plasticity clay and high plasticity organic silts and soils with a little of medium plasticity clays and low plasticity clay.

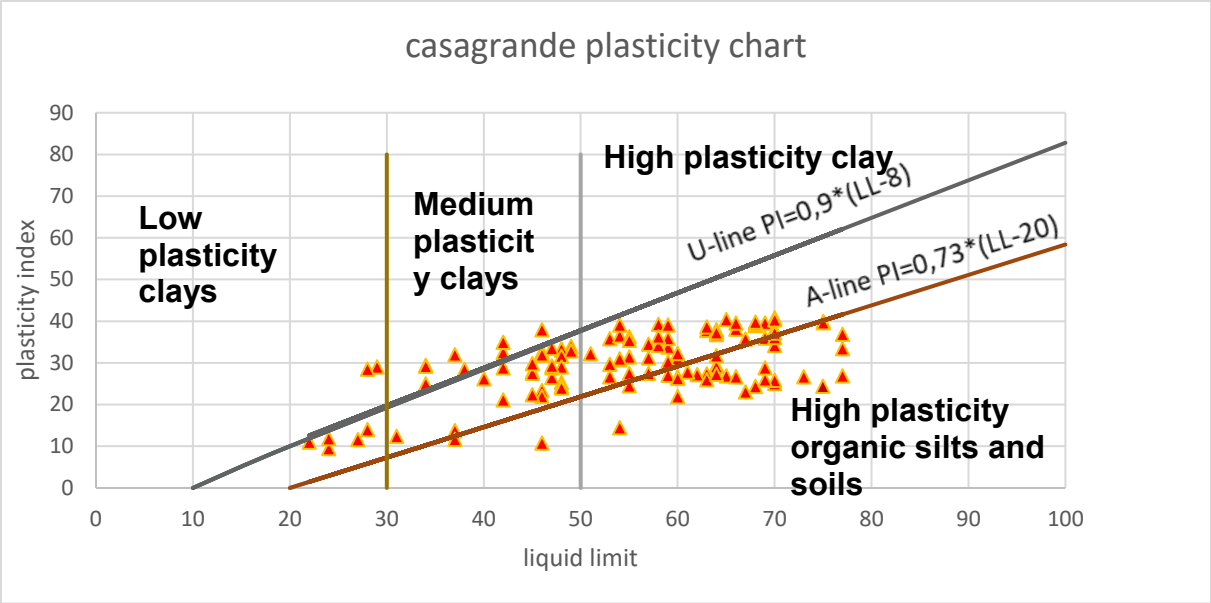

Figure 22: Classification of the supporting soil according to Casagrande chart (1948).

III.5.9. According to Dakshanamurthy and Raman classification

Dakshanamurthy and Raman (1973) were inspired by the plasticity diagram of Casagrande (1948) to provide a classification of the swelling potential. The graphical examination of the point cloud shows that soils have intermediate to high swelling potential.

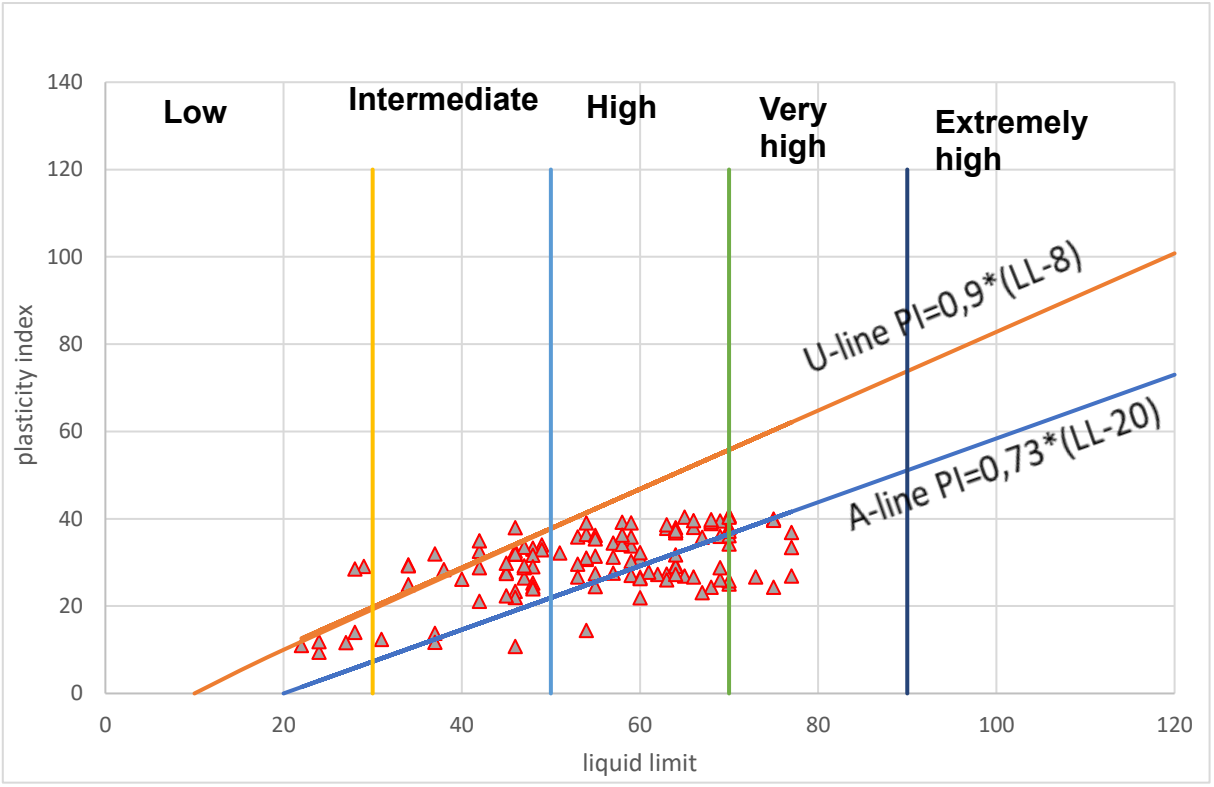

Figure 23: Classification of the support soil for the study area based on Dakshanamurthy and Raman (1973).

III.5.10. According to granularity (NF P 94-056)

The results of the granulometric analyses show that 100% of the soils analyzed have elements of dimensions less than 63 m. According to the GTR (version2023) method, these soils are classified in F1, F2, F3, F4 and I2, which means that the field is composed of limes little plastics, loess,alluvial silts,fine sands littlepolluted and arenaslittle plastic with Fine clay sands,silt, clayand marl littleplastics and arenasvery plastic, clay and claymowers, silts very plastic, sandand gravelvery silty.

