Oct 29th, 12:00 AM - Oct 30th, 12:00 AM

International Conference on Computer Science

University for Business and Technology - UBT

UBT

Leadership and Innovation

Education | Research | Training | Consulting | Certification

PROCEEDINGS

11th UBT ANNUAL INTERNATIONAL CONFERENCE

29-30

OCTOBER

UBT Innovation Campus

INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND COMMUNICATION ENGINEERING

Proceedings of the 11th Annual International Conference on Computer Science and Communication Engineering

Edited by
Edmond Hajrizi

Conference Book of Proceedings International Conference

Pristina,

ISBN 978-9951-550-78-9

International Conference on Business, Technology and Innovation Pristina, Kosovo 29-30

Editor: Edmond Hajrizi

Organizing Committee: Edmond Hajrizi, Hasan Metin, Visar Krelani, Hazir Cadraku, Retkoceri B, Selmani F, Muhamet Ahmeti, Selmani F, Muhamet Sherifi, Kastrati A, Mirlinda Reçica

Publisher, UBT

Editor Speech of IC - BTI

International Conference is the 11th international interdisciplinary peer reviewed conference which publishes works of the scientists as well as practitioners in the area where UBT is active in Education, Research and Development. The UBT aims to implement an integrated strategy to establish itself as an internationally competitive, research-intensive institution, committed to the transfer of knowledge and the provision of a world-class education to the most talented students from all backgrounds. It is delivering different courses in science, management and technology. This year we celebrate the 21th Years Anniversary. The main perspective of the conference is to connect scientists and practitioners from different disciplines in the same place and make them be aware of the recent advancements in different research fields, and provide them with a unique forum to share their experiences. It is also the place to support the new academic staff for doing research and publish their work in international standard level. This conference consists of sub conferences in different fields: - Management, Business and Economics - Humanities and Social Sciences (Law, Political Sciences, Media and Communications) - Computer Science and Information Systems - Mechatronics, Robotics, Energy and Systems Engineering - Architecture, Integrated Design, Spatial Planning, Civil Engineering and Infrastructure - Life Sciences and Technologies (Medicine, Nursing, Pharmaceutical Sciences, Phycology, Dentistry, and Food Science),- Art Disciplines (Integrated Design, Music, Fashion, and Art). This conference is the major scientific event of the UBT. It is organizing annually and always in cooperation with the partner universities from the region and Europe. In this case as partner universities are: University of Tirana – Faculty of Economics, University of Korca. As professional partners in this conference are: Kosova Association for Control, Automation and Systems Engineering (KA – CASE), Kosova Association for Modeling and Simulation (KA – SIM), Quality Kosova, Kosova Association for Management. This conference is sponsored by EUROSIM - The European Association of Simulation. We have to thank all Authors, partners, sponsors and also the conference organizing team making this event a real international scientific event. This year we have more application, participants and publication than last year.

Congratulation!

Edmond

Hajrizi, Rector of UBT and Chair of IC - BTI

Përmbajtja

Using matrices in cryptography

Faton Kabashi* Lamir Shkurti* Besnik Qehaja* Hizer Leka* Mirlinda Selimaj*

**** UBT, Higher Education Institution, Pristina, Kosova***

(e-mails: faton.kabashi@ubt-uni.net;lamir.shkurti@ubt-uni.net; besnik.qehaja@ubt-uni.net; hizer.leka@ubt-uni.net; mirlinda.slimaj@ubt-uni.net).

Abstract: Data encryption has become a necessity with the increase in sensitive data being stored and transmitted through computers. Cryptography is the science of encoding and decoding messages, which is commonly used in everyday life to store sensitive information such as credit card numbers. Cryptography has many purposes, the first being to ensure the confidentiality of communication, then the integrity and authenticity. Cryptology is the use of algorithms and codes to increase the security of data. The study of cryptography requires proficiency in various mathematical concepts such as algebraic theory, probability, statistics, discrete mathematics, algebraic geometry, complex analysis, number theory, algorithms, binary numbers, prime factorization, and other key areas. A mathematical discipline used for data encryption in Cryptography is Linear Algebra, especially matrices operations. In this paper, our aim is to explore the applications of matrices in the fields of cryptography. The purpose of this paper is to show that mathematical concepts, in this case "matrices" are playing a major role in computer graphics, computer science and robotics. This study would be important to other researchers because the research findings of this study will benefit researchers with literature review to expand their research in the application of matrices. It helps address and provide background information for researchers who would like to conduct further research in this area.

Keywords: Matrices, inverse matrices, cryptography, data encryption, data decryption, the Affine Cipher algorithm, the Hill Cipher.

I. INTRODUCTION

Matrices are a powerful mathematical tool that can be used in a variety of cryptographic applications. One common use of matrices in cryptography is in the encryption and decryption of messages using a matrix transformation. This technique is known as matrix encryption. In matrix encryption, a message is represented as a matrix, and then transformed using a key matrix [1]. The resulting matrix is then transmitted to the recipient, who can use the inverse of the key matrix to decrypt the message. Matrix encryption has several advantages over other encryption techniques. First, it is resistant to linear attacks, which means that an attacker cannot use linear algebra to break the encryption. Second, it allows for efficient encryption and decryption of large messages, which is important in many practical applications.

Another use of matrices in cryptography is in the creation of hash functions. A hash function is a mathematical function that takes an input (usually a message) and produces a fixed-length output, called a hash. Hash functions are used in a variety of cryptographic applications, such as verifying the integrity of data and ensuring the authenticity of messages [2].

Matrices can be used to create hash functions by representing the message as a matrix, and then applying a matrix transformation to the matrix. The resulting matrix is then reduced to a fixed-length hash using a variety of techniques, such as taking the determinant of the matrix.

In conclusion, matrices are a powerful mathematical tool that can be used in a variety of cryptographic applications, including encryption and hash functions. By using matrix transformations, it is possible to create secure and efficient cryptographic systems that can be used to protect sensitive data and ensure the authenticity of messages [3].

II. THEORETICAL BACKGROUND

1. Matrix Encryption:

Matrix encryption is a type of symmetric key encryption in which a message is transformed using a matrix. The most common type of matrix used in encryption is a square matrix with dimensions $n\ x\ n$, where n is the number of elements in the message. The key matrix used for encryption and decryption is also a square matrix with dimensions $n\ x\ n$ [4].

To encrypt a message using a key matrix, the message is first represented as a matrix with dimensions $n\ x\ n$. The key matrix is then multiplied with the message matrix using matrix multiplication. The resulting matrix is then transmitted to the recipient.

To decrypt the message, the recipient uses the inverse of the key matrix to recover the original message. The inverse of the key matrix is calculated using matrix inversion, which is a computationally intensive operation. However, once the inverse matrix is calculated, decryption can be performed efficiently using matrix multiplication.

Matrix encryption is resistant to linear attacks because the relationship between the input message and the encrypted message is non-linear. Linear attacks are a type of attack that uses linear algebra to find patterns in the encrypted message that can be used to recover the original message.

2. Hash Functions:

A hash function is a mathematical function that takes an input message of any length and produces a fixed-length output called a hash. Hash functions are used in many cryptographic applications, including digital signatures, message authentication codes, and password storage [5].

Matrices can be used to create hash functions by transforming the message matrix using a matrix transformation. The resulting matrix is then reduced to a fixed-length hash using a variety of techniques, such as taking the determinant of the matrix.

Matrix-based hash functions have several advantages over other hash functions. First, they are resistant to collision attacks, which are a type of attack that attempts to find two different input messages that produce the same hash. Second, they are computationally efficient, which makes them suitable for use in systems that require fast processing of large amounts of data.

3. Other Cryptographic Applications:

Matrices are also used in other cryptographic applications, such as digital signatures, key exchange, and error correction. For example, matrices can be used to create digital signatures by transforming a message matrix using a secret key matrix and then reducing the resulting matrix to a fixed-length hash. The hash is then signed using a private key and transmitted to the recipient.

In key exchange, matrices can be used to generate a shared secret key between two parties. The parties each generate a random matrix and then exchange the matrices. They then multiply their own matrix with the received matrix using matrix multiplication to generate a

shared secret key.

In error correction, matrices can be used to correct errors that occur during data transmission. The data is represented as a matrix, and then a parity check matrix is used to detect and correct errors in the data [6].

In conclusion, matrices are a versatile and powerful tool that can be used in many different cryptographic applications. They provide a way to transform and manipulate data in a way that is resistant to attacks and efficient to compute.

4. RELATED WORK

Through the Related work, we will review the existing studies related to on the use of matrices in cryptography. There are several applications of the use of matrices in cryptography, which we will present below.

Arpita Bose, et al. [7], in their survey paper provides an overview of various matrix-based cryptographic schemes, including matrix encryption, matrix-based hash functions, and other applications of matrices in cryptography. The authors discuss the advantages and disadvantages of using matrices in cryptography and provide a comparison of different matrix-based schemes. They also highlight some of the challenges and open problems in this área.

Xiaoyan Liu, et al. [7], in their paper proposes a new matrix-based encryption scheme based on nonassociative algebraic structures. The scheme uses a nonassociative algebraic structure called the octonion algebra to generate the key matrix, which is used to encrypt the plaintext matrix. The authors claim that this scheme is more secure than traditional matrix-based encryption schemes because it is resistant to certain types of attacks.

Jianhua Wang, et al. [7], in their paper proposes a matrixbased image encryption scheme that uses elliptic curve cryptography to generate the key matrix. The scheme is designed to be secure against various types of attacks, including brute-force attacks and differential attacks. The authors demonstrate the effectiveness of their scheme through simulations and comparisons with other matrix-based encryption schemes.

Muhammad Ilyas, et al. [7], in their paper proposes a matrix-based cryptographic scheme for secure communication in 5G wireless networks. The scheme uses matrix multiplication and modular arithmetic to encrypt and decrypt messages. The authors claim that their scheme provides high security and low computational complexity, making it suitable for use in resource-constrained wireless networks.

Baojiang Cui, et al. [7], in their paper proposes two efficient error correction schemes based on matrices in cryptography. The first scheme uses a matrix multiplication algorithm to correct errors in data transmission, while the second scheme uses a syndrome decoding algorithm based on the parity check matrix. The authors demonstrate the effectiveness of their schemes through simulations and comparisons with other error correction schemes.

Overall, these papers demonstrate the versatility and potential of using matrices in cryptography. Matrixbased schemes have been shown to provide high security and low computational complexity, making them suitable for use in a variety of applications. However, there are still open problems and challenges in this area that require further research and development.

5. CRYPTOGRAPHIC ALGORITHMS

The process of encrypting and decrypting data involves a mathematical function that utilizes keys, which can be in the form of a word, number, or phrase. The cryptographic algorithm utilizes one or more of these keys to encrypt the data, and the same piece of information can be encrypted using various keys to generate different ciphertext. The effectiveness of the encryption relies on both the cryptographic algorithm and the keys used in the encryption process [12]. In the current article, we will primarily rely on the Affine Cipher algorithm and the Hill Cipheras. The Affine Cipher algorithm and the Hill Cipher are two different cryptographic algorithms that use different techniques to encrypt plaintext messages [13]. The Affine Cipher algorithm is a type of substitution cipher that uses a combination of multiplication and addition to encrypting letters in a message, while the Hill Cipher uses matrix multiplication to encrypt the entire message at once. However, it is possible to combine these two algorithms by using the Affine Cipher with more than one key in the Hill Cipher.

A. USING THE AFFINE CIPHER ALGORITHM IN HILL CIPHER

Affine Cipher

The Affine Cipher is a type of monoalphabetic substitution cipher where each letter in the plaintext is mapped to its numeric equivalent, transformed by a mathematical function, and then mapped back to a letter to generate the ciphertext. The mathematical function used in the Affine Cipher is of the form:

$$E(x) = (ax + b)\ mod\ m$$

where $\boldsymbol{E(x)}$ is the encrypted value of the plaintext letter $\boldsymbol{x}$**,** $\boldsymbol{a}$ and $\boldsymbol{b}$ are the encryption keys, and m is the size of the alphabet (usually 26 for the English language). The keys $\boldsymbol{a}$ and $\boldsymbol{b}$ must be carefully chosen to ensure the cipher is secure.

To decrypt the ciphertext, we use the inverse function:

$$(y) = a^{-1}(y - b)\ mod\ m$$

where $\boldsymbol{D}(\boldsymbol{y})$ is the decrypted value of the ciphertext letter $\boldsymbol{y}$, and $\boldsymbol{a^{-1}}$ is the modular inverse of $\boldsymbol{a}$.

The Affine Cipher is relatively easy to implement, but it is vulnerable to cryptanalysis. For example, if the attacker knows any two plaintext-ciphertext pairs, they can easily recover the encryption keys and decrypt any message encrypted with the same keys.

Hill Cipher

The Hill Cipher is a type of polygraphic substitution cipher where blocks of n plaintext letters are mapped to blocks of n ciphertext letters using a matrix multiplication operation. The matrix used in the encryption and decryption process is called the "key matrix" and must be carefully chosen to ensure the cipher is secure.

The Hill Cipher works as follows: given a plaintext block $\boldsymbol{P}$ of $\boldsymbol{n}$ letters and a key matrix $\boldsymbol{K}$, we first convert $\boldsymbol{P}$ into a vector $\boldsymbol{p}$ of $\boldsymbol{n}$ numeric values using a pre-defined mapping of letters to numbers (e.g., A=0, B=1, ..., Z=25). We then encrypt $\boldsymbol{p}$ to obtain the ciphertext block $\boldsymbol{C}$ of n letters using the matrix multiplication operation:

$$C = K * p\ (mod\ m)$$

where mod m means that we take the result modulo m for each element in the matrix multiplication. Finally, we convert $\boldsymbol{C}$ back into a block of $\boldsymbol{n}$ letters using the reverse mapping.

To decrypt the ciphertext, we use the inverse matrix of **K**:

$$K^{-1} = (ad(K) * det(K)^{-1})\ mod\ m$$

where $\boldsymbol{adj}(\boldsymbol{K})$ is the adjugate matrix of $\boldsymbol{K}$, $\boldsymbol{det(K)}$ is the determinant of $\boldsymbol{K}$, and $*$ denotes matrix multiplication. We first compute the adjugate matrix and the modular inverse of the determinant, and then compute the inverse matrix $\boldsymbol{K^{-1}}$ using the above formula. We then apply the same matrix multiplication operation as in the encryption process to obtain the plaintext block $\boldsymbol{P}$.

The Hill Cipher is more secure than the Affine Cipher because it is resistant to simple crypt-analysis techniques, such as frequency analysis. However, it is still vulnerable to more advanced attacks, such as known plaintext attacks or chosen plaintext attacks, if the key matrix is poorly chosen.

Example 1. Encode the message (JCPHQOPNUZ) by using Hill cipher algorithm where the matrix is:

$$\begin{array}{llll} & 1 & 2 & 3 \\ A = (4\ 7 & 5 & 6) \\ & & 8 & 9 \end{array}$$

Solution.

1. Compute the inverse key matrix:

To decrypt the message, we need to compute the inverse of the key matrix A modulo 26. We can do this using modular arithmetic and matrix operations, or we can use an online calculator. Let's assume we have computed the inverse matrix A:

$$\begin{array}{llll} & 9 & 2 & 23 \\ A^{-1} = (16\ 15 & 22 & 1\ ) & mod(26)\ 17 \\ & 14 & & \end{array}$$

2. Convert the ciphertext to a matrix:

Let's say our ciphertext message is "JCPHQOPNUZ". We can group the letters into blocks of size 3:

JCP HQO PNU Z

We can then convert each block to a matrix using a substitution cipher. For example, we could use the following substitution cipher:

A B C D E F G H I J K L M 0 1 2 3 4 5 6 7 8 9 10 11 12 N O P Q R S T U V W X Y Z
13 14 15 16 17 18 19 20 21 21 23 24 25

We can then convert each block to a matrix using the same substitution cipher we used to encrypt the plaintext message. For example, the first block "JCP" would be represented by the matrix:

(9 2 15)

(7 16 15)

(15 13 20)

Similarly, we can represent the other blocks as matrices. 6. Decrypt the message using the Hill

Cipher:

To decrypt the message, we multiply each ciphertext block matrix by the inverse key matrix A^{-1} modulo 26. For example, if we have four ciphertext block matrices C1, C2, C3, and C4, the decrypted message would be: $P = (A^{-1} * C1) \mid\mid (A^{-1} * C2) \mid\mid (A^{-1} * C3) \mid\mid A^{-1} * C4)$

where || denotes matrix concatenation. Using matrix multiplication, we can compute:

$$\begin{array}{llllllll} & 9 & 2 & 15 & 5 & 7 & 10 & \\ A^{-1} * (\ 7 & & 16 & 15) = (11 & & 18 & 2\ ) & mod(26) = 16 \\ & 15 & 13 & 20 & 24 & 25 & & \end{array}$$

$$\begin{array}{llll} & 5 & 17 & 0 \\ = (11 & & 18 & 2\ ) \\ & 24 & 25 & 16 \end{array}$$

So, the decrypted first block is represented by the matrix [5 17 0], which corresponds to the plaintext block "WEA". We can similarly compute the other decrypted blocks.

4. Convert the decrypted matrix to plaintext:

To convert the decrypted message back to plaintext, we use the same substitution cipher we used to convert the plaintext to a matrix. For example, the decrypted block represented by the matrix [5 17 0] corresponds to the plaintext block "WEA". We can similarly convert the other decrypted blocks to plaintext.

So, the decrypted message would be "WELOVEMATH".

7. CONCLUSIONS

Matrices play a significant role in cryptography, especially in symmetric key encryption algorithms like the Hill Cipher. The Hill Cipher uses matrix multiplication to perform encryption and decryption operations, which makes it a powerful tool for securing information. The use of matrices in cryptography allows for the creation of more complex encryption algorithms that offer better security than simpler techniques. By using matrices, encryption algorithms can be designed to be resistant to known plaintext attacks and brute force attacks, which are common techniques used by attackers to break encryption.