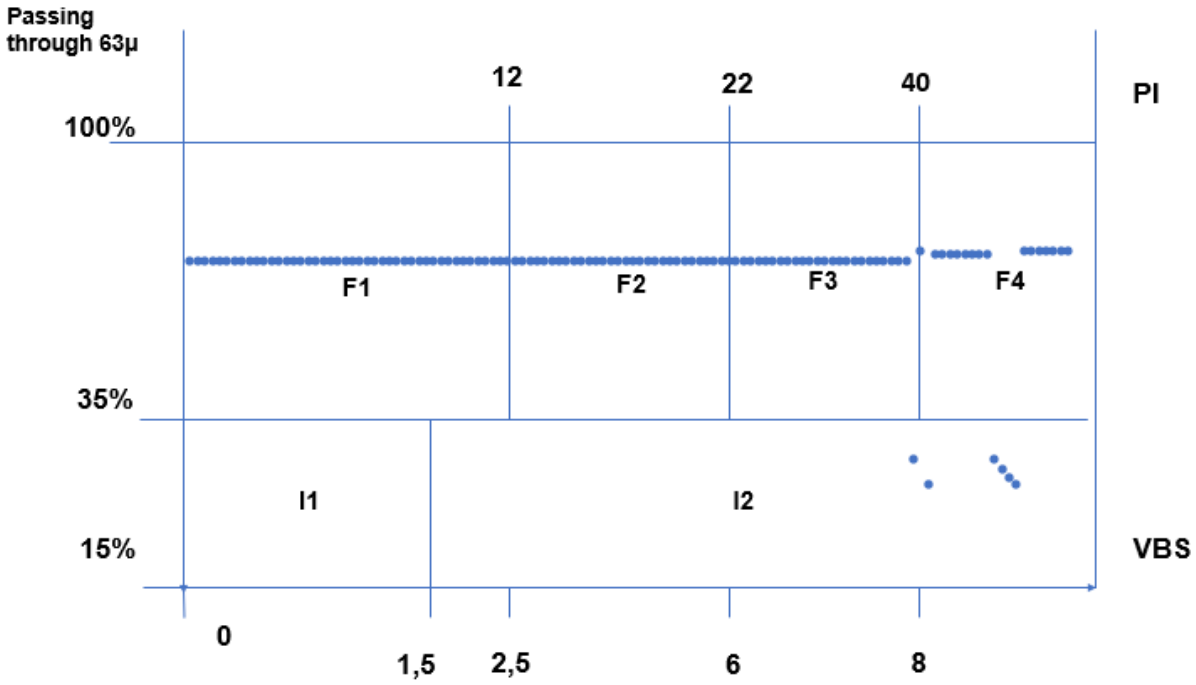

Figure 24: Classification of materials according to their nature (GTR, 2023).

III.6. Conclusion

- Tebessa is a semi-arid region in eastern Algeria, characterized by a complex geological and hydrogeological structure.
- The region experiences cold winters and hot summers with an average annual temperature of 16.14 °C and precipitation of 296.46 mm.
- Despite its semi-arid nature, the Tebessa basin has a dense river network with notable valleys like El Kebir, Serdies, and Chabro guiding runoff toward the center.
- Based on the geotechnical analyses and classifications presented in Chapter III, the soils in Tebessa are mainly fine-grained, plastic, and moderately to highly calcareous. They are dominated by silts, marls, clays, and fine sands, with variable swelling and bearing capacities, and are typical of semi-arid alluvial and sedimentary plains.

Chapter IV: Development of CBR Formula by Statistical Model and ANN Method

Chapter IV: Development CBR Formula by Statistical Model and ANN Method

IV.1. Introduction

The California Bearing Ratio (CBR) is a crucial parameter used to evaluate the strength of subgrade soils in road and pavement design. CBR is influenced by multiple geotechnical properties such as moisture content, plasticity index, compaction characteristics, and grain size distribution. Analysing and modelling the relationship between these variables and CBR can be challenging due to data complexity and inter-correlation among variables.

To address this, Principal Component Analysis (PCA) is used to reduce the number of input variables by transforming them into a smaller set of uncorrelated components that still capture most of the data's variance. These components represent the dominant patterns and factors affecting soil behaviour. After dimensionality reduction, regression analysis is applied to model the relationship between the principal components and the CBR value. This combined approach, often called Principal Component Regression (PCR), improves model accuracy and stability by eliminating multicollinearity and simplifying the data structure. Using PCA and regression together provides a powerful framework for predicting CBR from geotechnical parameters in a more efficient, interpretable, and reliable way especially when working with complex or limited datasets.

IV.2. CBR test

The California Bearing Ratio (CBR) is a penetration test used to evaluate the load-bearing capacity of soil and base materials for the design of pavements and other transportation infrastructure. It is defined as the ratio of the force per unit area required to penetrate a soil sample with a standard plunger to that required for a well-graded crushed stone, expressed as a percentage.

Mathematically, it is given by:

$$\boldsymbol{CBR}\ (\%) = \left(\frac{\boldsymbol{measured\ pressure\ on\ soil\ (at\ a\ given\ penetration)}}{\boldsymbol{standard\ pressure\ on\ crushed\ stone}}\right) \times \mathbf{100} \quad (\mathbf{40})$$

CBR testing is typically conducted at penetration depths of 2.5 mm and 5.0 mm, with the 2.5 mm value usually taken as the representative CBR unless the 5.0 mm value is higher. The resulting CBR value serves as an important input in determining the thickness and structural design of pavement layers.

California Bearing Ratio (CBR) is usually obtained through laboratory testing However, due to the time and resources required for such tests, researchers have developed predictive models using statistical and machine learning approaches. Regression analysis and artificial neural networks (ANNs) are commonly used to estimate CBR values based on easily measurable soil properties.

IV.3. Correlation analysis

Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. Specifically, in terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1.

IV.3.1. Interpretation of PCA results

IV.3.1.1. Interpretation of eigenvalues

The table and corresponding graph are linked to a mathematical object, the eigenvalues, which reflect the quality of the data projection which reflect the quality of the data projection. Each eigenvalue corresponds to a factor. The eigenvalues and corresponding factors are ranked in descending order in terms of the variability they represent.