The Affine Cipher and Hill Cipher are two important encryption algorithms used in cryptography. both the Affine Cipher and Hill Cipher are important encryption techniques that have their own strengths and weaknesses. They are useful tools for securing information, but their effectiveness depends on the specific use case and implementation. It is important to carefully consider the security requirements and limitations of each cipher when choosing an encryption algorithm for a given task.

REFERENCES

[1] Al Etaiwi, W. M. (2014). Encryption algorithm using graph theory. Journal of scientific research and reports, 3(19), 25192527.

[2] Kumar, S. N. (2015). Review on network security and cryptography. International Transaction of Electrical and Computer Engineers System, 3(1), 1-11.

[3] Behrouz A Forouzan, “Data Communications and etworking”, cGraw-Hill, 4th Edition.

[4] Larson, R., & Falvo, D. C. (2010). Fundamentos de álgebra lineal, Sexta edición.Cengage Learning Editores, S.A. de C.V.

[5] Maetouq, A., Daud, S. M., Ahmad, N. A., Maarop, N., Sjarif, N. N. A., & Abas, H. (2018). Comparison of hash function algorithms against attacks: A review. International Journal of Advanced Computer Science and Applications, 9(8).

[6] Huang, K. H., & Abraham, J. A. (1984). Algorithm-based fault tolerance for matrix operations. IEEE transactions on computers, 100(6), 518-528.

[7] Arpita Bose, "Matrix-based Cryptography: A Survey", 2017.

[8] Xiaoyan Liu, "Secure Matrix Encryption Scheme Based on NonAssociative Algebraic Structure", 2019.

[9] Jianhua Wang, "Design of Matrix-based Image Encryption Scheme Using Elliptic Curve Cryptography", 2020.

[10] Muhammad Ilyas, "Matrix-based Cryptography for 5G Wireless Networks", 2020.

[11] Baojiang Cui, "Efficient Error Correction Schemes Using Matrices in Cryptography", 2021.

[12] Rao, U. H., & Nayak, U. (2014). The InfoSec Handbook. Apress Media, LLC.

[13] D. Rachmawati and M. A. Budiman, "New approach toward data hiding by using affine cipher and least significant bit algorithm," 2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT), Kuta Bali, Indonesia, 2017, pp. 1-6, doi: 10.1109/CAIPT.2017.8320737.

Solar Energy Forecasting in Vlora using Artificial Neural Networks and Open Data

Dezdemona Gjylapi[1], Alketa Hyso[1], Eljona Proko[1] and Sokol Gjylapi[2]

[1]Computer Science Department, University "Ismail Qemali", Vlora, Albania

[2]State Police Department, Vlora, Albania

Abstract. Solar energy forecasting is considered an essential scientific aspect in supporting efforts to integrate solar energy into electricity grids. This is because grid operators need to know how much solar energy the system is producing so they can optimally engage solar and other energy sources to balance demand and production. Improving solar power forecasts will allow the electric grid to be more flexible and adapt to changing conditions. This will in return help minimize outages and the overall cost of service.

Artificial Neural Networks (ANN) are powerful tools for modeling and estimating solar energy even though they use few inputs. To train the networks, a dataset of daily meteorological time series for a period of 12.5 years (2010–2022) collected for the city of Vlora by Weather Data Service Visual Crossing, and publicly accessible, was used. The meteorological parameters used to estimate solar energy were daily values of the maximum, minimum, and average temperatures; relative humidity; daylight hours; cloud cover; solar radiation, and weather description as inputs. The output is the daily solar energy in MJ/m 2.

In this paper, we tested two main models: one model using the average solar radiation variable in addition to the others, and the second model using only weather data that are more easily measurable for almost every area and accessible to everyone. Various ANN models have been designed and implemented by combining the above-mentioned parameters.

The obtained results showed that the ANN model can be successfully used to estimate the daily solar energy for Vlora, and can be used for other locations too. The data used is Open Data, which makes the model very suitable to use for other regions as well.

Keywords: Solar radiation; Solar Energy; Artificial Neural Networks; Forecasting; Open Data;

1. Introduction

Electricity consumption has increased significantly in recent decades. Traditional energy, such as oil, coal, and nuclear, hurt the environment, causing major climate changes.

Climate change has brought an increasingly strong impact on governments for the sustainable development of renewable energy across the globe. This is reflected within the EU's 2030 agenda, which envisions a future where there is universal access to affordable, reliable, and sustainable energy.

For these reasons, researchers are focused on renewable sources of energy such as solar, marine, and wind energy.

One of the most important energies in recent years has been solar energy. Solar energy can be divided into two different types of energy: solar thermal energy, which converts solar radiation into thermal energy to power industrial processes, desalination plants, homes, or water purification plants, among others; and photovoltaic solar energy, through which solar radiation is converted into electricity that can be used throughout the electricity grid.

Scientists are trying to improve the efficiency of converting this energy into electricity, using different technologies. A drawback to using solar energy is that it depends on weather variables such as air pollution, wind velocity, and cloud cover. This weather variability makes solar energy potentially unreliable. So, the success of the application of these technologies is highly related to the amount of solar energy available. An efficient conversion and utilization of solar energy systems require accurate, detailed, and long-term knowledge of available solar energy data in various forms, depending on the related application [1].

The cost, the difficulty in measurement, and the lack of accuracy of the measuring equipment are the main elements that make solar radiation data not as available as meteorological data such as temperature, humidity, or wind speed. For this reason, it is essential to develop alternative ways to generate the required data [2].

Solar energy forecasting is considered an essential scientific aspect in supporting efforts to integrate solar energy into traditional electricity grids. This is because grid operators need to know how much solar energy the system is producing so they can optimally engage solar and other energy sources to balance demand and production. Improving solar power forecasts will allow the electric grid to be more flexible and adapt to changing conditions. This will in return help minimize outages and the overall cost of service.

In this study, an artificial neural network (ANN), which is a numerical modeling technique, was used to estimate daily solar energy. ANNs can take multiple input variables to predict multiple output variables. ANNs differ from statistical modeling techniques, as they can learn about the system to be modeled without prior knowledge of the process relationships [3].

The prediction by a well-trained ANN is normally much faster than the conventional simulation programs or mathematical models as no lengthy iterative calculations are needed to solve differential equations using numerical methods. By the way, the selection of an appropriate neural network topology is important in terms of model accuracy and model simplicity. In addition, it is possible to add or remove input and output variables in the ANN algorithm if it is needed [4].

Many studies are implemented to develop for predicting global solar radiation (GSR) using different techniques such as ANN, fuzzy control, and empirical models. These techniques depended on different types of datasets (such as meteorological and geographical) [5].

Numerous meteorological and geographical variables such as maximum temperature, relative humidity, sunshine duration, cloud cover, latitude, longitude, and altitude have been used to develop ANN models for solar energy or global solar radiation prediction [6].

2. MATERIALS AND METHODS

1. Data

Despite the large spectrum of applications demanding solar radiation data, such direct measurements of solar energy are not widely available, rendering the use of numerical techniques an essential alternative. With such indirect techniques, other observed meteorological data are mathematically exploited to estimate the amounts of GSR reaching the earth. Except for GSR, other meteorological data are parameters that are routinely recorded at a large number of climatological stations (manned and automatic), due to the low cost of the respective recording instrumentation and the easiness of data acquisition [2].

Measured daily data for 12.5 years for the period of 2010–2022 were collected from https://www.visualcrossing.com/weather-data for the Vlora region.

The meteorological parameters used to estimate solar energy were daily values of maximum, minimum, and mean temperatures (◦C); relative humidity (%); daylight duration (h); cloud cover (%); weather description (text values converted in numerical values from 0-5) average solar radiation (W/m^2 day). It also used the number of day in the year as an input parameter.

Visual Crossing measures solar radiation as the amount of solar radiation per unit area per second. This is sometimes named 'solar irradiance' and is typically measured in Watts per meter squared (W/m^2).

The total amount of solar radiation energy for a day is found by summing the individual solar radiation values for the day. As measurements are not recorded for every second, it is generally assumed that the same solar energy was recorded for the whole previous interval. Energy is typically expressed in megajoules per square meter (MJ/m^2).

The default daily solar radiation display is the mean value of the solar radiation for the day. For solar energy, the daily value is the sum of the hourly values.

The data used is Open Data, which makes the model very suitable to use for other regions as well.

2. **Neural Networks**

Artificial Neural Networks (ANNs) provide powerful models for statistical data analysis. Their most prominent feature is their ability to "learn" dependencies based on a finite number of observations [7].

ANN is a mathematical model that performs a computational simulation of the behavior of its biological counterpart. Neural networks have been used in many business applications for pattern recognition, forecasting, prediction, and classification, due to their ability to "learn" from the data, their nonparametric nature, and their ability to generalize [8].

The first step in developing an ANN comprises the definition of the network architecture, which is defined by the basic processing elements (i.e. neurons) and by how they are interconnected (i.e. layers). An ANN may have one or more layers of neurons, which may be fully or partially connected. Each link between two nodes has a weight w_{ij}, which summarizes the knowledge of the system. The processing of the existing cases with the inputs: x_1, x_2, xj, and the expected results, will adjust these weights based on the difference between the actual and the expected results. The input layer nodes are passive, doing nothing but relaying the values from their single input to multiple outputs, while the hidden layer nodes and output layer nodes are active [9]. The number of neurons in the input layer depends on the independent variables, whereas the number in the output layer on the dependent variable.

The second step deals with the definition of NN Learning, which implies that a processing unit is capable of changing its input or output behavior as a result of changes in the environment i.e. adjusting the weights based on input vector values.

The third step finalizes the definition of the data used for training, testing, and validating the neural network. The dataset is divided into training, validation, and testing subsets. The data used for validation and testing were not used during training.

The network training process follows these steps:

1. Initialize weights with random values and set other parameters, such as learning rate, momentum, and bias.
2. Read the desired input and target vectors.
3. The estimated output is calculated, passing forward through the layers where the input values in the neuron are subjected to the selected activation function. In this study, the activation function was sigmoid and linear in the hidden and output layers, respectively.
4. The error is calculated as the difference between the target and output values.

5. The weights are modified by working backward from the output layer through the hidden layers

The following Fig.1 gives the error flow (backpropagation) for a neuron.

Fig. 1. Error flow (backpropagation) for a neuron

Since ANNs are constructed with layers of units, they are termed multilayer ANNs. Multilayer Perceptron (MLP) is perhaps the most common type of feed-forward network. MLP networks are used in a variety of problems especially in forecasting because of their inherent capability of arbitrary input-output mapping [2].

In this study, MATLAB R2021a software is used for designing and testing ANN models. It is used a two-layer feed-forward network trained with a backpropagation (BP) algorithm, and error minimization is obtained by Levenberg–Marquardt (LM) procedure.

Fig.2 shows a typical configuration that consists of the input layer, hidden layer, and output layer.

Fig. 2. A typical feedforward neural network (MLP). Tmin, Tmax, Tavg: temperature, H: humidity, Cc: Cloud Cover, Dh: daylight hours, Ny: number of day in the year, and Dw: description of the weather

3. ANN models for solar energy forecasting

To consider the effect of each input variable on the prediction of solar energy, different models were designed. Each model was built with a different number of neurons in the input layer, starting with three parameters and adding every time a different one. If adding a parameter as an input to the model improved the performance of the model, the parameter was kept, otherwise, the parameter was removed from the model. The number of neurons in the hidden layer is calculated as 2/3 of the number of neurons in the input layer plus the number of neurons in the output layer. We have also tested some models increasing the number of neurons in the hidden layer. What we noticed was that the performance did not improve.

To avoid using the average daily solar radiation, due to unavailability or errors in estimation, in different locations, different models were designed and implemented that took as input simple parameters to measure (weather conditions).

To evaluate the difference between measured and predicted values by ANN models, mean absolute error (MAE), mean square error (MSE), and correlation coefficient (R) were determined.

The following Table 1 gives a summary of the results obtained for ANNs using weather condition parameters as input.

Table 1. Different ANN-based models using weather conditions and their performance

Input Parameters	ANN architecture	MSE	MAE	R
Tmax, Tmin, Cc	3-2-1	23.023	3.995	0.831
Tmax, Tmin, Cc, Dh	4-4-1	4.844	1.484	0.966
Tmax, Tmin, Cc, Dh, Dw	5-4-1	4.9	1.4851	0.966
Tmax, Tmin, Cc, Dh, H	5-4-1	4.784	1.4886	0.967
Tmax, Tmin, Cc, Dh, Ny	5-5-1	4.441	1.4686	0.969

Tmax, Tmin, Cc, Dh, Ny, Tavg	6-5-1	4.682	1.5438	0.968
Tmin, Tmax, Tavg, H, Cc, Dh, Ny, Dw	**8-6-1**	**4.225**	**1.4**	**0.971**

The following Table 2 gives a summary of the results obtained for ANNs using daily average solar radiation as input.

Table 2. Different ANN-based models using also solar radiation (Sr) and their performance

Input Parameters	ANN architecture	MSE	MAE	R
Tmin, Tmax, H, Sr, Dh, Dw	6-5-1	0.0192	0.0867	0.999
Tmin, Tmax, Sr, Ny, Dh	**5-4-1**	**0.0161**	**0.0864**	**0.999**

The optimal model for estimating solar energy was an ANN with one hidden layer where the inputs were daily numerical values for maximum and minimum temperature, daylight hours, average solar radiation, and the number of day in the year (***ANN1***: 5-4-1 topology). Its mean absolute error (MAE), mean square error (MSE), and regression value (R) were found to be 0.999, 0.019, and 0.0864, respectively.

The best model with simple weather conditions as input parameters but still very good performance was an 8-6-1 neural network, where the maximum, minimum, and average temperature; daylight hours; humidity, cloud cover, weather description, and the number of day in the year were used as input parameters (***ANN2***: 8-6-1 topology). For this model, the R, MSE, and MAE values were found to be 0.971, 4.225, and 1.4, respectively.

Fig.3 presents the regression analysis of daily values calculated from the selected ANN1 model for the training dataset, valuation dataset, testing data set, and all the data for the ANN1 model. It can be seen, the selected ANN model performs perfectly well compared to solar energy measurements.

Fig. 3. Regression R values for ANN1 model

Fig.4 shows the comparison between measured values of solar energy and predicted values for this ANN. It can be seen that the evolution is similar and one line is practically superimposed over the other.

Fig. 4. Comparing the daily measured solar energy with ANN1 model estimated solar energy.

Fig.5 shows the regression for the training dataset, valuation dataset, testing data set, and all the data for the ANN2 model, and it can be seen, the selected ANN model performs very well compared to solar energy measurements.

Fig. 5. Regression R values for ANN2 model

Fig.6 shows the comparison between measured values of solar energy and predicted values for ANN2. It can be seen that the evolution is similar to the real data values but not as good as the estimation done by the ANN1 model. Anyway, it is a very good one.

Fig. 6. Comparing the daily measured solar energy with the ANN2 model estimated solar energy

3. Conclusions

In this paper, we presented an approach to forecasting solar energy. The forecasting performance of various ANNs was assessed by comparing their predictions to actual data available online at https://www.visualcrossing.com/weather-data. One of the advantages of the technique used in this paper is that it requires minimal preprocessing of the input data.

ANN with maximum and minimum temperature, daylight hours, average solar radiation, and the number of day in the year as inputs lead to maximum R and minimum MSE and MAE. The topology of this ANN is 5-4-1; a single hidden layer with 4 neurons showed the best structure among other models.

Also, ANN with maximum, average, and minimum temperature, daylight hours, weather description, cloud cover, humidity, and the number of day in the year as inputs resulted to be a very good forecasting model too, with the structure 8-6-1.

The results indicate that the ANN model seems promising for evaluating the solar energy potential in Vlora, but also in other regions where there are no monitoring stations because they can predict very well even using only weather conditions.

References

1. Sozen, A., E. Arcaklioglu, and M. Ozalp. (2004). Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. Energy Conversion and Management 45:3033–52.
2. Zeynab Ramedani, Mahmoud Omid & Alireza Keyhani (2013) Modeling Solar Energy Potential in a Tehran Province Using Artificial Neural Networks, International Journal of Green Energy, 10:4, 427-441, DOI: 10.1080/15435075.2011.647172
3. Barma, S. D., B. Das, A. Giri, S. Majumder, and P. K. Bose. (2011). Back propagation artificial neural network (BPANN) based performance analysis of diesel engine using biodiesel. Journal of Renewable and Sustainable Energy 3:013101.Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
4. Ramedani Z., Omid M., and Keyhani A. (2013). Modeling Solar Energy Potential in a Tehran Province Using Artificial Neural Networks. International Journal of Green Energy Volume 10, 2013 - Issue 4. Pages 427-441
5. Mohamed, Z.E. Using the artificial neural networks for prediction and validating solar radiation. *J Egypt Math Soc* **27**, 47 (2019). https://doi.org/10.1186/s42787-019-0043-8
6. J. Mubiru, and E. J. K. B. Banda (2008), "Estimation of monthly average daily global solar irradiation using artificial neural networks," Solar Energy, vol. 82, pp. 181-187
7. Antanasijević, Davor; Pocajt, Viktor; Popović, Ivanka; Redžić, Nebojša; Ristić, Mirjana: The forecasting of municipal waste generation using artificial neural networks and sustainability indicators, Sustainability Science; Jan2013, Vol. 8 Issue 1, p37
8. Priyanka Gaur, Neural Networks in Data Mining, International Journal of Electronics and Computer Science Engineering, ISSN- 2277-1956, Volume 1, Number 3 (p 1449-p1453)
9. Gjylapi, D.; Durmishi, V. (2014) "Artificial Neural Networks in forecasting tourists' flow, an intelligent technique to help the economic development of tourism in Albania",

Academicus: International Scientific Journal Vol.10, PP 102-112 ISSN: 2079-3715, eISSN:2309-1088

Graphic presentation and statistical data analysis of the number of children in the families of 7th grade students in a lower secondary school in the Republic of Kosovo

Nazmi Misini [1]

[1] **UBT,Higher Education Institution,Kosova**

[1] **nazmi.misini@ubt-uni.net**

Abstract

Statistics is a science that researches, examines and studies the quantitative conditions of various phenomena: demographic, social, cultural, educational, economic, agricultural, commercial, industrial, communicative, physical, chemical, etc. Therefore, statistics is classified in the social sciences, which object the study has mass phenomena and the characteristics of their variation. These mass phenomena include, for example, the population,economy, agriculture, industry, communication, culture, education, etc., each in their quantitative aspect and the corresponding quantitative characteristics, including the time and space factor.