In this study, we can see that the first eigenvalue represents 90,8% of the total variability of the total variability. This means that if we plot the data on a single axis, we will still be able to see 90,6% of the total variability in the data with two axes, we will see 94,5% of the total variability, which is very good. The fewer the variables and the more correlations there are between variables, the higher the representation of the first two axes.

In this study, it is not necessary to take into account the 3rd axis which represents only 2,7% of the total variability.

Table 9: Representation of eigenvalues according to PCA.

	F1	F2	F3
Valeur propre	8498,730	338,257	253,000
Variabilité (%)	90,876	3,617	2,705
% cumulé	90,876	94,493	97,198

IV.3.1.2. Interpretation of correlation cercle

The correlation circle is a visualization displaying how much the original variables are correlated with the first two principal components according to the figure γ_d and CA are positively correlated with CBR, finers and IP are negatively correlated with CBR.

So, the most effected parameters to CBR values are γ_d, CA, passing and IP.

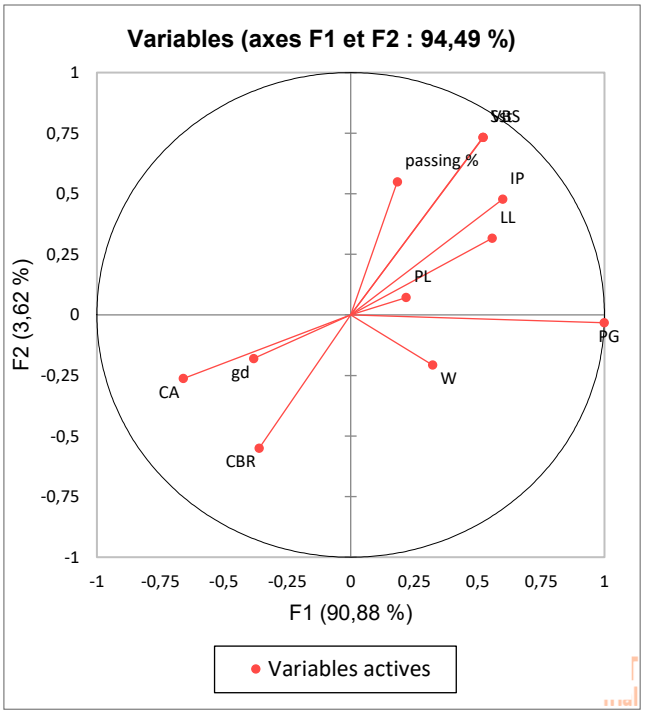

Figure 25: Correlation circle.

IV.3.1.3. Interpretation of correlation matrix

To understand the relationship between soil parameters, the correlation matrix was used to study the dependence of several geotechnical parameters at the same time. The table spreads out the correlation coefficients between each variable in the form of a matrix. The last row of the correlation matrix describes the level of regression between CBR and other geotechnical parameters mentioned above. This shows that CA and $\gamma_d$are strongly positively correlated with coefficients of 0,57 and 0,30 respectively. In addition, finers and IP are strongly negatively correlated to CBR with a correlation coefficient of -0,82 and -0,64 respectively.

	Sst	*PG*	*VBS*	*CA*	*IP*	*W*	*γd*	*Passing %*	*LL*	*PL*	*CBR*
Sst	1										
PG	0,4993984	1									
VBS	0,9999368	0,5012703	1								
CA	-0,351552	-0,6440613	-0,3541111	1							
IP	0,5240834	0,5814723	0,5266762	-0,631538	1						
W	-0,1118952	0,3263638	-0,1110587	-0,3832404	0,200136	1					
γd	-0,3051779	-0,3764872	-0,3068027	0,2574337	-0,3355267	-0,0690058	1				
Passing	0,2129301	0,1661547	0,2137416	-0,370316	0,5484961	0,1108904	-0,1953867	1			
LL	0,3318836	0,5409574	0,333502	-0,605107	0,5719637	0,211727	-0,3552812	0,358461	1		
PL	0,0278805	0,2100292	0,0278285	-0,3864564	-0,0190972	0,1876148	-0,1209038	0,1179084	0,5437208	1	
CBR	-0,3233565	-0,3413236	-0,3250475	0,5760344	-0,6491017	-0,1232969	0,3074942	-0,8260858	-0,5088918	-0,2132541	1

Sst total specific surface (m^2), Ps swelling pressure (KN/m^2), VBS methylene blue value, CA carbonates content (%), PI plasticity index (%), W water content (%), γ_d dry unit weight (Mg/m^3), fraction of particles smaller than 0,080mm (%), LL liquid limit (%), PL plastic limit (%), CBR California bearing ratio (%)

Figure 26: Correlation matrix.

IV.4. Statistical Analysis

IV.4.1 Multiple Linear Regression Analysis (MLR)

Multivariate linear regression was used to study how different input variables affect the CBR value. At first, all the input variables were included in the model. Then, one variable was removed at a time to test different combinations. This helped find the best set of variables that gave the most accurate and reliable results. The independent variables were blue methylene (VBS), swelling pressure (SP), specific surface total (SST), carbonates content (CA), plasticity index (PI), water content (W), dry unit weight (γ_d), finers (%), liquid limit (LL) and plastic limit (PL) and CBR was taken as dependent variable.

It has been observed from the MLR analysis that the most important factors affecting the CBR are finers (%), carbonates content (CA), plasticity index (PI) and dry unit weight (γ_d).

The predictive model for CBR containing the minimum variables and giving significant value of coefficient of determination derived by MLR analysis is given below:

$$\boldsymbol{CBR = 39,9026 - 0,433512 finers + 0,135481 CA - 0,253576 \frac{PI}{\gamma_d}} + \mathcal{E} \qquad \textbf{(41)}$$

Where$\mathcal{E}$ the mean-zero Gaussian random error term,γ_d is in KN/m^{3}and all other parameters are in %.

The coefficient of determination (R^2) for the equation is 0,7805, and the corresponding adjusted value is 0,7748. The adjusted coefficient of determination, denoted as R^2(adj), represents the adjustment for the degree of freedom in the estimating equation and avoids the upward bias in the unadjustedR^2, when the number of samples is relatively smaller than that of explanatory variables in the model.