Today, statistics is not limited to numerical information needed for the state, but has entered all social pores and natural sciences. Without statistics, they cannot function efficiently: accounting, finance, marketing, production, economy, biology, physics, sociology, psychology, pedagogy, astronomy, medicine, etc. Every day we come across numbers of statistical data, expressed in numbers or percentage.

The graphic presentation is one of the most efficient tools both for the visual description of the results of multiple surveys of one or several characteristics of a statistical population, as well as for discovering the relationships and connections between these characteristics or between changes in time and space of phenomena. Graphical representation facilitates understanding much more quickly than the presentation of a multitude of figures, performing a great service to science and constituting a very valuable aid to statistical studies.Statistical analysis is of particular importance because through it we can compare data and research results for two or more phenomena, in time and space.

This paper will deal with the main concepts in statistics, including concrete examples. First, we will explore the statistics of birth data at the level of the Republic of Kosovo, and then we will separately analyze our case of the number of children in families of students in a class in a lower secondary school in a village in Kosovo.

Keywords: children, family, class, school

Entry

Family-wikipedia (lat. familia, from famulus – house slave, servant) is the smallest social unit consisting of two or more persons who are related by marriage, blood or adoption; who live in the same environment for a relatively long time, have common cultural and economic relations and care for each other.

According to Article 16(3) of the Universal Declaration of Human Rights "The family is the basic group unit of society and is given protection by society and the state"

According to FGJSSH-FAMILY p. sh.- Small unit of organization of social life, consisting of husband, wife, children or other close people, who live and live together.

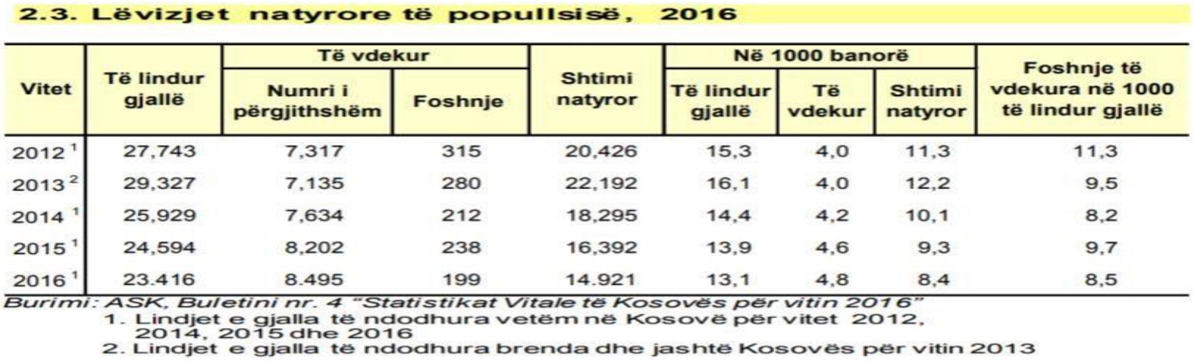
CHILDREN-wikipedia

A child is a human being in its developmental period between birth and adulthood (which includes newborns, infants, young children...).

According to FGJSSH-CHILD m. sh.

Birth of a woman; baby; minor boy or girl. Healthy (weak) child. Smart (educated) child. Darling child. Children's voices. Children's language.

The natural movement of the population from 2012 to 2016 is given by the table:



2.3. Lëvizjet natyrore të popullsisë, 2016

Vitet	Të lindur gjallë	Të vdekur		Shtimi natyror	Në 1000 banorë			Foshnje të vdekura në 1000 të lindur gjallë
		Numri i përgjithshëm	Foshnje		Të lindur gjallë	Të vdekur	Shtimi natyror	
2012 [1]	27,743	7,317	315	20,426	15,3	4,0	11,3	11,3
2013 [2]	29,327	7,135	280	22,192	16,1	4,0	12,2	9,5
2014 [1]	25,929	7,634	212	18,295	14,4	4,2	10,1	8,2
2015 [1]	24,594	8,202	238	16,392	13,9	4,6	9,3	9,7
2016 [1]	23.416	8.495	199	14.921	13,1	4,8	8,4	8,5

Burimi: ASK, Buletini nr. 4 "Statistikat Vitale të Kosovës për vitin 2016"

1. Lindjet e gjalla të ndodhura vetëm në Kosovë për vitet 2012, 2014, 2015 dhe 2016
2. Lindjet e gjalla të ndodhura brenda dhe jashtë Kosovës për vitin 2013

From the table above, we see that for every year we have a decrease in the birth of children in Kosovo. From the data of KAS, there is a decrease in the number of the population due to the decrease in the number of births. From the data of the resident population in Kosovo, it results that the total population in the country at the end of 2014 was 1,804,944 inhabitants, while for 2015 was 1,771,604.

The size of the family economy is 5.71 people.

The population density for 2015 was 162.41 inhabitants per km^2

The average age of unmarried people for 2015 was 29.1 years for women 27.3 years, while for men 30.8 years.

Fertility has decreased compared to previous years from 3 children per woman in 2003 to 2 children in

2014.

Meanwhile, according to the gender structure, the total resident population in 2015 is as follows: Males:
49.8% and Females: 50.2%.

The average age of the population of Kosovo was 30.2 years, while the average life expectancy in Kosovo was estimated to be 79.4 years for women and 74.1 years for men.

Leposaviq
Mitrovicë
Zubin Potok
Zveçan
Mitrovicë Veriore
Podujevë
Istog
Vushtrri
Skënderaj
Pejë
Obiliq
Prishtinë
Klinë
Gllogoc
Fushë Kosovë
Kamenicë
Deçan
Novobërdë
Lipjan
Ranillug
Junik
Malishevë
Shtime
Gjilan
Gjakovë
Rahovec
Suharekë
Ferizaj
Partesh
Viti
Mamushë
Prizren
Shtërpcë
Kaçanik
Kllokot
Hani i Elezit
Dragash
0
20
40
km

The number of students over the years in the said school is as follows:

year	2008/09	09/10	10/11	11/12	12/13	13/14	14/15	15/16	16/17	17/18
No of students	481	482	440	405	385	381	376	344	338	316

THE MAIN QUESTION IN THIS PAPER IS: HOW MUCH IS THE AVERAGE NUMBER OF CHILDREN IN THE FAMILIES OF CLASS 7 STUDENTS OF THE GIVEN SCHOOL?

Sub-questions: What is the number of children in the dominant families? What is the smallest number of children in the family, and what is the largest?

DATA PRESENTATION

Quantitative (quantitative) population: The number of students in the lower secondary school: "Nazim Hikmet" in the village of Dobërçan, municipality of Gjilan is 316 students. Albanian students and those from the Turkish community study in this school.

Quantitative sample: Seventh grade students: (25 students) Cl. VII$_1$ (10 nx.), VII$_2$ (13 nx.), VII$_T$ (2 nx.)

First, the students were asked how many children are in their family. They stated:

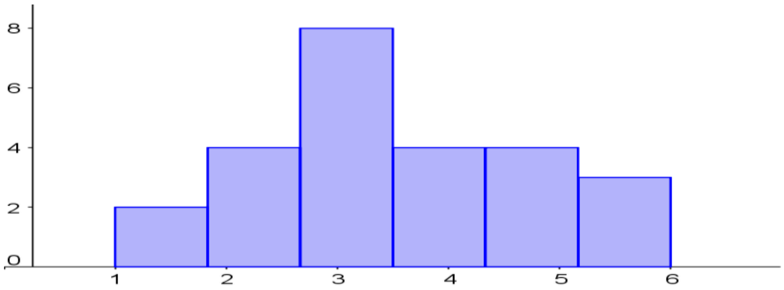
The first method: **Listing data:**2,2,3,4,3,5,3,4,5,3,3,4,5,5,1,6,2,3,3,3,4 ,6,2,1,6,or

{2,2,3,4,3,5,3,4,5,3,3,4,5,5,1,6,2,3,3,3,4,6,2,1,6 }

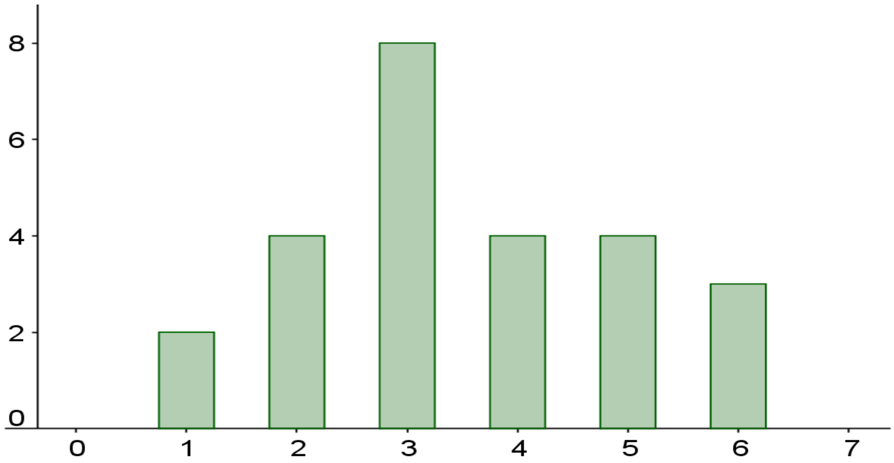
Second method: Table of frequencies (frequency):

x	1 child	2 childs	3 childs	4 childs	5 childs	6 childs
f	2	4	8	4	4	3

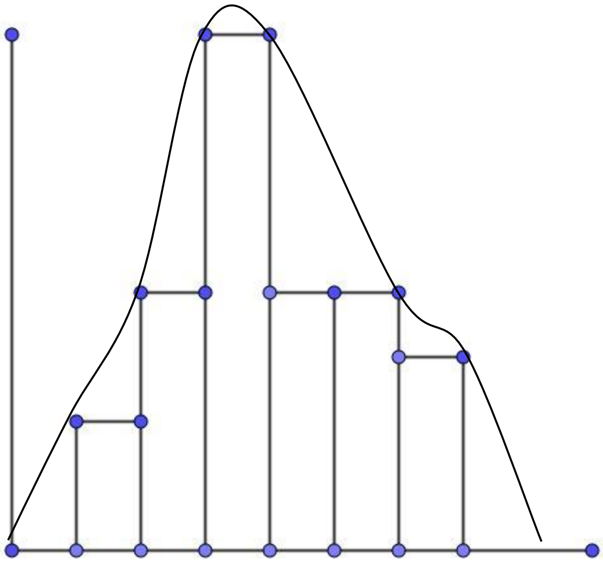
The number of children in the family	1 ch	2 ch	3 ch	4 ch	5 ch	6 ch
Number of families	2	4	8	4	4	3



Histogram of frequencies:



BAR CHART

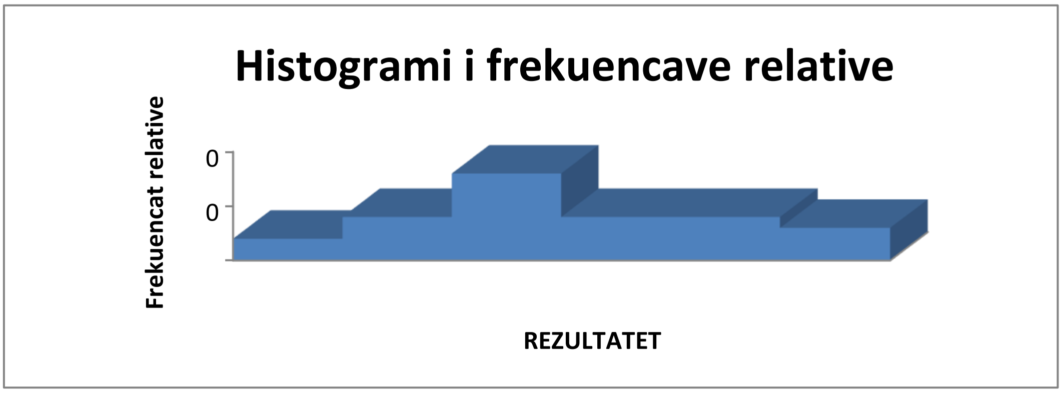


Small sample

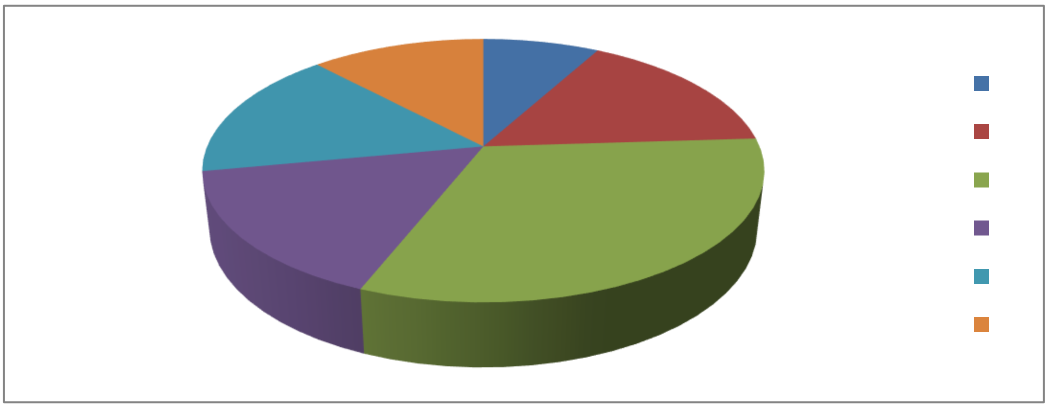
Table of relative frequencies:

x	1f	2f	3f	4f	5f	6f
$\frac{f}{n}$	0.08	0.16	0.32	0.16	0.16	0.12

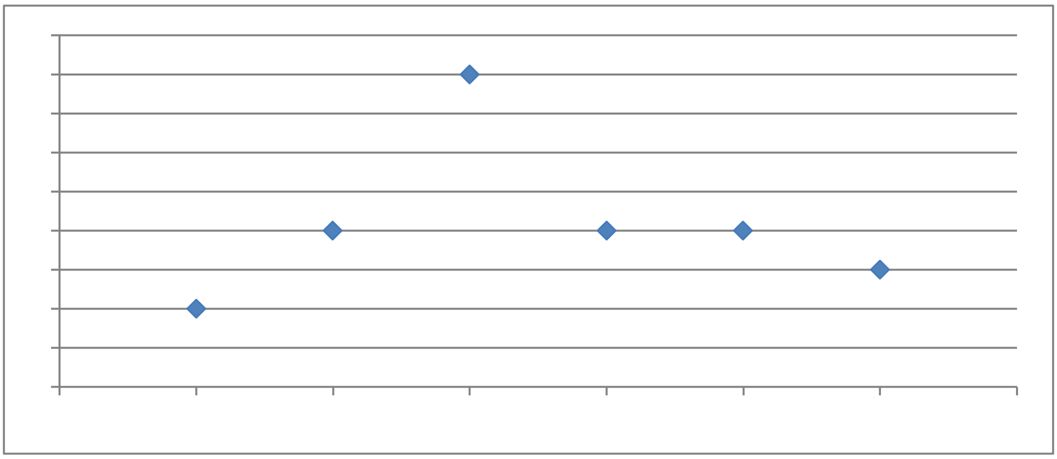
$\frac{2}{25}$ □ 0.08 , $\frac{4}{25}$ □ 0.16 , $\frac{8}{25}$ □ 0.32 , $\frac{4}{25}$ □ 0.16 , $\frac{4}{25}$ □ 0.16 , $\frac{3}{25}$ □ 0.12



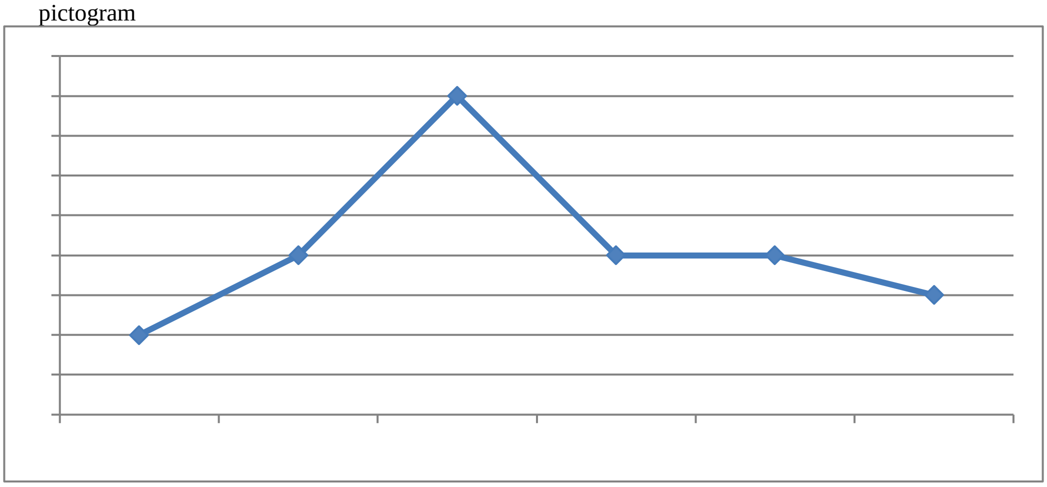
Histogram of relative frequencies:



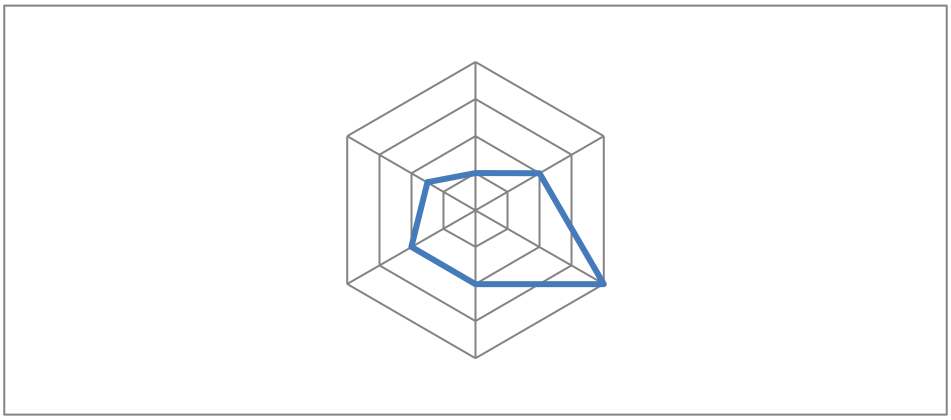
CIRCULAR DIAGRAM



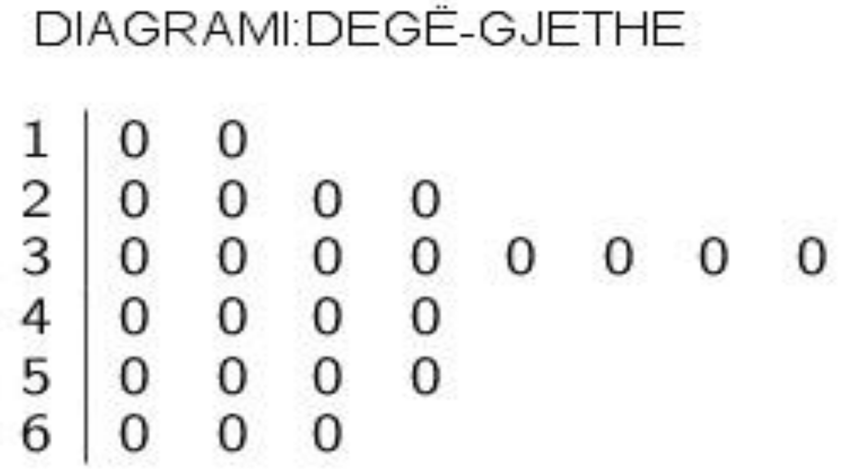
pictogram



DISTRIBUTION POLYGON



Polar diagram



DIAGRAMI:DEGË-GJETHE

1	0	0						
2	0	0	0	0				
3	0	0	0	0	0	0	0	0
4	0	0	0	0				
5	0	0	0	0				
6	0	0	0					

Branch-leaf diagram

Mean, median, mode:

$$\frac{\qquad}{25}$$

$$\square \quad \frac{\quad | \quad | \; 24 \; | \; 16 \; | \; 20 \; | \; 18}{25}$$

$$\square \quad \frac{88}{25}$$

□

x

n

□x □ 2 1□ □ 4 2□ □8 3□ □ 4 4□ □ 4 5□ □ 3 6□

□

x

□

3

.

5

2

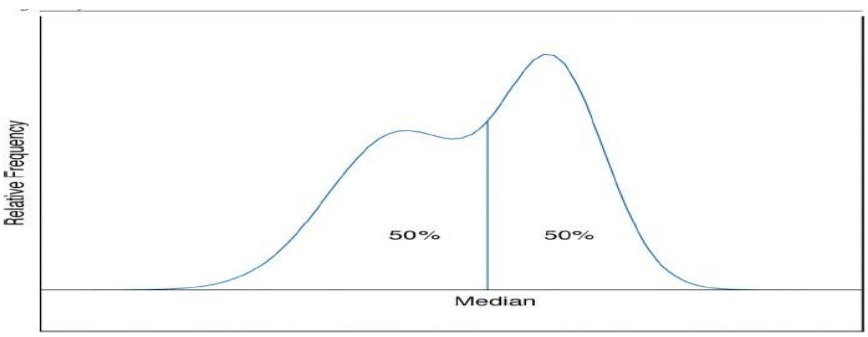
n

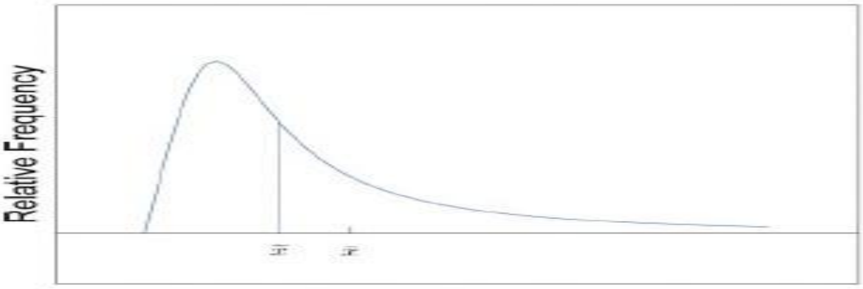
x□3.52

Sample median:

$\boxed{3}$1,1,2,2,2,2,3,3,3,3,3,3,,3,4,4,4,4,5,5,5,5,6,6,6

x□3





$(c)\ \bar{x} > \widetilde{x}$

In our case the mean is greater than the median

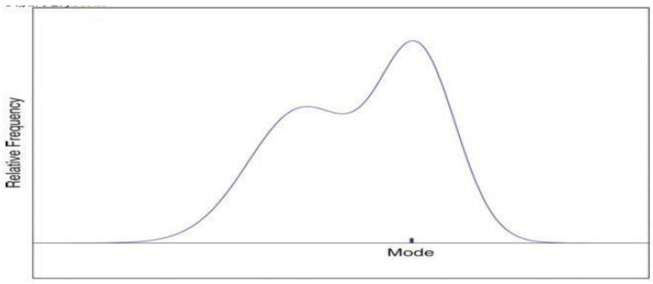
Data mode:

3

1,1,2,2,2,2,3,3,3,3,3,3,,3,4,

4,4,4,5,5,5,5,6,6,6 mode is

number 3



mode

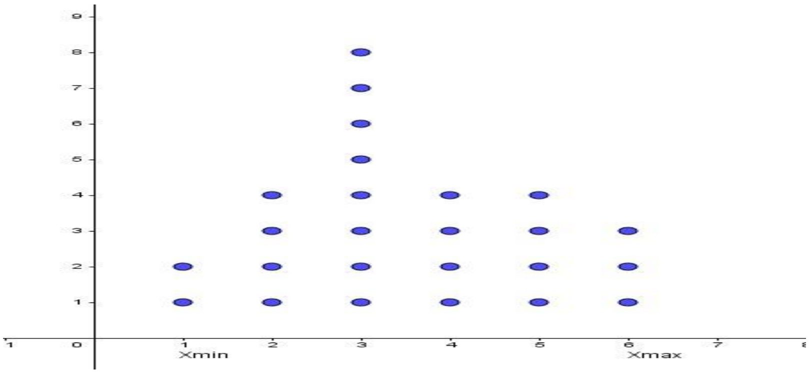
MEASURES OF VARIABILITY

Range, variance and standard deviation

3 1,1,2,2,2,2,3,3,3,3,3,3,3,,3,4,4,4,4,5,5,5,5,6,6,6

Range: R=x$_{max}$-x$_{min}$=6-1=5

R=5



Data collection

VARIANCE AND STANDARD DEVIATION

Variance

— 2 2

2 □□*nx x*□□1 □*ose s*2 □□ □*x*^{2} □*n*□*n*1
□1 *x*□ *s* □

Since $\bar{x}$ =3.52, then we will have:

1- 3.52=-2.52

2- 3.52=-1.52

3- 3.52=-0.52

4- 3.52=0.48

5- 3.52=1.48

6- 3.52=2.48

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{2(-2.52)^2 + 4(-1.52)^2 + 8(-0.52)^2 + 4(0.48)^2 + 4(1.48)^2 + 3(2.48)^2}{25-1}$$

$$s^2 = \frac{2\cdot 6.3504 + 4\cdot 2.3104 + 8\cdot 0.2704 + 4\cdot 0.2304 + 4\cdot 2.1904 + 3\cdot 6.1504}{24}$$

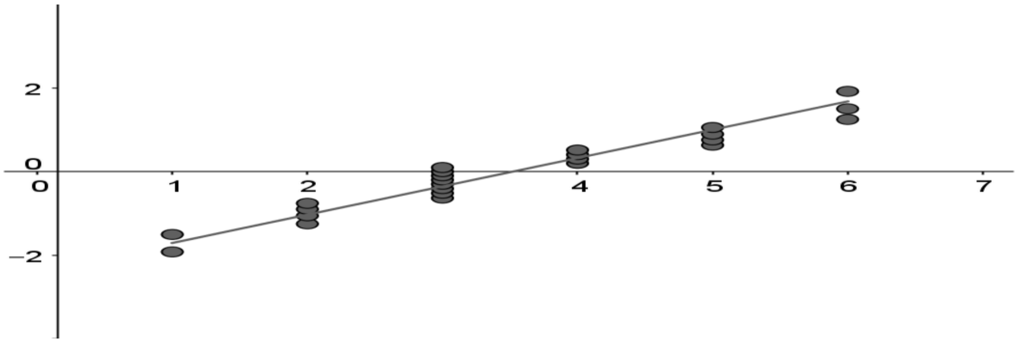
$$s^2 = \frac{12.7008 + 9.2416 + 2.1632 + 0.9216 + 8.7616 + 18.4512}{24}$$

$$s^2 = \frac{52.24}{24} = 2.176666667 \approx 2.17$$

Standard Deviation:

$\sqrt{\square \; x\bar{x}}$ *os* $\sqrt{\square\square}$

dhes $\sqrt{2.1}$ *dmt* 1.47535306



NORMAL QUANTILE

R

ELATIVE DATA POSITION

PERCENTILE:

1,1,2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,4,5,5,5,

5,6,6,6 A family with one child is:

$\frac{1}{25}$□ 0.04 *ose*4% *e*25

$\frac{2}{25}$□ 0.08 *ose*8% *e*25 that is the family with 1 child is the 4th and 8th percentile

Families with 4 children:

$\frac{15}{25}$□ 0.6 *ose*60% *e*25

$\frac{16}{25}$□ 0.64 *ose*64% *e*25

$\frac{17}{25}$□ 0.68 *ose*68% *e*25

$\frac{18}{25}$□ 0.72 *ose*72% *e*25 that is 4 children is the 60th, 64th, 68th and 72nd percentile

SECOND QUARTILE IS THE MEDIAN: $Q_{2=}$ *x*□3
We arrange the data from the smallest to the largest:

1,1,2,2,2,2,3,3,3,3,3,3,$\boxed{3}$,3,4,4,4,4,5,5,5,5,6,6,6

Q_2= *x*□3

The bottom set: $\mathbf{Q}_{\mathbf{1}:}$ 1,1,2,2,2,2,3,3,3,3,3,3

$\mathbf{Q}_{\mathbf{1}}$ is the median of the lower set:

2,3

B_p={1,1,2,2,2,,3,3,3,3,3}

Q

1

□

2
□
3
□
□
5

2
.
5
2

2

The first quartile is Q_1=2.5

Upper set Bs ={3,4,4,4,4,5,5,5,5,6,6,6}

Q_3 □ —— □□ $\frac{1}{}$ 5

The 5 summary numbers are: Xmin, Q1, Q2, Q3, Xmax.

In our case they are: **1; 2.5; 3; 5; 6**

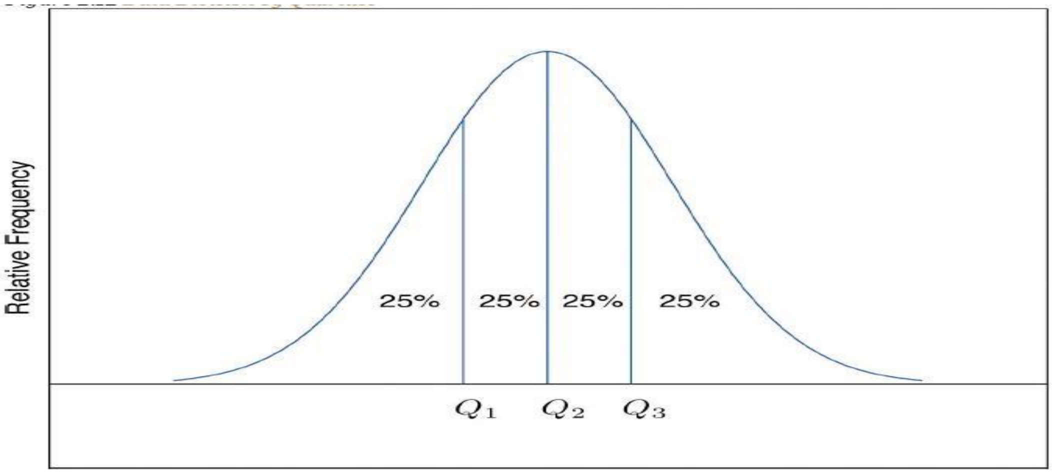
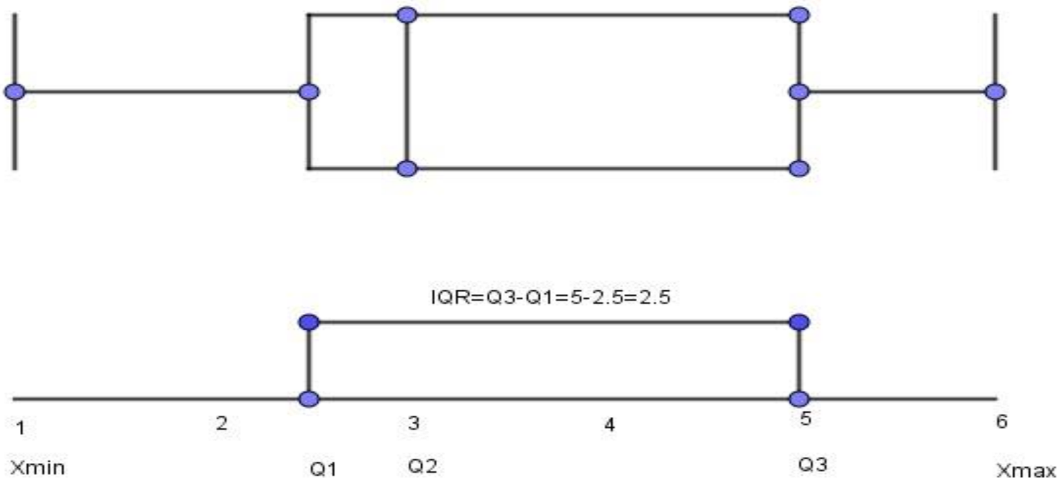


DIAGRAM “ WHISKER PLOT”



INTERQUARTILE RANGE: IQR=Q3-Q1=2.5

IQR=Q3-Q1=5-2.5=2.5

Z-scores

The Z-Score indicates how much standard deviation a particular data point is away from the data center-mean.

x x□ —

Z □ ,x□{1,2,3,4,5,6},x□ — 3.52,

s□1.48

s —

x x□ 1□3.52 □2.52

Z □ □ □ □□1.7027 s 1.48 1.48

x x□— 2 □ 3.52 □1.52

Z □ □ □ □□1.027 s 1.48 1.48 x

x□ 3□3.52 □0.52

Z □ □ □ □□3.5135 s 1.48 1.48 x

x□ 4 □ 3.52 0.48

Z □ □ □ □0.3243 s 1.48

1.48 x x□ 5□3.52

1.48 —

—

Z □ □ □ □1 s
1.48 1.48

Z □ x x□ □ 6□3.52 □ 2.48 □1.6756 s
1.48 1.48

Since Z in the first three cases is negative, while in the last three cases it is positive, then we say that:

Families with 1,2,3 children are below average, while those with 4,5 and 6 children are above average.