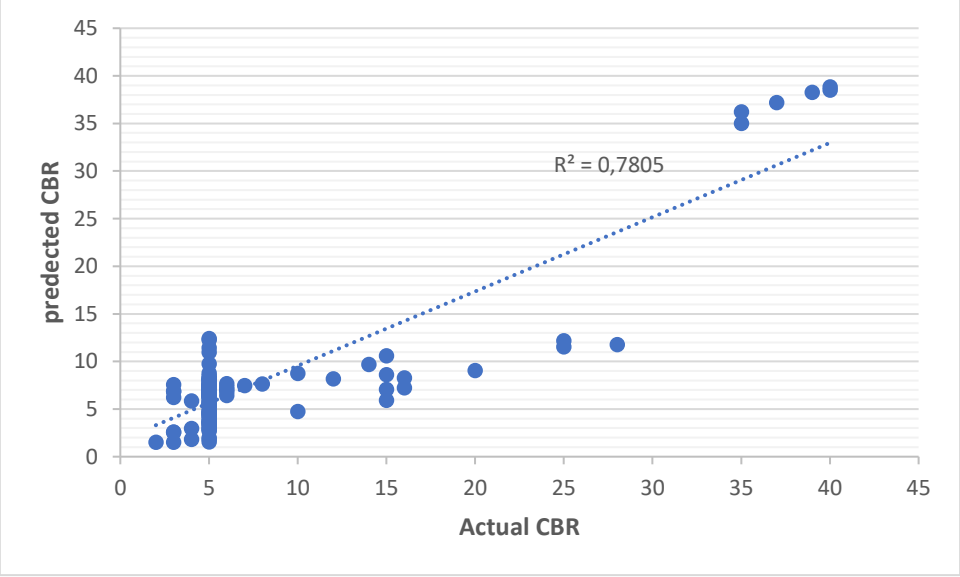

Figure 27: Measured versus predicted values of CBR.

The T statistics and the corresponding *p* values for the equation are shown in Table 10. Each variable in the equation is significant as long as its *p* value is less than 0.05.

Table 10: Results of regression analysis relevant to Eq. (41) significant at 5% level.

Terms	coefficient	T statics	P values
Constant	39,9026	11,6752	0,000
Passing	-0,4335	-12,9547	0,000
CA	0,1355	4,3117	0,000
IP/γ_d	-0,2536	-2,6726	0,009

IV.4.2. Artificial Neural Network analysis (ANN)

Artificial Neural Networks, or ANNs for short, are a type of artificial intelligence that learn by looking at data—kind of like how people learn from experience. Instead of being told exactly what to do, they figure out how things are related by finding patterns in the data they're given. Once they've learned these patterns, they can make smart guesses or predictions about new situations. This makes neural networks useful for solving problems where it's hard to write exact rules.

IV.4.2.1. ANN results

IV.4.2.1.1. CBR metrics

The figure bellow presents performance metrics for two models used in predicting the California Bearing Ratio (CBR): a Neural Network (Enhanced) model and a Linear Formula model. The Neural Network (Enhanced) model demonstrates superior performance across all evaluation metrics. It achieves a Mean Absolute Error (MAE) of 0.70, indicating that its predictions deviate from actual values by an average of 0.70 units. Its Mean Squared Error (MSE) is 2.02, and the Root Mean Squared Error (RMSE) is 1.42, both reflecting relatively low prediction errors. Most notably, it has a coefficient of determination (R^2) of 0.893, suggesting that the model explains approximately 89.3% of the variability in the CBR data.

In contrast, the Linear Formula model performs less favorably, with a higher MAE of 1.53, MSE of 3.06, and RMSE of 1.75. Its R^2 value is 0.857, which, while still strong, indicates a slightly lower ability to explain the variance in the data compared to the neural network. Overall, these results clearly indicate that the Enhanced Neural Network model provides more accurate and reliable CBR predictions than the traditional Linear Formula.

IV.4.2.1.1.1. Metrics elements

a) Mean Absolute Error (MAE)

Mean Absolute Error, or MAE, is an evaluation metric used to evaluate the performance of the regression model. It measures the average of the errors' magnitude between the predicted and actual values.

The mathematical formula for Mean Absolute Error is:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^{'}\right|} \qquad \textbf{(42)}$$

Where:

$\boldsymbol{n}$**:** number of observations.

$\boldsymbol{y_i}$**:** the actual value of i^{th} observation.

$\boldsymbol{y_i^{'}}$**:** the predicted value of the i^{th} observation.

b) Mean Squared Error (MSE)

Mean Squared Error, as the name suggests, calculates the average of the squares of the errors or residuals. In the context of machine learning, it quantifies the average squared difference between the actual values and the values predicted by the model.

Mathematically, MSE is defined as:

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^{'}\right|^2} \qquad \textbf{(43)}$$

Where:

$\boldsymbol{n}$**:** number of observations.

$\boldsymbol{y_i}$**:** the actual value of i^{th} observation.

$\boldsymbol{y_i^{'}}$**:** the predicted value of the i^{th} observation.

c) Root Mean Square Error (RMSE)

RMSE represents the square root of the average squared differences between predicted and observed outcomes. It is a metric predominantly utilized in regression analysis and forecasting, where accuracy matters significantly. The lower the RMSE, the better the model's ability to predict accurately. Conversely, a higher RMSE signifies a greater discrepancy between the predicted and actual outcomes.

Mathematically, RMSE is defined as:

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2}{n}}} \qquad \textbf{(44)}$$

d) R -squared (coefficient of determination)

R-squared, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by one or more independent variables in a regression model. In simpler terms, it shows how well the data fit a regression line or curve.

Mathematically, R-squared is defined as:

$$\boldsymbol{R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2}{\sum_{i=1}^{n}(y_i - \mu)^2}} \qquad \textbf{(45)}$$

where:

RSS: sum of squares of residuals.

TSS: total sum of squares.

IV.4.2.1.2. Loss vs Epochs

The graph shows how the training and validation loss change over 350 epochs. At the start, both losses are high, meaning the model wasn't performing well. As training continues, the training loss keeps going down steadily, which means the model is learning from the training data. However, the validation loss goes down at first but then stays around the same level and even jumps up and down a bit. This shows the model is not improving much on unseen data after a certain point.

In simple terms, the model is learning the training data well, but it's not getting better at handling new data — a sign of overfitting.