THE EMPIRICAL RULE AND CEBISHEV'S THEOREM

Average $\bar{x}$ = 3.52, standard deviation *s* □ 1.48

If the data has a well-shaped frequency histogram, then:

$\bar{x}$□1*s*□□ (3.52 □1.48) □ 2.04□ *x*□1*s*□□□□ (3.54 □1.48) □ 5

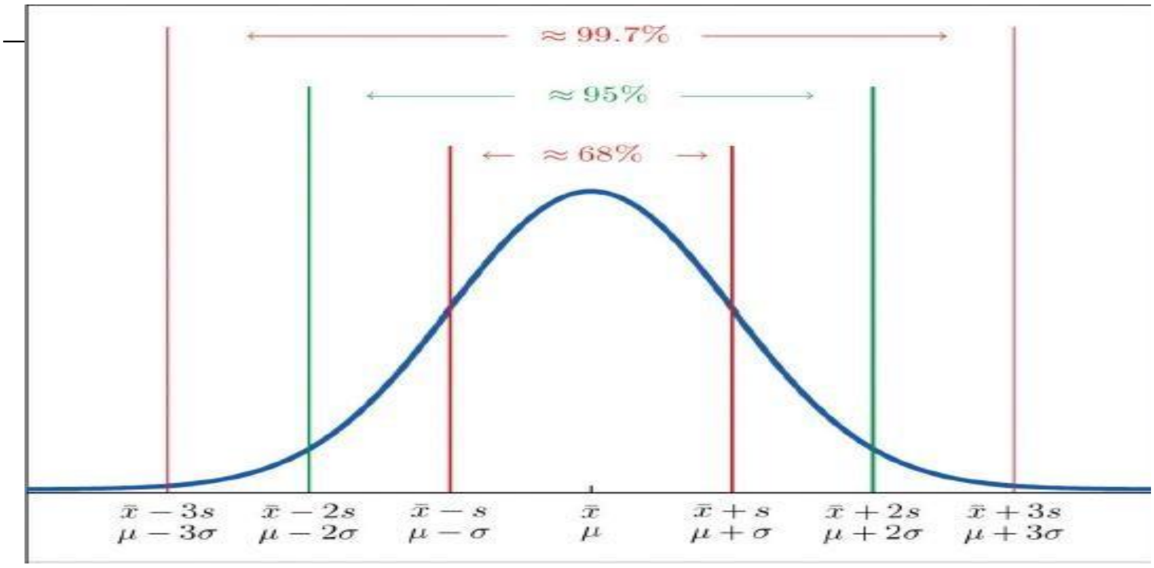
□□□ (2.04;5) □ 68% $\bar{x}$□ 2*s*□□ (3.52 □ 2.96) □ 0.56□

□□ □□ (0.56;6.48) □ 95% *x*□ 2*s*□□ (3.54 □ 2.96) □ 6.48□

$\bar{x}$□3*s*□□ (3.52 □ 4.44) □□0.96□

□□ □□ □(0.92;7.96) □ 99.7%

$\bar{x}$□ 3*s*□□ (3.54 □ 4.44) □ 7.96□

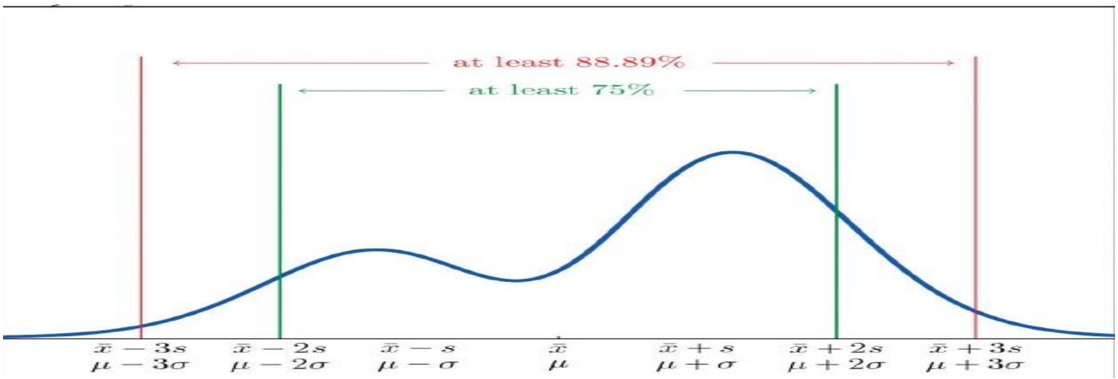


-0.92 0.56 2.04 3.52 5 6.48
7.96

In the first interval $(\bar{x}-1s, \bar{x}+1s)$ there are 20 of our data, while in the second interval

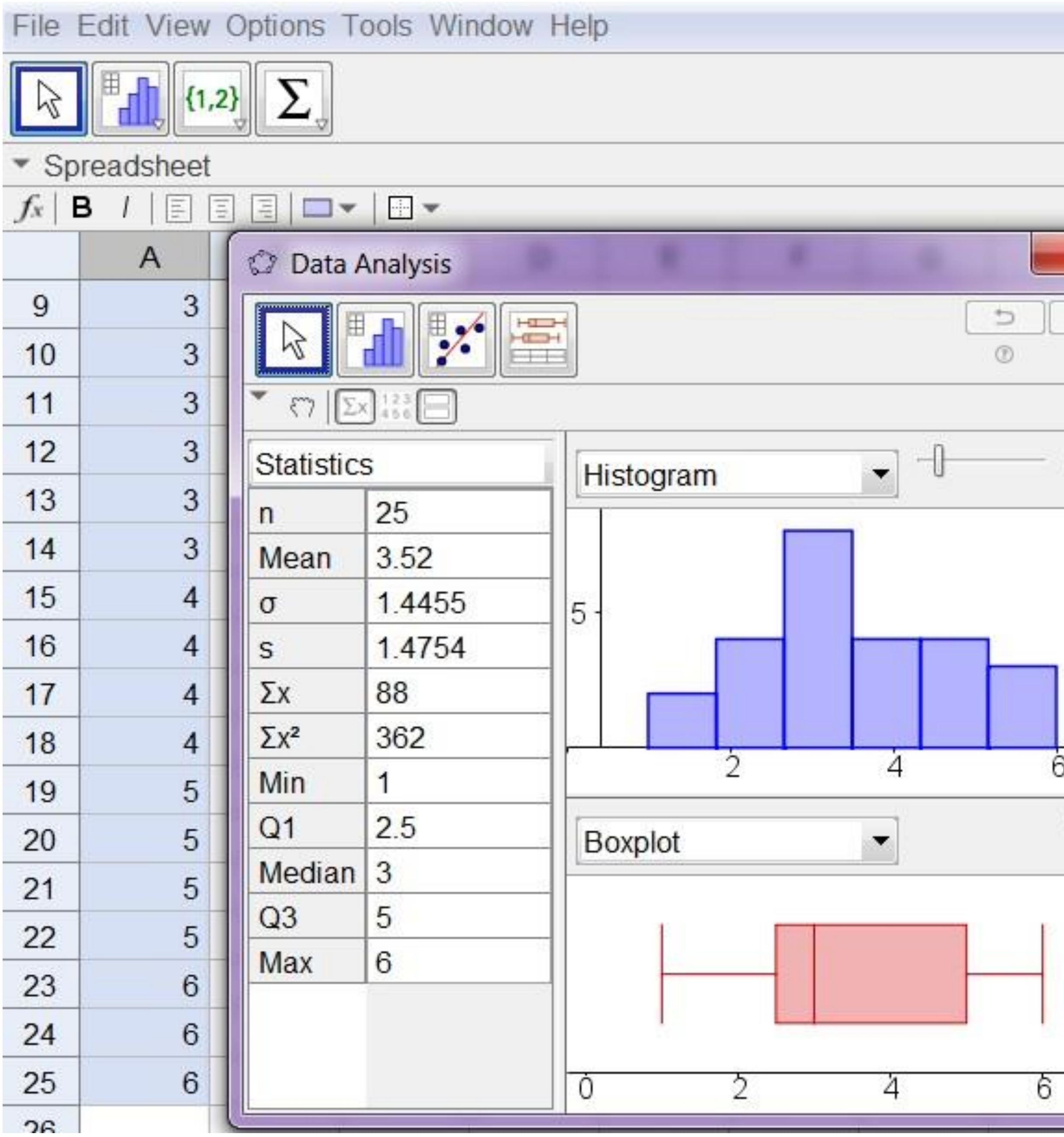
$(\bar{x}-2s, \bar{x}+2s)$ all data is located.

CHEBYSHEV'S THEOREM



-0.92 0.56 2.04 3.52 5 6.48 7.9

For any set of numerical data, at least: $\frac{3}{4}$ of data are in the interval $(\bar{x}-2s, \bar{x}+2s)$, in our case it is all the data in this interval.



	A
9	3
10	3
11	3
12	3
13	3
14	3
15	4
16	4
17	4
18	4
19	5
20	5
21	5
22	5
23	6
24	6
25	6
26	

Statistics

n	25
Mean	3.52
σ	1.4455
s	1.4754
Σx	88
Σx²	362
Min	1
Q1	2.5
Median	3
Q3	5
Max	6

Recommendation

In order that in the future we do not have a decrease in the population of Kosovo, where as a consequence of this is the decrease in the number of births in Kosovo, then there is the emigration of young people to the outside world, recommend that the Government of Kosovo and the institutions of the Republic of Kosovo to do everything possible: either through the regulation of legislation, or in other forms to help families in different forms of benefits.

These funds would be distributed to families on the occasion of the birth of children, the granting of child allowances, the regulation of health insurance, or the creation of opportunities to open job vacancies, in order for families to have a better future. safe and comfortable for increasing the fertility of children in Kosovar families.

References:

prof.Dr. Sadri Shkodra:Statistika,Universiteti “Gjilani”,Gjilan,2008.

https://sq.wikipedia.org/wiki/Familja https://sq.wikipedia.org/wiki/F%C3%femija FGJSSH http://ask.rks-gov.net/media/3385/vleresimi-i-popullsise-se-kosoves-2016.pdf http://ask.rks-gov.net/media/3598/kosova-n%C3%AB-shifra-2016.pdf http://klankosova.tv/ka-renie-te-numrit-te-popullsise-ne-kosove

Introductory Statistics

Healthcare Application for Blood Donation

Greta Ahma[1], Elton Boshnjaku[2], Edmond Hajrizi[3]

[1,2,3]UBT, Prishtina, Kosova

1 2 greta.ahma@ubt-uni.net, elton.boshnjaku@ubt-uni.net,

3 edmond.hajrizi@ubt-uni.net

Abstract. The healthcare system is one of the most important and complex industries that provides various healthcare services to meet people's health needs. It also includes efforts to influence the determinants of health, as well as more direct health-improving activities. This paper presents a high-level system to bridge the gap between blood donors and people in need of blood. The Blood Donation System application is a way to synchronize people and hospitals with their care. It is an online network through which the Hospital uses to check the availability of the required blood and can send for blood to the nearest hospital for blood or blood donation requests. The person who wants to donate blood can find the nearest blood hospital by using the blood donation application.

Keywords: Healthcare, Blood Donation System, Donor, Hospital, Patient.

Introduction

The safety of the amount of blood needed in emergency cases remains an important public health concern in Kosovo. The availability of blood of all types and their provision ensures public confidence in its excellent healthcare system. Through the use of the online blood donation system, the safety of blood transfusion is expected to increase or improve. The risks of improper documentation of blood donors and erroneous registrations can be minimized or avoided altogether. Also, the processes that include the collection of blood bags, storage, and inventory will be systematized and organized, thus improving health care management. The paper is organized as follows. Section 2 describes the conceptual design of the application. Implementation details are presented in Section 3 through code snippets extracted from the source code. In Section 4, the validation process performed on a personal computer is developed. Section 5 shows our future plans. The relevant works are developed in section 6 and finally the paper closes with concluding remarks. The Blood Donation System project was created using the Python Django Framework. The system is completely built on the Django Framework on the back-end and HTML, CSS and JavaScript on the front-end.

Conceptual design

Admin after login, can see the blood unit of any available blood group, donor number, blood request number, approved request number and total blood unit in the panel. Can view, update and delete Donor. Can view, update and delete Patient. Admin Can view the donation request made by the donor and can approve or reject that request based on the illness of the donor. If the donation request is approved by the administrator, then that unit of blood is added to the blood stock of that blood type. If the Donation Request is rejected by the administrator, then 0 units of blood are added to the store. Admin can view blood request made by donor / patient and can approve or reject that request. If the Blood Request is approved by the administrator, then this unit of blood is subtracted from that blood group's blood stock. If the Blood Request is rejected by the administrator, then 0 units of blood are reduced from the stock. Admin Can view blood request history, as well as can update the specific blood type unit.

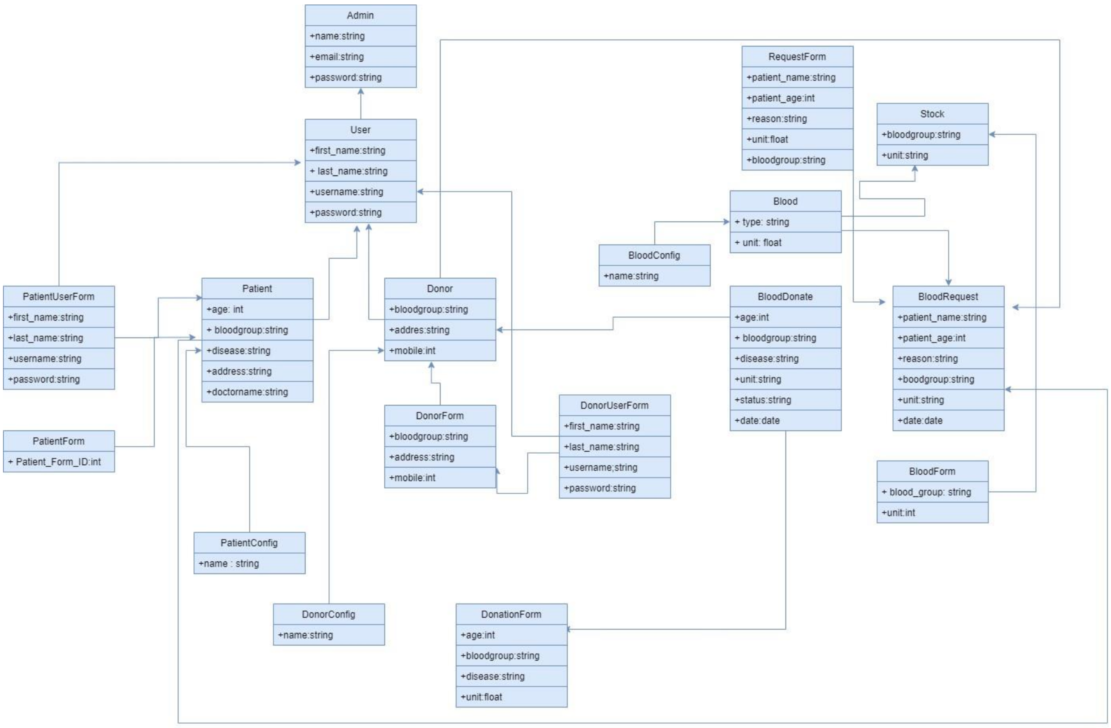


Figure 1- Class Diagram

Donor can create account by providing basic details. After logging in, the donor can donate blood. After approval by the administrator only, the blood will be added to the blood stock. Donor can view their donation history with status (Pending, Approved, Rejected). Donor can also request blood from blood stock, can see his blood request history with status, can see number of blood request made, approved, pending or rejected by admin in his dashboard. At the same time, the donor can donate blood and can also ask for blood. While Patient can create account (No admin approval required, can login after registration). After login, can see number of blood request made, approved, pending or rejected by admin on their dashboard. The patient can request blood of specific blood group and unit from the blood stock. At the same time, the patient can view the blood request history with status (pending, approved, rejected).

Implementation

The modules that are included in this application are: admin, donor and patient. To log a donor/patient into the application, they must first be registered. The part of the code that applies to registering a donor looks like this:

```
    class DonorUserForm(forms.ModelForm):
class Meta:        model=User
```

```
fields=['first_name','last_name','username','pass-
word']        widgets = {
          'password': forms.PasswordInput()
          }

    class Donor-
Form(forms.Mod-
elForm):     class
Meta:
          model=models.Donor
          fields=['bloodgroup','address','mo-
bile','profile_pic']
    def donor_signup_view(request):
userForm=forms.DonorUserForm()     do-
norForm=forms.DonorForm()
mydict={'userForm':userForm,'donor-
Form':donorForm}     if re-
quest.method=='POST':
          userForm=forms.DonorUserForm(re-
quest.POST)        donorForm=forms.DonorForm(re-
quest.POST,request.FILES)        if
userForm.is_valid() and donorForm.is_valid():
             user=userForm.save()           user.set_pass-
word(user.password)           user.save()           donor=donor-
Form.save(commit=False)           donor.user=user
donor.bloodgroup=donorForm.cleaned_data['bloodgroup']
donor.save()           my_donor_group = Group.ob-
jects.get_or_create(name='DONOR')           my_do-
nor_group[0].user_set.add(user)        return HttpRe-
sponseRedirect('donorlogin')
       return render(request,'donor/donorsignup.html',context=mydict)
```

From the figure it can be seen that the fields that must be completed during the registration of a donor are: first name, last name, username, password, blood group, address, phone number and a photo. The same applies to the registration of a patient, only that in the fields of the patient's registration, the patient's age and disease must also be filled in.

When the donor logs in to the system, the blood donation form appears, where he must fill in the blood group, how many milliliters of blood he will donate, if he has any diseases and his age. The piece of code that enables this looks like the following:

```
    class Donation-
Form(forms.Model-
Form):     class Meta:
            model=models.BloodDonate
            fields=['age','bloodgroup','disease','unit']

    def donate_blood_view(request):
        dona-
tion_form=forms.Do-
nationForm()     if re-
quest.method=='POS
T':
            donation_form=forms.Donation-
Form(request.POST)         if dona-
tion_form.is_valid():
                blood_donate=dona-
tion_form.save(commit=False)
blood_donate.bloodgroup=dona-
tion_form.cleaned_data['bloodgroup']
donor= models.Donor.ob-
jects.get(user_id=request.user.id)
blood_donate.donor=donor
blood_donate.save()            return HttpRe-
sponseRedirect('donation-history')
        return render(request,'donor/donate_blood.html',{'donation_form':donation_form})
```

Also, even though he is registered as a donor, he has the possibility of making a request for blood in case of need:

```
    class Re-
questForm(forms.Mo
delForm):     class
Meta:
            model=models.BloodRequest
            fields=['patient_name','patient_age','reason','bloodgroup','unit']

    def make_request_view(request):     bloodArray = [bmod-
els.Blood.OPositive.value, bmodels.Blood.ONegative.value, bmod-
els.Blood.APositive.value,      bmodels.Blood.ANegative.value, bmod-
els.Blood.BPositive.value, bmodels.Blood.BNegative.value,      bmod-
els.Blood.ABPositive.value, bmodels.Blood.ABNegative.value]     re-
quest_form=bforms.RequestForm()     if request.method=='POST':
          request_form=bforms.Re-
questForm(request.POST)        if re-
quest_form.is_valid():
             blood_request=request_form.save(commit=False)
blood_request.bloodgroup=request_form.cleaned_data['bloodgroup']
donor= models.Donor.objects.get(user_id=request.user.id)
blood_request.request_by_donor=donor           blood_request.save()
return HttpResponseRedirect('request-history')       return render(re-
quest,'donor/makerequest.html',{'request_form':request_form,
    'bloodArray':bloodArray})
```

When the patient logs in to the system, the blood request form is displayed, which is almost the same as the donor blood request form.