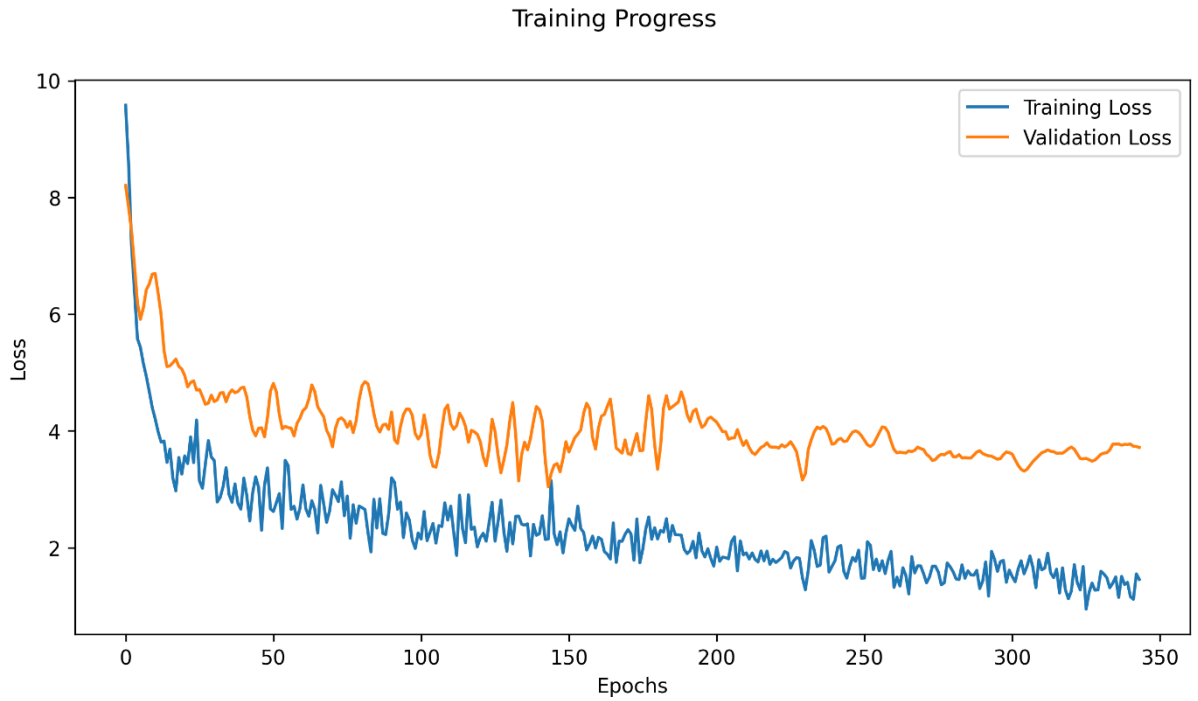

Figure 28: Loss vs Epochs.

IV.4.2.1.3. Neural Network vs Linear Formula Comparaison

The bar chart compares how well a Neural Network and a Linear Formula perform using four metrics: MAE, MSE, RMSE, and R^2. In all four, the Neural Network does better. It has lower error values (MAE, MSE, RMSE), which means its predictions are more accurate and closer to the actual results. It also has a slightly higher R^2 value, showing it fits the data a little better. Overall, the Neural Network gives better and more reliable results than the Linear Formula.

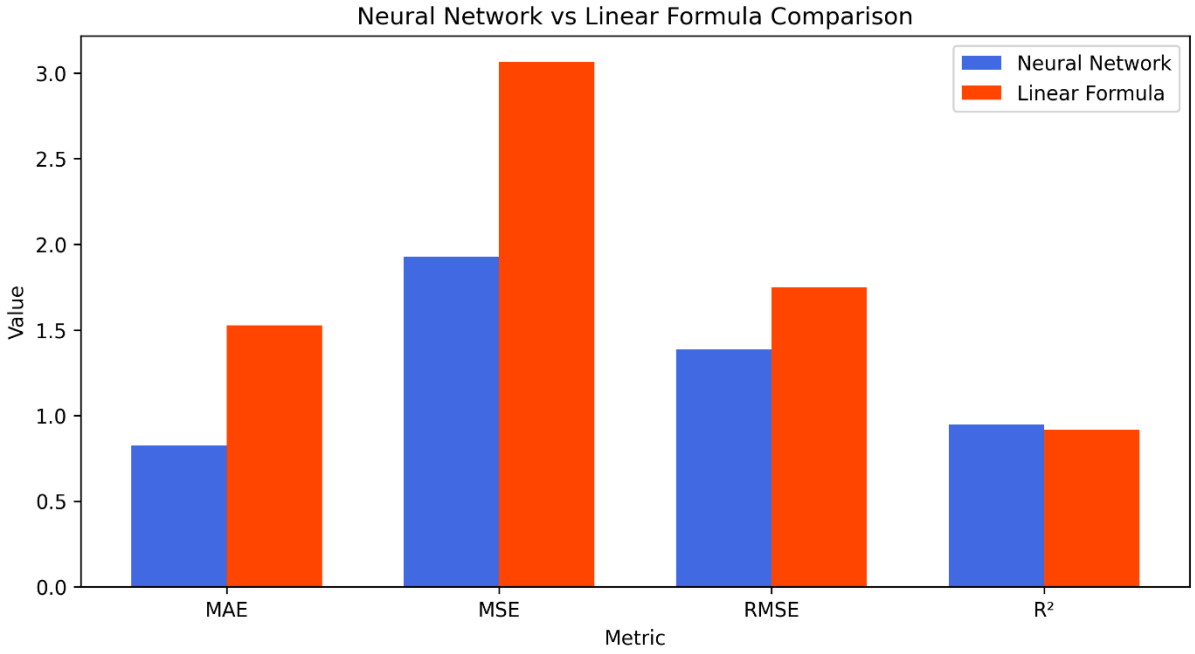

Figure 29: Neural Network vs Linear Formula Comparison.

IV.4.2.4. Actual vs Predicted CBR Comparaison

The scatter plot titled *"Actual vs Predicted CBR Comparison"* compares the prediction accuracy of a Neural Network and a Linear Formula in estimating California Bearing Ratio (CBR) values. Each point on the plot represents a prediction, with blue circles indicating the Neural Network predictions and orange representing the Linear Formula predictions. Ideally, accurate predictions should lie close to the imaginary diagonal line where predicted values equal actual values. In this plot, the Neural Network predictions are more closely aligned with this diagonal trend, showing that they are generally more accurate and consistent. On the other hand, the Linear Formula predictions are more scattered and tend to deviate further from the actual values. This visual comparison clearly indicates that the Neural Network provides more reliable and precise CBR predictions than the Linear Formula.