When the admin is logged in to the system, requests for blood from the patient or donor, and blood donation from the donor can be approved or rejected by the admin. For the rejection of the patient's blood request/donor's blood donation, the part of the code applies, as follows:

```
    @login_re-
quired(login_url='adminlo-
gin') def update_reject_sta-
tus_view(request,pk):
req=models.BloodRe-
quest.objects.get(id=pk)
req.status="Rejected"
req.save()
       return HttpResponseRedirect('/admin-
request')
```

The part of the code for the approval of the blood donation from the donor looks like the following, where the amount of blood donated will be added to the stock upon approval by the admin:

```
    @login_re-
quired(login_url='ad-
minlogin') def ap-
prove_dona-
tion_view(request,pk):
      donation=dmodels.BloodDonate.ob-
jects.get(id=pk)     dona-
tion_blood_group=donation.bloodgroup
donation_blood_unit=donation.unit
       stock=models.Stock.objects.get(bloodgroup=dona-
tion_blood_group)     stock.unit=stock.unit+dona-
tion_blood_unit     stock.save()     donation.status='Ap-
proved'     donation.save()
      return HttpResponseRedirect('/admin-
donation')
```

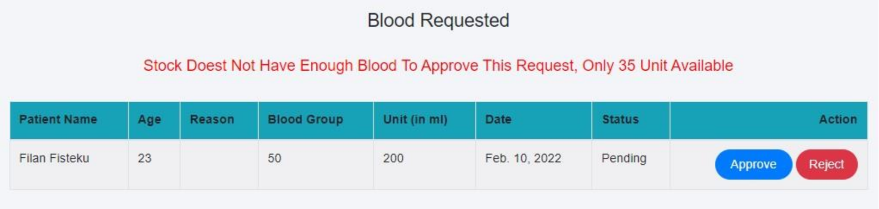
Whereas, if the request for blood from the patient is approved, if there is enough blood in stock, the request will be approved, but if there is not enough blood in stock as requested, then there is no possibility to approve that request:

```
    @login_required(login_url='adminlo-
gin') def update_approve_status_view(re-
quest,pk):     req=models.BloodRequest.ob-
jects.get(id=pk)     message=None     blood-
group=req.bloodgroup     unit=req.unit
stock=models.Stock.objects.get(blood-
group=bloodgroup)     if stock.unit > unit:

stock.unit=st
ock.unit-unit
stock.save()
req.sta-
tus="Ap-
proved"
else:
         message="Stock Doest Not Have Enough Blood To Approve This Request, Only
    "+str(stock.unit)+" Unit
Available"     req.save()
       requests=models.BloodRequest.objects.all().filter(status='Pending')     return render(re-
quest,'blood/admin_request.html',{'requests':requests,'message':message})
```

Validation

As a result of the work, after using the application, it is observed that some of the validations that have been done, which are also mentioned in the implementation part, look as follows:

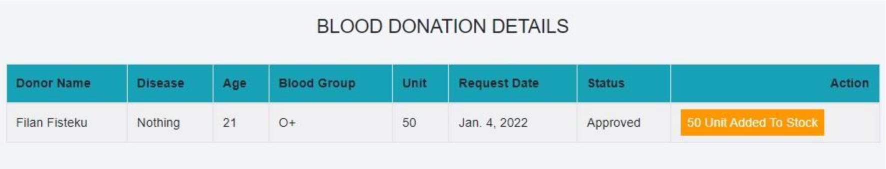


Blood Requested

Stock Doest Not Have Enough Blood To Approve This Request, Only 35 Unit Available

Patient Name	Age	Reason	Blood Group	Unit (in ml)	Date	Status	Action
Filan Fisteku	23		50	200	Feb. 10, 2022	Pending	Approve Reject

From the figure it can be seen that if the request for blood is approved, but there is not enough blood in stock as requested by the patient, then the approval of the request is not possible.



BLOOD DONATION DETAILS

Donor Name	Disease	Age	Blood Group	Unit	Request Date	Status	Action
Filan Fisteku	Nothing	21	O+	50	Jan. 4, 2022	Approved	50 Unit Added To Stock

If the blood donation is approved by the donor, then the amount of blood donated will be immediately added to the stock. Then the registration of the patient/donor cannot be done if all the fields that are necessary to was completed (required).

Related works

OOP Applications are widely used applications for managing large amounts of data and sensitive data. In our case, such applications that offer the possibility of managing blood donors have been created quite a lot in recent years. Such applications mainly include the simple management of a user's data, i.e. mainly enable the storage of personal data and these data include data such as: the fact whether the user in question has ever donated blood or not, or even vice versa, if the user in question was one of the people in need of blood transfusion. Also, the existing applications are similar to each other, however, each of them has some special functionality, but based on a study carried out about the existing applications, it was concluded that these applications have some properties that do not provide efficient service in the form of blood donation management. A large number of applications, according to the study (Ouhbi, Fernandez-Aleman, Toval, Idri, Pozo, 2015), are applications that do not require user registration, only offer the possibility of being informed whether there is blood in stocks or not (without indicating the amount of blood in stock), are applications that do not provide any notification system for a user who needs blood or will want blood when the need arises, etc. As prime examples of blood donation systems we have the blood donation systems of Red-Cross America and Red-Cross

Australia. The content and functions that these systems offer are mainly basic information about the forms of blood donation and their types. At the same time, they enable the appointment of blood donation appointments, but both in different ways. The Red-Cross America system offers the possibility of booking an appointment for blood donation based on the date or even on the basis of the destination, the possibility of displaying the user's donation history, managing the reserved appointment as well as applying for the donor card [9]. However, Red-Cross Australia, apart from offering the possibility of choosing the date of the appointment for blood donation, as well as booking the appointment based on the preferred destination, the system also allows donors/users to designate the type of blood donation (blood, plasma, etc.) [8]. So, as systems, they provide different opportunities for better organization and management of the blood transfusion process, but in some key parts they lack information and functionality.

It is worth mentioning one of the projects that has been proposed by a group of researchers in India (Srivastava, Tanwar, Krishna Rao, Manohar, Singh, May 5th, 2021). Although as a project it is still nonexistent, the idea for such an application template is quite innovative. According to this study, in contrast to our approach to the problem (which includes the possibility of saving personal data starting from the physical state of health to saving data on any existing accompanying diseases of users (specifically donors), offering the possibility of information for approval or rejection of the request for blood transfusion, offering the possibility of information about the amount of blood based on the group, etc., i.e. all this in real time), it was concluded that the best form for the realization of an application for donating blood, it is by offering the possibility of searching for blood based on the location where the user is located, so that the provision of blood is faster.

Conclusion

Seeing the great need for blood transfusion, the creation of an effective system which would enable efficiency in the search and donation of blood, is actually essential in the field of medicine. Based on the studies we have done and which we have presented and reasoned in this paper, we have come to the conclusion that a blood donation application needs some key functionalities for regular and flexible operation. On the basis of the existing systems so far, as in the case of Red-Cross America and Red-Cross Australia, we have managed to see that the functionalities offered through them are mainly registration as a blood donor, reservation of appointments for blood donation (determining time and destination (hospital center), when and where the donation will be made), management of these terms and management of the user's donation history, as well as the main information about the blood donation process is provided. In order to mark innovation in this field, based on the study of existing systems and their functionalities, in our work we managed to determine additional functionalities which would facilitate and optimize the blood donation process. Such functionalities include the possibility to register each user, including the main personal data and additional data about the physical state of health (combining diseases, cases of deterioration of the state of health in recent times, etc.) Also, the possibility of donor application is offered for blood donation, this application is then approved or rejected based on the data of the user in question, by competent persons, examining the case based on the data of the donor who made the request. Such a system, which offers the possibility of information about the amount of blood in stock, for the specification of requests for blood donation, for the possibility of information about the approval or rejection of requests for blood donation or transfusion, etc., is an approach of new and quite innovative in this field, since through the credibility and advantages that OOP offers, it will be possible to create an application that will offer security, privacy, great usability and an increase / improvement in the field in question. Also, as mentioned above, from the studies done in recent years, the possibility of examining requests for blood donation and transfusion based on the location of the user in question has emerged as an additional and essential requirement, so that the process is faster, which represents a function of vital importance, especially in emergency cases. However, adding such a function to our application remains one of the future goals, which depends on the progress of the development of the application and on the studies that will be done regarding the functionality in question, so that the requirements around it are exactly specified.

References

[1]. https://www.researchgate.net/publication/357234718_A_CrossPlatform_Blood_Donation_Application_with_a_Real-Time_Intelligent_and_Rational_Recommendation_System

[2]. https://www.researchgate.net/publication/353838299_Matching_Algorithms_for_Blood_Donation

[3]. https://www.researchgate.net/publication/342136863_BDonor_A_Geolocalised_Blood_Donor_Management_System_Using_Mobile_Crowdsourcing

[4]. https://www.researchgate.net/publication/346767243_Analysis_of_blood_supply_service_advertisemen ts_in_print_media_on_the_example_of_Sumy_regional_blood_supply_service_center

[5]. https://www.researchgate.net/publication/344930632_Analysis_Of_The_Marketing_Activities_In_The_ Blood_Service_Bibliometric_Analysis

[6]. https://www.researchgate.net/publication/273067813_Free_Blood_Donation_Mobile_Applications

[7]. https://ijcrt.org/download.php?file=IJCRT2105420.pdf

[8]. https://www.lifeblood.com.au/

[9]. https://www.redcrossblood.org/

Evaluating and comparing web scraping tools and techniques for data collection

Vesa Morina[1], Shqipe Sejdiu[2]

[1]UBT – Higher Education Institution,

Prishtine, Kosovo

vesa.morina@ubt-uni.net

[2]UBT – Higher Education Institution,

Prishtine, Kosovo

ss36902@ubt-uni.net

Abstract. The purpose of this paper is to evaluate and compare various web scraping tools for data collection. Data collection is necessary for various platforms to function, and web scraping tools offer a solution for those who want to access structured web data in an automated way. In this paper we worked and established software comparison metrics. We compare some of the most popular web scraping tools `Selenium' and 'Beautiful Soup'. Evaluation includes performance, features, reliability, and usability. Performance is measured in terms of runtime, CPU usage, and memory usage. The feature evaluation is based on implementing and completing tasks. The usability evaluation takes into account the installation process, official tutorials, and documentation. Both tools are useful and viable, but the results show that which tool performs better depends on the website.

Keywords: Data Scraping; Automation; Web Scraping Tools; Data Collection

Introduction

The use of data in decision-making has become increasingly important across various industries, and the volume of data being generated and made available on the internet is growing exponentially. However, manually collecting data from the web can be a time-consuming and resource-intensive task.

Web scraping tools offer a solution by automating the data collection process. With these tools, individuals or organizations can extract structured data from websites in a more efficient and systematic way. The use of web scraping tools is especially relevant for businesses that rely on data to drive their operations, such as marketing, research, and finance.

In this paper, we evaluate and compare two of the most popular web scraping tools, Selenium [8] and Beautiful Soup. We establish software comparison metrics based on performance, features, reliability, and usability, to help users determine which tool is best suited for their specific data collection requirements. By providing this information, we hope to assist individuals and organizations in making informed decisions about web scraping tools and the data collection process [1].

Performance Evaluation

The performance evaluation metric is a critical factor in determining the effectiveness of web scraping tools. It measures the tools' efficiency in terms of runtime, CPU usage, and memory usage. In our evaluation, we tested both Selenium and Beautiful Soup on various websites with different sizes and complexity to determine their performance.

To evaluate runtime, we measured the time it took each tool to extract data from the websites. Our results showed that Beautiful Soup outperformed Selenium in smaller and simpler websites, with faster runtime. However, as the website size and complexity increased, Selenium was able to handle dynamic content and complex JavaScript code better than Beautiful Soup. Therefore, the performance of each tool is dependent on the website's size and complexity.

CPU usage and memory usage were also crucial factors in determining the performance of the web scraping tools. High CPU and memory usage can lead to slow performance and potential crashes. Our evaluation showed that Beautiful Soup had lower CPU and memory usage compared to Selenium in smaller and simpler websites. However, as the website's size and complexity increased, Selenium's CPU and memory usage were more efficient than Beautiful Soup. Therefore, the selection of a web scraping tool should take into account the website's size and complexity to ensure optimal performance.

In conclusion, the performance evaluation of both web scraping tools, Selenium and Beautiful Soup, showed that they are both effective tools with varying performance capabilities. Beautiful Soup outperforms Selenium in smaller and simpler websites, with faster runtime and lower CPU and memory usage. However, Selenium performed better in larger and more complex websites, where it was able to handle dynamic content and complex JavaScript code better than Beautiful Soup. Therefore, users should consider the website's size and complexity when selecting a web scraping tool to ensure optimal performance [2].

Feature Evaluation

The feature evaluation metric is an important tool for assessing the capabilities of web scraping tools in implementing and completing tasks. When it comes to data collection, there are many different factors to consider, such as the type of data being collected, the source of the data, and the complexity of the data. Two of the most popular web scraping tools available are Selenium and Beautiful Soup, each of which offers unique features that cater to different types of data collection tasks.

Beautiful Soup is a well-known choice for HTML parsing, which is the process of extracting information from HTML and XML documents. It is particularly effective for extracting structured data from static HTML content. One of the key strengths of Beautiful Soup is its ability to quickly and easily navigate HTML documents and extract relevant information using a range of filtering and parsing techniques.

On the other hand, Selenium is a powerful tool for web automation tasks. It is especially adept at handling dynamic web pages with interactive content and JavaScript. Selenium can be used to interact with web pages in a way that mimics human behavior, allowing it to perform tasks like filling out forms, clicking buttons, and even logging in to websites.

Based on our evaluation, we have found that Beautiful Soup is more effective than Selenium in extracting structured data from static HTML content. However, when it comes to handling dynamic web pages with interactive content and JavaScript, Selenium is the clear winner. In other words, the choice between these two tools will depend on the specific needs of your data collection task. If you are working with static HTML content, Beautiful Soup is likely the better choice. But if you are dealing with dynamic web pages, Selenium may be the more effective option [3].

Reliability Evaluation

The reliability evaluation metric is an important aspect of assessing the stability and consistency of web scraping tools. It is crucial that the tools used for data collection are reliable, as errors or crashes can result in incomplete or inaccurate data. To determine the reliability of web scraping tools, we conducted tests on both Selenium and Beautiful Soup in different web environments and scenarios.

Our evaluation revealed that both tools are reliable, with minimal errors and crashes. However, we did note that the reliability of Selenium is often dependent on the complexity of the website being scraped. In more complex websites with dynamic content and multiple interactive elements, Selenium may encounter more issues and errors. On the other hand, Beautiful Soup's reliability can be affected by changes in website structure and HTML format. If the website's structure or format changes, Beautiful Soup may not be able to correctly parse the HTML and extract the necessary data [4].

Despite these potential issues, both Selenium and Beautiful Soup are generally reliable tools for web scraping. However, it is important to consider the specific needs of your data collection task and choose a tool that is most suited to the complexity of the website being scraped and the potential changes in website structure and HTML format. Additionally, it is important to regularly monitor the reliability of web scraping tools and make any necessary adjustments to ensure that data collection is accurate and reliable over time [5].

Usability Evaluation

The usability evaluation metric is an important aspect of assessing the ease of use and accessibility of web scraping tools. It is important that the tools used for data collection are user-friendly, with clear installation processes and comprehensive documentation. To evaluate the usability of both Selenium and Beautiful Soup, we considered factors such as the installation process, official tutorials, and documentation.

Our evaluation revealed that both tools have straightforward installation processes, with detailed official tutorials and comprehensive documentation. This makes it easier for users to install and use the tools, even if they have limited experience with web scraping. However, we did note that there is a steeper learning curve associated with using Selenium, as it requires knowledge of web automation and scripting languages such as Python. This may make it more difficult for beginners to use, but can also provide greater flexibility and customization options for more experienced users.

In contrast, Beautiful Soup is more straightforward and easy to use, with a simpler syntax and less reliance on external libraries. This makes it a good choice for users who are just starting out with web scraping or who do not have a lot of experience with scripting languages. However, it may be less suitable for more complex data collection tasks that require automation or interaction with web pages [6].

Overall, both Selenium and Beautiful Soup are usable tools for web scraping, with detailed official tutorials and comprehensive documentation. However, the choice between the two will depend on the specific needs of your data collection task and your level of experience with web automation and scripting languages. If you are just starting out with web scraping or have limited experience with programming, Beautiful Soup may be the better choice. But if you require more advanced automation and customization options, Selenium may be the more suitable tool [7].

Conclusion

After considering the various aspects of web scraping tools, it can be concluded that both Selenium and Beautiful Soup are effective and reliable solutions for automated data collection. The suitability of either tool largely depends on the size and complexity of the website being scraped and the specific data collection requirements. For smaller websites with static HTML content, Beautiful Soup is an ideal option. It is a lightweight library that can be used to parse HTML and XML documents and extract relevant data with ease. It offers a simple and intuitive interface and is easy to learn and use. Beautiful Soup also supports regular expressions and can handle malformed HTML, making it a robust solution for scraping small to medium-sized websites.

However, for larger and more complex websites with dynamic and interactive content, Selenium is a more suitable option. Selenium is a powerful web scraping tool that allows automation of browser actions, including clicking buttons, filling forms, and interacting with web elements. It can handle JavaScript-rendered pages and AJAX calls, which are common in modern websites. Selenium also supports multiple languages, making it a versatile option for web scraping.

When choosing a web scraping tool, it is crucial to consider various software comparison metrics based on performance, features, reliability, and usability. Some of these metrics include execution speed, memory usage, stability, and documentation. It is also essential to evaluate the specific data collection requirements and choose a tool that can meet those requirements effectively.

In summary, both Selenium and Beautiful Soup are reliable solutions for web scraping, but their effectiveness depends on the website's size and complexity. Beautiful Soup is best suited for smaller websites with static HTML content, while Selenium is ideal for larger and more complex websites with dynamic and interactive content. It is crucial to evaluate the software comparison metrics and data collection requirements before choosing a web scraping tool.