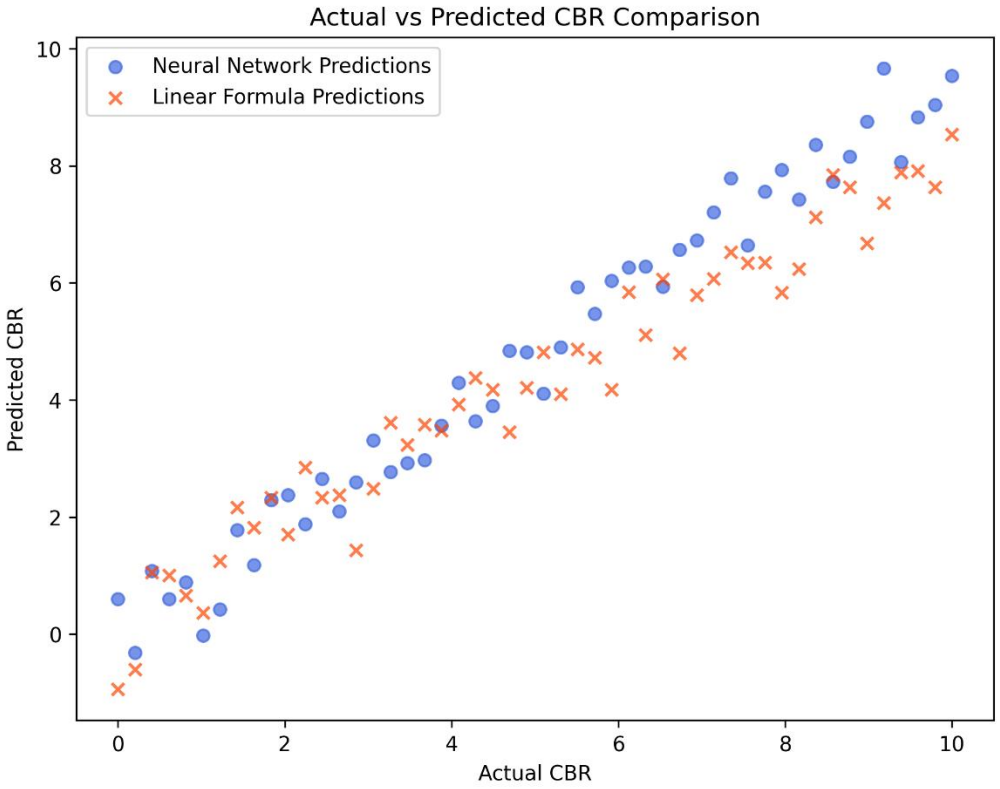

Figure30: Actual vs Predicted CBR Comparaison.

IV.4.2.5. Prediction Comparison

This chart compares how well two methods predict California Bearing Ratio (CBR) values. The blue dots show predictions made by a Neural Network, and the orange X's show predictions from a Linear Formula. The horizontal axis shows the actual CBR values, while the vertical axis shows the predicted CBR values.

Most of the actual CBR values are small (around 2 to 6), and both methods give fairly similar predictions in this range. However, there is one large actual CBR value (about 35), and here both methods predicted higher values than the actual. The Neural Network predicted a much higher value than the Linear Formula. In summary, both methods work okay for small CBR values but are less accurate for larger ones.

IV.5. Conclusion

This chapter focused on predicting the California Bearing Ratio (CBR) — a key soil strength parameter used in road and pavement design — using statistical methods and Artificial Neural Networks (ANN).

- The first principal component (PC1) alone explains 90.88% of data variation.
- Most influential factors on CBR: positive: Carbonate content (CA), Dry unit weight (γd) and negative: % Passing (fine particles), Plasticity Index (PI).
- Multiple Linear Regression (MLR) (R^2 = 0.7805, Adjusted R^2 = 0.7748 and All predictors are statistically significant ($p < 0.05$)).

- Artificial Neural Network (ANN) (Superior prediction performance, R^2 = 0.893 (explains 89.3% of CBR variability) and MAE = 0.70, RMSE = 1.42.
- ANN predictions closely match actual CBR values across all ranges.
- ANN outperformed MLR in all metrics.
- Recommended for more accurate and reliable CBR prediction, especially when dealing with complex soil behavior.

General Conclusion

This study combined detailed geotechnical investigation and predictive modeling to evaluate and estimate the California Bearing Ratio (CBR) in the Tebessa region. Tebessa is a semi-arid region in eastern Algeria, characterized by a complex geological and hydrogeological structure.

The region experiences cold winters and hot summers with an average annual temperature of 16.14 °C and precipitation of 296.46 mm. Despite its semi-arid nature, the Tebessa basin has a dense river network with notable valleys like El Kebir, Serdies, and Chabro guiding runoff toward the center.

Based on the geotechnical analyses and classifications presented in Chapter III, the soils in Tebessa are mainly fine-grained, plastic, and moderately to highly calcareous. They are dominated by silts, marls, clays, and fine sands, with variable swelling and bearing capacities, and are typical of semi-arid alluvial and sedimentary plains.

The predictive modeling phase showed that Principal Component Analysis (PCA) effectively reduced data dimensionality and highlighted the most influential parameters affecting CBR: carbonate content, dry unit weight, finers, and plasticity index. A Multiple Linear Regression (MLR) model was developed with a determination coefficient R^2=0.7805, but the Artificial Neural Network (ANN) model achieved significantly higher accuracy with R^2=0.893, confirming its superiority in predicting CBR from soil properties.

References

Subhabaha. P (2023), history of statistics, linked in, https://www.linkedin.com/pulse/history-statistics-dr-subhabaha-pal, last seen 02/20th/2025 at 15:54.

Chappelow. J (2024), statistics: definition, types, and importance, investopedia, https://www.investopedia.com/terms/s/statistics.asp, last seen 02/20/2025 at 16:31.

Frost. J (2018), mean, median, and mode: measures of central tendency, statistics by Jim, https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode, last seen 02/20/2025 at 19:24.

Hargrave. M (2024), standard deviation formula and uses vs. Variance, investopedia, https://www.investopedia.com/terms/s/standarddeviation.asp, last seen 02/21/2025 at 1:45.

Hayes. A, what is variance in statistics? Definition, formula, and example, investopedia, https://www.investopedia.com/terms/v/variance.asp, last seen 02/21/2025 at 15:35.

Greenacre. M, j. F. Groenen. P, Hastie. T, Iodice D'enza. A, Markos. A and Tuzhilina. E (2022), principal component analysis, research gate, pp 1-24,doi: 10.1038/s43586-022-00184-w .

Jaadi. Z (2024), principal component analysis (pca): a step-by-step explanation, built in, https://builtin.com/data-science/step-step-explanation-principal-component-analysis, last seen 02/21/2025 at 15:49.

Singh. H (2025), principal component analysis, link springer,https://link.springer.com/article/10.1007/s10661-022-10555-1, last seen 02/23/2025 at 9:32.

Beers. B (2024), regression: definition, analysis, calculation, and example, investopedia, https://www.investopedia.com/terms/r/regression.asp, last seen 02/23/2025 at 9:39.

Bevans. R (2020), simple linear regression | an easy introduction & examples, scribbr, https://www.scribbr.com/statistics/simple-linear-regression, last seen 02/23/2025 at 10:06.

Hayes. A (2024), multiple linear regression (mlr) definition, formula, and example, investopedia, https://www.investopedia.com/terms/m/mlr.asp, last seen 02/23/2025 at 10:22.

Kumar. A (2024), understanding statistical analysis: techniques and applications, simplilearnhttps://www.simplilearn.com/what-is-statistical-analysis-article, last seen 02/23/2025 at 17:37.

Jacquez. G (2024), what is geostatistics? Biomedware. Https://biomedware.com/what-is-geostatistics, last seen 02/23/2025 at 17:58.

Goovaerts. P (2023); kriging vs co-kriging: understanding two data interpolation methods, biomedware,https://biomedware.com/kriging-vs-cokriging-understanding-two-data-interpolation-methods, last seen 02/24/2025 at 12:57.

Wackernagel. H (2003), multivariate geostatistics.

Paláncz. B, l. Awange. J, völgyesi. L and h. Lewis. R (2023), mathematical geosciences: hybrid symbolic-numeric methods, research gate, doi: 10.1007/978-3-030-92495-9.

Jiang. Y, li. X, luo. H, yin. S and kaynak. O (2022), discover artificial intelligence. Google scholar,https://scholar.google.com/scholar?Q=jiang.+y,+li.+x,+luo.+h,+yin.+s+and++kaynak.+o+(2022),+discover+artificial+intelligence.&hl=fr&as_sdt=0&as_vis=1&oi=scholart, last seen 02/24/2025 at 13:08.

Glover. E (2024), what is artificial intelligence (ai)? Alpinum consulting https://alpinumconsulting.com/blogs/fpga-front-runner/april-2024-using-ai-in-fpga-development-and-products/renishaw-an-overview-of-products-and-an-insight-into-the-application-of-ai-technology-and challenges/#:~:text=glover%2c%20(2024)%20highlights%20that,resources%20for%20higher%2dpriority%20activities, last seen 02/26/2025 at 17:31.

The investopedia team (2025), what is artificial intelligence (ai)? Investopedia, https://www.investopedia.com/terms/a/artificial-intelligence-ai.asp, last seen 03/02/2025 at 8:45.

Coursera staff (2024), what is artificial intelligence? Definition, uses, and types, https://www.coursera.org/articles/what-is-artificial-intelligence, last seen 03/02/2025 at 9:02.

Crabtree. M (2024), what is machine learning? Definition, types, tools & more, datacamp, https://www.datacamp.com/blog/what-is-machine-learning, last seen 03/02/2025 at 10:20.

Lev. C (2024),what is machine learning? Guide, definition and examples, tech target, https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ml, last seen 03/02/2025 at 11:25.

Brown. S (2021), machine learning, explained, mitsloan, https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained, last seen 03/02/2025 at 21:45.

Chen. J (2024), what is a neural network? Investopedia, https://www.investopedia.com/terms/n/neuralnetwork.asp, last seen 03/02/2025 at 22:11.

B.j. Copeland, (2025), artificial intelligence, Britannica. Https://www.britannica.com/technology/artificial-intelligence, last seen 03/02/2025 at 22:39.

Oppermann. A (2023), artificial intelligence vs. Machine learning vs. Deep learning: what's the difference? Built in, https://builtin.com/artificial-intelligence/ai-vs-machine-learning, last seen 03/04/2025 at 14:56.

La'aro Bolaji. A, Tajuddin Khader. A, Azmi al-Betar. M, Awadallah. M (2013), artificial bee colony algorithm, its variants and applications: a survey, journal of theoretical and applied information technology,jatit&lls, pp434-459.

Balwant. K, Dharmender. K (2013), a review on artificial bee colony algorithm, international journal of engineering and technology, ijet, pp175-186.

Sharkawy. A-n(2020), principle of neural network and its main types: review, ournal of advances in applied & computational mathematics, pp 8-19, doi: 10.15377/2409 5761.2020.07.2

Oliver. M, webster. R, basic steps in geostatistics: the variogram and kriging, springer briefs, pp 15-37, doi 10.1007/978-3-319-15865-5_2.

Djalali. A, sarker. D, benghazi. Z, rais. K (2022), geospatial-based approach for susceptibility assessment of expansive soils using a new multicriteria classification model, arabian journal of geosciences, pp 1-13, doi.org/10.1007/s12517-022-11024-2.
Gringarten. E, Deutsch. C, variogram interpretation and modeling, international association for mathematical geology, pp 507-533, doi 0882-8121/01/0500-0507$19.50/1

Rashad. M. Z, El tahlawi. M. R, Ahmed. S, Anwer. A (2007), election of variogram model for elshagara gold mine, eastern desert, Egypt, pp 33-44.

Sedrati. N, Djabri.l (2014), contribution of hydrochemistry to the characterization and assessment of groundwater resources: the case of Tebessa alluvial aquifer (Algeria), iahs press, pp 458-463, doi: 10.5194/piahs-364-458-2014.

Achou. A-a (2024), mémoire variation et monotonie de la resestivite électrique des aquifères de la région de Tébessa : etat passif et actif, pp4.

Legrioui. R, Baali. F, Amor. H, Abdeslam. I, Mouici. R (2017), water quality at a karstic aquifer in the region of Tebessa, northeast -Algeria, science direct, Elsevier ltd, pp 356-366, doi:10.1016/j.egypro.2017.07.119.

Cheikhne cheikh el Mehdi. S-m (2024), mémoire application des méthodes statistiques multivariées pour l'évaluation de la qualité des eaux souterraines de la nappe alluviale Tébessa-Bekkaria-el hammamet, pp17-25.