References

1. I. U. Khan, S. Afzal, and J. w. Lee, “Human Activity Recognition via Hybrid Deep Learning Based Model”, 2022
2. Mitchell, R. (2015). Web Scraping with Python: Collecting More Data from the Modern Web. O'Reilly Media. https://www.oreilly.com/library/view/web-scraping-with/9781491910283/
3. Jarmul, K., & Lawson, R. (2020). Python Web Scraping - Second Edition: Hands-on data scraping and crawling using PyQT, Selenium, HTML and BeautifulSoup. Packt Publishing. https://www.packtpub.com/product/python-web-scraping-second-edition/9781786462589
4. DataFlair Team. (2021). Web Scraping with Beautiful Soup: A Beginner's Guide. DataFlair. https://data-flair.training/blogs/python-web-scraping-tutorial/
5. Real Python. (n.d.). Web Scraping with Beautiful Soup: A Tutorial. https://realpython.com/beautiful-soup-web-scraper-python/
6. Morreal, J. (2020). Web Scraping with Python and BeautifulSoup. Towards Data Science. https://towardsdatascience.com/web-scraping-with-python-and-beautifulsoup-4b01facb8f45
7. DataCamp. (n.d.). Introduction to Web Scraping with BeautifulSoup. https://www.datacamp.com/community/tutorials/web-scraping-using-beautifulsoup
8. SeleniumHQ. (n.d.). Selenium WebDriver Documentation. https://www.selenium.dev/documentation/en/webdriver/

A Review on the Emerging Technologies for Air Pollution Monitoring and Management

Zhilbert Tafa

University for Business and Technology, 10000 Prishtina, Kosovo

zhilbert.tafa@ubt-uni.net

Abstract. Air pollution (AP) is one of the main causes of lung cancer and stroke. In order to minimize the negative health impacts, AP should be properly monitored and managed. Conventional systems are expensive and sparsely deployed. As such, they cannot provide the required spatiotemporal resolution. This paper reviews the emerging technologies for real-time AP monitoring based on Wireless Sensor Networks (WSNs). The review is focused on data acquisition and dissemination, as well as on the design and implementation issues. The role of Machine Learning (ML) in AP monitoring and management is also considered.

Keywords: WSN, Air Pollution, Sensors, Machine Learning.

Introduction

The development of human societies and the accompanying growth in the consumption of natural resources, especially over the past couple of centuries, has given rise to a multitude of human-induced environmental problems [1]. Air pollution is becoming one of the greatest threats to the human health. Traditionally, it has been monitored by using sparsely deployed expensive instruments. However, the AP has a much higher spatiotemporal resolution. Precisely, due to the metrological (temperature, humidity, wind, etc.) and terrain conditions, the pollutant concentration may quickly vary over a small portion of an area and may not be necessarily accumulated in the close vicinity to the AP source. Some experimental justifications can be found [2] and [3]. Consequently, data acquired from just one station or a few devices may be insufficient to describe the pollution distribution over a given region. The development of high-sensitivity multiparameter, densely deployed, and inexpensive real-time monitoring systems appear to be of crucial importance to the appropriate pollution prevention and/or management.

The advances in wireless communications and low-cost embedded systems have opened opportunities of ubiquitous computing [4]. Small smart sensors can sense, locally process, and transfer environmental data, virtually from anywhere. Being densely deployed, network can provide a more precise AP distribution over a given region. Such a georeferenced and time-stamped data can further be processed with ML techniques for AP source localization and AP distribution prediction.

In order to eliminate, decrease, or prevent the AP; the monitoring and management information systems work in the inverse feedback manner. They provide real-time sensing and data transmission, remote data visualization and analytics, AP forecast, early warning, etc. Main aspects of such a system may be summarized in: (a) data acquisition, (b) data dissemination, (c) energy management, (d) data utilization.

This article presents a short review on data acquisition and dissemination for AP monitoring. After the introduction, the target parameters and the overall architecture of an AP monitoring system are presented. The modules of a sensor node are described along with the main issues related to the design and implementation. The available wireless technologies and possible topologies are presented in the similar manner. Energy-related issues, as being an important aspect of the Wireless Sensor Networks (WSNs), are also shortly presented. In order to exemplify the aforementioned concepts, two prototype systems developed in the laboratory are presented as well. Some notes on using ML for different tasks in AP monitoring and management are given before the conclusion. Finally, last section concludes the article.

WSNs for Air Quality Monitoring

Air quality analysis involves the examination of the biological, physical, and chemical properties of the air. Although public concern has been mostly focused on urban areas, due to cooking, spraying pesticides etc., indoor areas may be also highly contaminated. Traditional air quality analysis involves periodical sensing and data reporting from the accurate and reliable instruments. These instruments are expensive (of the order of few thousand dollars) and of large physical dimensions (Fig. 1), typically occupying an area of few m^2. As such, only one or a few of them are typically used to cover an urban area (e.g., a city).

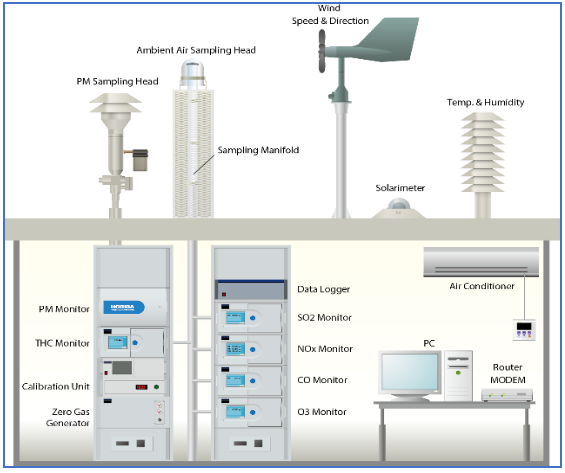


Fig. 1: A conventional AP monitoring station [5]

WSN nodes can be placed anywhere. They are low-cost (of order of few hundred dollars), power autonomous, and of small physical dimensions. As such, they can provide sensing systems of high sampling and spatial resolution. Equipped with wireless transceivers, they can report readings in real time. However, the implementation of the WSNs has limitations, such as energy issues, wireless link instability, the environmental and electromechanical influences, etc.

Air Quality Indicators

Different countries and organizations have set different standards for AP evaluation. Most of the outdoor systems are focused on measuring: carbon monoxide (CO), nitrogen dioxide ($NO_2)$, ground level ozone (O_3), ammonium (NH_4) , particulate matter (PM), sulfur dioxide ($SO_2)$, lead (Pb), temperature, and humidity. Indoors, carbon dioxide ($CO_2)$, Particulate Matter (PM), Volatile Organic Compound (VOC), temperature, and humidity, are mostly measured. In order to provide a single understandable information, so-called Air Quality Index (AQI) is extracted from the measured parameters. AQI is usually measured as worst index of separately calculated indices for each pollutant. US EPA categorizes QA in six categories, in scale from 0 to 500. In EU countries, Air Quality Framework Directive specifies AQI in range of 0-100. Some countries (e.g., Canada and UK) use 10-point scale to quantify the overall AQI etc.

The Elements of a WSN-based Systems for AP Monitoring and Management

The emerging systems for AP monitoring and management rely on WSN infrastructure. A WSN is a data acquisition and dissemination platform that enables data reporting from smart sensors to a web server and/or to data center. The information system should provide the means of data visualization in real time. Also, it should contain modules for data analytics, prediction and early warning. The workflow scheme of such a system is depicted in Fig. 2.

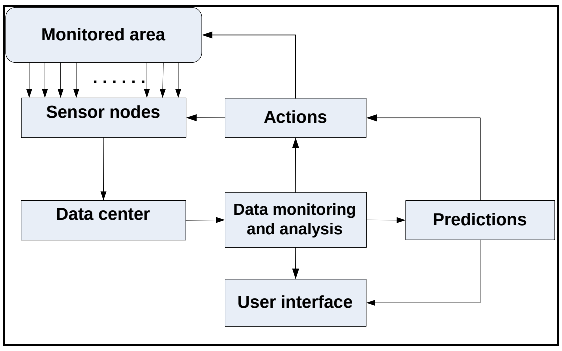


Fig. 2: An AP monitoring system contains data acquisition module, data dissemination infrastructure, data center for data visualization and utilization, user interface, and (sometimes) actuators for AP and node's management.

In order to minimize the negative influence of the AP to human health, after identifying the pollution sources, or possible future pollution hotspots, the authorities can act towards the elimination or minimization of the contaminant's concentrations or AP sources.

Data Acquisition Architecture and Issues

Main components of a sensor node are: (a) sensors, (b) signal conditioning module, (c) computing module, (d) communication module, (e) energy management module. The core elements are depicted in Fig. 3.

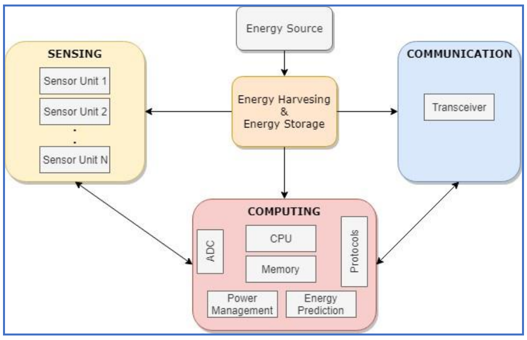


Fig. 3: Signals are conditioned, sampled and digitized. Then, they are forwarded to the communication module for wireless transmission. Energy management module is focused on the energy optimization.

The existing AP low-cost monitoring nodes are either based on off-the-shelf components or are delivered as fully integrated plug-and-play solutions. The last ones are easier to integrate and have robust electro-mechanical design, but lack in flexibility (Fig. 4).

Fig. 4: Libelium Air Quality station [6] – an example of integrated low-cost sensor for AP monitoring. This module also enables for the online calibration by using artificial intelligence.

Continuous AP reporting based on low-cost sensors has several constrains. Most of them are related to sensors' performances. Precisely, sensors come with somewhat unknown and unpredictable settings. Prior to their integration into AP measurement system, they need to be evaluated in terms of accuracy, selectivity, sensitivity, and precision. The evaluation should be followed with the calibration process. Because of the aging effects, they should also be periodically rechecked on the aforementioned metrics and recalibrated.

The sensor performances vary from sensors to sensors. Also, some of them might show good performances regarding specific parameters, but they may be not as accurate to measure other parameters. For instance, for

NO_2, data acquired from Alphasense [7] show high R^2 correlation with data obtained from standard instruments. On the other hand, the correlation is not satisfying regarding the O_3 parameter.

Low-cost sensors for AP monitoring are highly sensitive to the operating conditions and impose aging drift in time. Their readings are influenced by wind, temperature, humidity, etc. They are also influenced by external or internal noise, such as mobility and low sensitivity, respectively.

The electromechanical gas sensors consume a considerable amount of energy. As compared to the humidity or temperature sensors, they are much greater energy consumer. This questions the nodes' autonomy in some applications. The energysaving schemes often include lowering the duty-cycle. However, turning off the gas sensor for a longer time is not suitable, because they need some time to heat-up before they become operable again.

Regardless of the aforementioned limitations, a study performed in [8] shows that, if the reason of using low-cost sensors is not to measure the absolute concentration of AP values, but to indicate the quality of the atmospheric environment through different health impact levels (such as AQI), then low-cost sensor devices may successfully fit this purpose.

The radio-characteristics of a wireless communication module has direct impact to the network coverage and topology, but also directly influence the node's lifetime. For instance, as compared to ZigBee technology, 3G/4G modules can provide much broader coverage, but they consume much more energy.

In contrast to gas sensors and wireless transceivers, modern microcontrollers have very low power consumption. Some of the open-source electronic prototyping platforms integrate high performance low-power microcontrollers (ATmega, PIC, MSP430 etc) to provide design's flexibility and modularity.

Energy management module manages power consumption and power supply of a node. Software-triggered routines enable dynamic duty-cycling as well as dynamic sampling and transmission rate adjustment. Also, when communication is achieved via multi-hop transmissions, energy-aware MAC (Media Access Control) and routing algorithms are often required to minimize power consumption. Finally, energy management modules may encompass some energy-harvesting techniques, such photovoltaic, wind, vibration, etc.

Wireless transmission

With power-autonomous nodes, the design of a data dissemination topology and the selection of a most suitable wireless technology is of a crucial importance. Precisely, without considering other (heavy) consumers such as electrochemical gas sensors, a wireless transceiver typically consumes around 70% of the node's energy. The choice of wireless technology and network topology impact the optimal balance between the network coverage, energy efficiency, and network performance. For instance, an "energy-efficient" wireless routing protocol aims to route via the most energyefficient path (instead via the shortest path). These protocols might impose long endto-end delays or may fail to find a route [9].

In most of the emerging AP monitoring applications, smart sensors can be powersupplied from the existing power distribution systems, i.e., AP readings can be sent from the positions where some kind of stabile energy source is available. For instance, nodes can be installed on traffic lights, the roof of the buildings, vehicles, etc.

If energy consumption is not a system's limitation, AP systems may use some of the long-range technologies (3G, 4G, GPRS, etc.) to transfer data to the data center. If nodes are energy-constrained, there are two near-optimal sub-scenarios. If a local gateway can be supplied from the power distribution system, then data are locally transferred (to the gateway) via some low-power short-range technology (such as Zigbee, Bluetooth, 6LoWPAN, etc.). The gateway may then use some WAN technology (e.g., 4G, or even cable modem) to transfer data to the data center. If the power distribution system is far from the points of installation, some of the point-topoint Low Power WAN technologies (such as LoRa, NB-IoT, or Sigfox) may be used.

Wireless infrastructure for AP data dissemination can be based on: Static Sensor Networks (SSNs) [10,11], Community Sensor Networks (CSNs) [12,13], Vehicle Sensor Networks (VSNs) [14,15], or combination of the aforementioned [16]. A comprehensive review and comparison of the wireless architectures and topologies for AP monitoring is presented in [17].

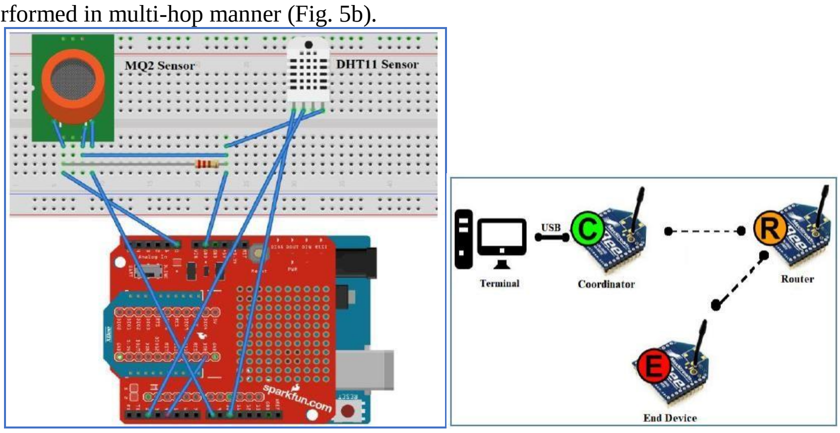
Examples - WSNs for AP monitoring

In the proceeding subsections, two experimental studies on the implementation of WSNs for AP monitoring are shortly presented. The first one was designed for indoor implementation, while the second one has been implemented in an urban area.

A Multi-Hop Indoor AP Visualization System

According to the Environmental Protection Agency (EPA), indoors, air may be 2-5 times more polluted than outdoors. An indoor off-the-shelf system for AP monitoring, developed in UBT laboratory, is presented in [18]. The data acquisition module is based on Arduino Uno platform, which utilizes Atmel ATmega328P microcontroller and interfaces a MQ2 gas sensor, DHT temperature/humidity sensor, and a ZigBee transceiver

(Fig. 5a). In order to cover a larger indoor area, the data dissemination is performed in multi-hop manner (Fig. 5b).



(a) (b)

Fig. 5: A system for indoor air quality monitoring. (a) sensor node, (b) multi-hop wireless network.

Although the proposed system provides a satisfying accuracy for the measurement of LPG, CO, and smoke, it shows some of the general aforementioned weaknesses related to the implementation of low-cost gas sensors. For instance, current drain in active mode is ≈140mA. With a specific duty cycle, system can achieve power autonomy of a few weeks. The system should be thoroughly tested and improved in accuracy.

A VSN-based AP Monitoring System

A number of VSN-based systems for AQM have been proposed in literature. They are either implemented as standalone VSNs or combined with SSNs. Different vehicles have been used to carry nodes, and different technologies have been used to sample, process, and transmit data. Some examples are given in [19, 20].

An experimental study has been conducted in the area of Prizren, Kosovo [21]. The MQ gas, temperature and humidity sensors are interfaced to the Arduino Uno board and are attached on roof of the Taxi vehicles (Fig. 6). Data are sent to the server via GPRS technology and are stamped in space and time with GPS. The system combines SSNs and VSNs to collect data for real-time visualization, historical view, and further analysis. The nodes are powered-up from the vehicle's power supply system. Hence, the system is not power constrained. Micro SD card module is installed on the board to store samples when the GPRS connection is unavailable.

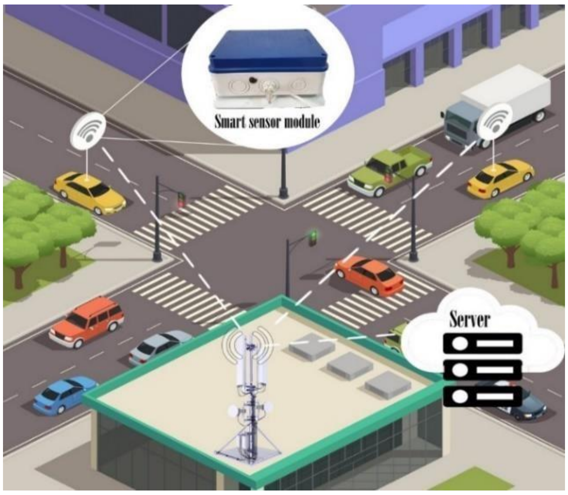
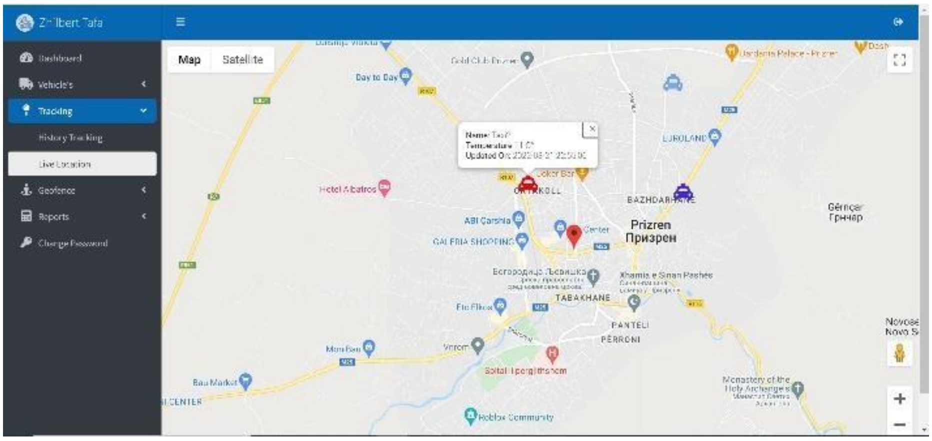


Fig. 6: Sensor are installed at the roof of the taxi vehicles and are power supplied from the vehicles' battery. The readings are combined with those obtained from a conventional AP monitoring station in the region of Prizren, Kosovo.