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي و البحث العلمي
جامعة العربي التبسي- تبسة

مقرر رقم: 199 مؤرخ في: 04 جوان 2025

يتضمن الترخيص بمناقشة مذكرة الماستر

إنّ مدير جامعة العربي التبسي بتبسة،

- بموجب القرار الوزاري رقم 318 والمؤرخ في 05ماي 2021المتضمن تعيين السيد "قواسمية عبد الكريم" مديرا لجامعة العربي التبسي - تبسة،

-و بمقتضى المرسوم التنفيذي رقم : 12- 363 مؤرخ في 8 أكتوبر 2012، يعدل و يتمم المرسوم التنفيذي رقم 09 - 08 المؤرخ في : 04 جانفي 2009 و المتضمن إنشاء جامعةالعربي التبّسي بتبسة،

-وبمقتضى المرسوم التنفيذي رقم 08-265 المؤرّخ في 17 شعبان عام 1429 الموافق 19 غشت سنة 2008 الذي يحدّد نظام الدراسات للحصول على شهادة الليسانس وشهادة الماستروشهادة الدّكتوراه، لاسيما المادة 9 منه،

-وبموجب القرار رقم 362 المؤرّخ في 09 جوان 2014 الذي يحدّد كيفيات إعداد ومناقشة مذكّرة الماستر، لاسيما المادة 7 منه،

-- وبموجب القرار رقم 375 المؤرّخ في 15 جوان 2020 المعدل لملحق القرار1080 المؤرّخ في 13 أكتوبر 2015 والمتضمّن تأهيل ماستر الفروع ذات تسجيل وطني بجامعة تبسة، اختصاص جيوتقني

-وبموجب المقرّر رقم 199 المؤرّخ في2025/06/04 والمتضمّن تعيين لجنة مناقشة مذكّرة الماستر،

-وبعد الاطّلاع على مقرر تعيين لجنة مناقشة مذكرة الماستر المؤرّخ في 2025/06/04

يقرّر ما يأتي:

المادة1:يُرخّصُ للطالب عزري حنين،المولودة بتاريخ18 ماي 2002بمرسط بمناقشة مذكّرة الماستر والموسُومة بـ

Indirect estimation of california bearing ratio with statistical and machine learning approach

المادة 2: يكلّف رئيس قسم المناجم و الجيوتكنولوجيابتنفيذ هذا المقرّر الّذي يسلّم نسخة عنه إلى الطّالب المعني بالمناقشة وأعضاء لجنة المناقشة فور توقيعه، وبضمان نشره عبر فضاءات المؤسّسة المادية والرقمية.

المادة 3: تُحفظ نسخة عن هذا المقرّر ضمن الملفّ البيداغوجي للطّالب المعني وينشر في النّشرة الرّسمية لجامعة العربي التبسي.

حُرّر ب تبسة، في: 04 جوان 2025

عن المدير، وبتفويض منه
مدير المعهد

مدير معهد المناجم
د.عوسي روبير

جامعة الشهيد الشيخ العربي التبسي
معهد المناجم

Scanned with CamScanner

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة العربي التبسي- تبسة

مقرر رقم : 200 مؤرخ في : 04 جوان 2025

يتضمن تعيين لجنة مناقشة مذكرة الماستر

إنّ مدير جامعة العربي التبسي بتبسة،

- بموجب القرار الوزاري رقم 318 والمؤرخ في 05ماي 2021المتضمن تعيين السيد "قواسمية عبد الكريم" مديرا لجامعة العربي التبسي - تبسة،

-و بمقتضى المرسوم التنفيذي رقم : 12- 363 مؤرخ في 8 أكتوبر 2012، يعدل و يتمم المرسوم التنفيذي رقم 09 - 08 المؤرخ في : 04 جانفي 2009 و المتضمن إنشاء جامعةالعربي التبّسي بتبسة،

-وبمقتضى المرسوم التنفيذي رقم 08-265 المؤرّخ في 17 شعبان عام 1429 الموافق 19 غشت سنة 2008 الذي يحدّد نظام الدراسات للحصول على شهادة الليسانس وشهادة الماستروشهادة الدّكتوراه، لاسيما المادة 9 منه،

-وبموجب القرار رقم 362 المؤرّخ في 09 جوان 2014 الذي يحدّد كيفيات إعداد ومناقشة مذكّرة الماستر، لاسيما المادتان 10 و11 منه،

-وبموجب القرار رقم 1380 المؤرّخ في 09 أوت 2016 والمتضمّن موائمة التكوينات في الماستر، اختصاص جيوتقني،

-وبعد الاطّلاع على محضر المجلس العلميلمعهد المناجمرقم 03 المؤرّخ في 12 ديسمبر 2024،

يقرّر ما يأتي:

المادة 1:تُعيِّنُ بموجب هذا المقرّر لجنة مناقشة مذكّرة الماستر المحضَّرة من طرف الطّالبة:

عزري حنين،المولودة بتاريخ18 ماي 2002بـ مرسط – تبسة -

والموسُومة بـ: Indirect estimation of california bearing ratio with statistical and machine learning approach

والمسجّل (ة) بمعهدالمناجم

المادة 2: تتشكّل اللجنة المشار إليها في المادة الأولى من الأعضاء الآتي ذكرهم:

رقم	الاسم واللّقب	الرتبة	مؤسّسة الانتماء	الصّفة
1	براح ياسين	أستاذ محاضر–أ	جامعة العربي التبسي - تبسة	رئيسا
2	جلالي عادل	أستاذ	جامعة العربي التبسي - تبسة	مشرفا
3	حمدان عادل	أستاذ مساعد – أ	جامعة العربي التبسي - تبسة	مناقشا

المادة 3: يكلّف رئيس قسم المناجم والجيوتكنولوجيا بتنفيذ هذا المقرّر الّذي يُسلّم نسخةٌ عنه إلى كلّ من الطّالب المعني والمشرف على المذكّرة وأعضاء لجنة المناقشة فور توقيعه.

المادة 4: تحفظ نسخة عن هذا المقرّر في الملفّ البيداغوجي للطالب المعني، وينشر في النّشرة الرّسمية لجامعة العربي التبسي.

حُرّر ب تبسة، في: 04 جوان 2025

عن المدير ، وبتفويض منه

مديرالمعهد

مدير معهد المناجم

د.خولمي زوبير

جامعة الشهيد الشيخ العربي التبسي تبسة

Scanned with CamScanner