A server receives data and stores them in a database for further web-based visualization and data analysis. Google API and JQuerry library are used for map and value entries visualization (Fig.7). The interface also provides the preview of the historical data (Fig. 8)



Name: Taxi7
Temperature: 11 C°
Updated On: 2022-08-31 22:50:00

Fig. 7: User-friendlyGUI for real-time AP vizualisation

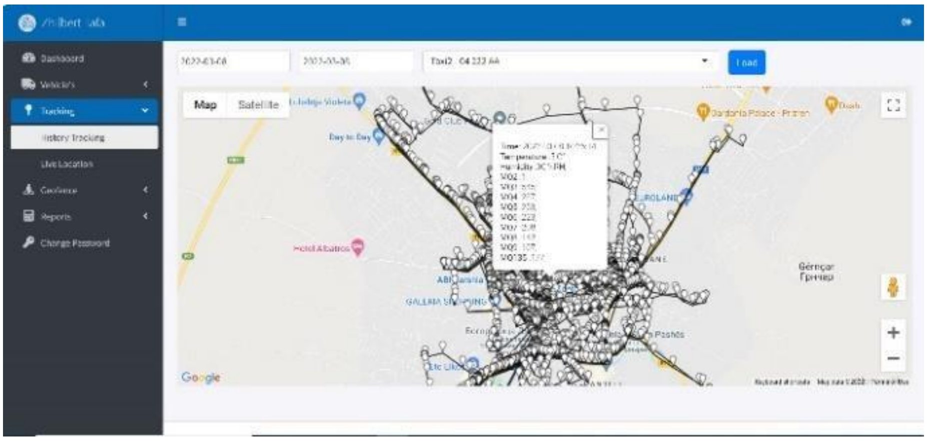


Fig. 8: Historical view of AP parameters

The system shows satisfying performances in terms of overall functionality and continuity, with a small number of transmission errors. It provides a satisfying spatiotemporal resolution for the region of Prizren, of an area of cca. 640 km^2. Future work will cover the implementation of ML for accuracy improvement and AP forecasting.

ML algorithms in Emerging AP Monitoring Systems

Calibration consists of setting a mathematical model to describe the relationship between low-cost sensor data and reference measurements [22]. Recently, calibration has been mostly performed by using statistical and ML approaches, such as multilinear regression, Support Vector Machine (SVM), Artificial Neural Networks (ANN) etc. [23,24]. The ML techniques take into account the influences of the temperature, humidity, air pressure, solar radiation etc. on the deviations of readings.

Machine learning techniques have been also widely investigated for anomaly detection [25], AP estimation [26], and AP prediction [27,28]. The follow-up of the study presented in [21] will include the aforementioned extensions. Future work will also combine sparse data acquired from fixed and mobile nodes for ML-based estimation of the AP for every point on the map.

Conclusions

WSNs have a potential to fill the gap between AP data acquisition requirements and the respective economically feasible technological solutions. Although some issues (as presented in this article) have slowed down the wider implementation of these technologies, recent extensive research show that WSNs can successfully improve spatiotemporal sensing resolution at a relatively low cost of implementation.

Moreover, ongoing research show that implementation of the ML techniques can greatly contribute to the improvement and optimization of the WSN-based AP monitoring and management information systems.

References

1. V. Evagelopolous et. al, "Cloud-Based Decision Support System for Air Quality Management," *Climate*, 10(3), 39, 2022. doi: 10.3390/cli10030039
2. G. Gualtieri et al., "An integrated low-cost traffic and air pollution monitoring platform to assess vehicles' air quality impact in urban areas, "*Transportation Res. Proc.*, vol. 27, pp. 609-606, 2017. doi: 10.1016/j.trpro.2017.12.043
3. A. M. Popoola et al., "Use of networks of low-cost air quality sensors to quantify air quality in urban settings," *Atmospheric Environment*, vol. 194, pp. 58-70, 2018. doi: 10.1016/j.atmosenv.2018.09.030
4. Z. Tafa, "Ubiquitous Sensor Networks," in *App. and Multidisciplinary Aspects of Wireless Sensor Networks*, L. Gavrilovska, et al., Eds., ed: *Springer London*, 2011, pp. 267-268. doi: 10.1007/978-184996-510-1_13
5. Online: https://www.horiba.com/pol/scientific/products/detail/action/show/Product/aqms-1560/ (accessed on Sept, 2022)
6. Online: https://www.libelium.com/iot-products/air-quality-station/ (accessed on Sept, 2022)
7. Online: https://www.alphasense.com/ (accessed on Sept, 2022)
8. G. C. Spyropoulos, P. T. Nastos, and K. P. Moustris, "Performance of Low-Cost Sensors for Air Pollution Measurements in Urban Environments. Accuracy Evaluation Applying the Air Quality Index (AQI)," *Atmosphere*, 12(10), 1246, 2021. doi: 10.3390/atmos12101246
9. Z. Tafa, "WSNs in environmental monitoring: Data acquisition and dissemination aspects." *Advances in Computers*, vol. 126, 2022. doi: 10.1016/bs.adcom.2021.11.010
10. D. Yaswanth & S. Umar, "A study on pollution monitoring system in wireless sensor networks," *International Journal of Computer Science & Engineering Technology*, vol. 3, pp. 324-328, 2013.
11. P. J. A. John, "Wireless Air Quality and Emission Monitoring," *Master Thesis*, Uppsala University, 2016.
12. Available online: https://airqualityegg.com/home (accessed on Sept, 2022)
13. M. Nyarku et al. "Mobile phones as monitors of personal exposure to air pollution: Is this the future?" *Plos ONE*, vol. 13, no. 2, pp. 1-18, 2018.
14. A. Anjomshoaa, B. Duarte, D. Rennings, T. Matarazzo, P. de Souza, and C. Ratti, "City Scanner: building and scheduling a mobile sensing platform for smart city services, " *IEEE Internet of things journal*, pp. 2327-4662, 2018.

15. S. Kaivonen & E. C.-H. Ngai, “Real-time air pollution monitoring with sensors on city bus,” *Digital Communications and Networks,* vol. 6, iss. 1, pp. 23-30, 2020.
16. S. Kaivonen and E.C.H. Ngai,” Real-Time Air Pollution Monitoring with Sensors on City Bus,” *Digital Communications and Networks*, vol. 6, pp. 23-30, 2020.
17. W.Y.Yi, K..M.Lo, T.Mak, K.S.Leung, Y.Leung, and M.L.Meng, “A Survey of wireless sensor network based air pollution monitoring systems,” *Sensors*, vol. 15, no. 12, 2015.
18. Z. Tafa &F. Cakolli, "Design of a Multi-Hop Wireless Network to Continuous Indoor Air Quality Monitoring." *Journal of Communications* 14.5 (2019).
19. A. Anjomshoaa, B. Duarte, D. Rennings, T. Matarazzo, P. de Souza, C. Ratti, “City Scanner: Building and Scheduling a Mobile Sensing Platform for Smart City Services,” *IEEE IoT Journal*, pp. 2327– 4662, 2018.
20. O. Saukh, D. Hasenfratz, A. Noori, T. Ulrich, and L. Thiele, “Demo Abstract: Route selection of mobile sensors for air quality monitoring,” *in: 9th European Conference on Wireless Sensor Networks*, pp. 10–11, 2012.
21. B. Zherka & Z. Tafa, “A Vehicle Sensor Network for Real-Time Air Pollution Analysis,” *Accepted, to appear in Journal of Advanced Information Technology*, 2022.
22. F. Karagulian, M. Gerboles, M. Barbiere, A. Kotsev, F. Lagler, and A. Borowiak, “Review of sensors for air quality monitoring,” EUR 29826 EN, Publications Office of the European Union, Luxembourg, 2019.
23. C. -C. Chen *et al.*, "Calibration of Low-Cost Particle Sensors by Using Machine-Learning Method," *2018 IEEE Asia Pacific Conf. on Circuits and Systems*, 2018, pp. 111-114.
24. D. Topalovic et al, “In Search of an Optimal Calibration Method of Low-Cost Gas Sensors for Ambient Air Pollutants: Comparison of Linear, Multilinear and Artificial Neural Network Approaches,” *Atmospheric Environment*, vol. 213, pp. 640-658, 2019.
25. D. -D. Dajic & N. -R. Grigoric, “Reliable Low-Cost Air Quality Monitoring Using Off-The-Shelf Sensors and Statistical Calibration,” Elektronika ir Elektrotechnika, vol. 26, no. 2, pp. 32-41, 2020.
26. K. Hu, A. Rahman, H. Bhrugubanda and V. Sivaraman, "HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation from Fixed and Mobile Sensors," in *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3517-3525, 2017.
27. I. U. Samee, M. T. Jilani and H. G. A. Wahab, "An Application of IoT and Machine Learning to Air Pollution Monitoring in Smart Cities," *2019 4th International Conference on Emerging Trends in Engineering, Sciences and Technology (ICEEST)*, 2019, pp. 1-6.
28. K. Kumar & B. -P. Pande, “Air Pollution Prediction Eith Machine Learning: A Case Study of Indian Cities,” *International Journal of Environmental Science and Technology*, 2022.

Using Machine Learning Tools to Study the Unemployment and Output Relationship in Albania

Blerina VIKA[1], Ilir VIKA[2], Kozeta SEVRANI[3]

[1, 3]Department of Statistics and Applied Informatics, Faculty of Economy, University of Tirana, Albania

[2]Research Department, Bank of Albania

blerina.vika@unitir.edu.al; ivika@bankofalbania.org, kozeta.sevrani@unitir.edu.al;

Abstract

The field of economics is nowadays increasingly employing artificial intelligence to complement and further improve the tools for analyzing and making future decisions. One of the most common machine-learning techniques that are used in forecasting economic indicators is the recurrent neural network method, which has often proven to be useful in capturing non-linearities in data series. This article applies the long short-term memory (LSTM) technique to test for the Okun relationship in Albania. Apart from examining movements in unemployment as predicted by developments in aggregate demand, we test whether a disaggregated version of the Okun's law – by decomposing aggregate demand into various expenditure components of GDP – provides better predictions for changes in unemployment. In-sample estimations suggest that the Okun's law may hold in Albania, but the response of unemployment to output performance is found to vary over different time periods. Non-linear model forecast evaluations show that unemployment rate movements in the second half of 2010s could be more related to private investment and government spending, while private consumption and external trade developments seem to predict it less.

Keywords: Machine Learning, LSTM model, unemployment, Okun relationship, Albania.

Disclaim: This research article is considered as a preliminary work that aims at stimulating debate and critical comments. It expresses the views of the authors and does not necessarily represent those of the institutions where they work. The authors are grateful to comments and suggestions received from participants at the "11th International Conference on Business, Technology and Innovation 2022", organized by the University for Business and Technology - UBT, Pristina, Kosovo, in October 29-30, 2022.

1. Introduction

There is a general consensus today that inequitable economic growth might have adverse effects in the society. For that reason, governments around the globe have increased their awareness on the importance of inclusive growth and are pushing their political agenda to make sure that economic growth could deliver more jobs and promote higher living standards for their citizens. Emerging and developing economies have been growing rapidly for a number of decades. Yet, recent empirical findings reveal that their satisfactory performance has been in the large part not necessarily translated into more employment and equal opportunities.

This paper uses the Okun relationship to examine movements in unemployment as predicted by developments in aggregate demand in Albania. In addition, it test whether a disaggregated version of the Okun's law – by decomposing aggregate demand into various expenditure components of GDP – provides better predictions for changes in unemployment. The hypotheses are tested by using the econometric least square method as well as a rather novel technique in this regard by employing the long short-term memory (LSTM) model. The latter is a machine-learning technique that has often proven to be useful in capturing non-linearities in economic data series.

To preview the results, the OLS estimations appear to be in favor of the Okun relationship in Albania. Estimating different sub-samples suggests that the response of unemployment to output could be time-varying, while faster domestic growth may be needed to keep unemployment rate from rising. On the other hand, the preliminary results in this ongoing project show that adding output growth to the LSTM forecast model does not improve the unemployment predictions that are generated from a univariate model. Nevertheless, the LSTM model predictions with GDP components reveal that unemployment could be more related to private investment and government spending, and perhaps not so much linked with private consumption and external trade.

The remainder of the article is organized as follows. In section 2 we provide a glimpse at the empirical literature on unemployment and growth in advanced and developing countries. Section 3 estimates the Okun coefficient in a simple OLS regression to examine the contemporaneous relation of unemployment with aggregate demand in Albania. In addition, to shed light on the stability of their relationship, the Okun coefficient is estimated for i) the early transition years, ii) the faster growth period, and iii) the post-global crisis decade. Section 4, then, reassesses the relevance of economic growth in the Okun relationship by relying on the information embedded in GDP and its expenditure components to improve upon the benchmark univariate forecasts of the unemployment rate. Moreover, as the in-sample estimations in section 3 evidenced a time-varying Okun coefficient, the uncommon cross-checking analysis in section 4 is based on the nonlinear LSTM model. The concluding remarks are available in section 5.

2. A brief literature review

Despite some theoretical limitations of Okun's law, Prachowny (1993) and Blinder (1997) conclude that Okun relationship is a useful principle of macroeconomics in which 'we should all believe'. In his seminal paper on the relationship between GDP developments and changes in unemployment, Okun (1962) originally treated unemployment as the exogenous variable while output as the dependent variable. However, the rich empirical literature estimating the Okun relationship has been mostly interested in the reverse direction, hence assuming the output growth as the driving force of unemployment.

The link between unemployment and output is generally assessed by transforming the variables in gaps or in differences. The "Gap" version relates the divergence of current unemployment (U_t) fromt its "natural" rate (U*) with the deviation of actual output (Y_t) from its potential level (Y*) [$U_t - U^* = \beta(Y_t - Y^*) + \varepsilon_t$]. On the other side, "Difference" version relates changes in unemployment rate to economic growth [$\Delta U_t = \alpha + \beta \Delta Y_t + \varepsilon_t$], thus assuming that the "natural" rate of unemployment is zero [$\Delta U^*=0$] while potential GDP growth is constant [$\Delta Y^*=g$].

The original Okun's paper estimated that β equaled negative 0.3 in the U.S. during the 1947-60 period. However, the recent empirical literature on the Okun's law has found heterogeneous evidence with respect to the size of the Okun coefficient. The difference among countries might have much to do with the very definition of unemployment, as the term does not follow a worldwide standard definition of measuring employment, nor can data availability be comparable between formal and shadow economies. Also, labor market rigidities and government social protection policies on employment have often been emphasized when trying to explain the heterogeneous sensitivity of unemployment to output decline, particularly among advanced economies [see e.g. IMF, 2010; Cazes, Verick, & Hussami, 2013].

A review by Pizzo (2019) shows Okun's coefficients in rich economies as estimated with the difference (gap) version are around -0.29 (-0.39), while in developing and emerging market economies they are -0.18 (-0.20). Other studies have similarly concluded that unemployment is less sensitive to production in developing than in advanced economies [see e.g., Ball et al., 2019; An et al., 2017; Bartolucci et al., 2018]. Using a sample of 176 countries for the period 1993 until 2015, Farole et al. (2017) evidence that Okun's coefficient is not constant across countries. For every percentage point increase in GDP, unemployment rates would decline on average for the high income, upper-middle income, lower middle-income, lower income countries by 0.21, 0.08, 0.03, and 0.005 pp, respectively.

Furthermore, reassessments after the Global Crisis period indicate that Okun coefficients all over the world have declined when comparing the pre-crisis (1992-2007) with the post-crisis (2010-2017) period (Lee et al., 2020). Yet, Farole et al. (2017) find that in transition economies like Ukraine and Croatia unemployment seems to react asymmetrically over the business cycle: the size of coefficients show relatively strong reaction during downturns, but much weaker upside response during periods of upward growth.

Using only the "difference" version and running pooled OLS for ILO data sets on unemployment, Lee et al. (2020) find that during the 1992-2017 period the Okun's coefficients are -0.22 in twenty nine NSW European countries and -0.15 in ten Eastern European economies. Anyhow, Farole et al. (2017) find a large difference among similar income countries in CE Europe. For example, the Okun coefficient is estimated 0.85 in Poland, but just 0.03 in Hungary. In Albania, Garo (2020) finds that "for a 1 p.p. drop in the output growth, the annual change in the unemployment rate increases about 0.184 p.p. over 4 consecutive quarters".

3. Does Okun's law hold in Albania? An OLS approach with aggregate demand

The link between unemployment and output is typically examined by assessing their contemporaneous relationship:

$$\Delta U_t = \alpha + \beta_y \Delta Y_t + \varepsilon_t \qquad (1)$$

where ΔU denotes annual change in unemployment rate; ΔY is annual growth rate of output; α is an intercept coefficient that captures the trend growth in unemployment rate ($\alpha = -\beta\Delta Y^*$); β is the so-called "Okun coefficient", denoting the sensitivity of unemployment rate to GDP growth, which is expected to be negative, since a higher GDP growth rate should lead to lower unemployment rate; ε is an error term.

We test for the Okun relationship using quarterly data from 1996q1 to 2019q4. As Albanian economy has experienced significant fluctuations during this period, it may be possible that the correspondence between variables be both nonlinear and time-varying. Thus, we test for changes in the Okun relationship during the i) early transition period of 1996q1-2003q4; ii) fast economic growth from 2003q1-2010q4; and the post-global financial crisis decade from 2011q1-2019q4.

Table 1. Okun coefficients for Albania: aggregate relationship

Table 1. Okun coefficients for Albania: aggregate relationship

	(1) Full Sample 1996q1:2019q4		(2) Early Transition 1996q1:2003q4		(3) Fast Growth 2003q1:2010q4		(4) Post-Global Crisis 2011q1:2019q4	
Dependent Var. ΔU	Coef.	Prob.	Coef.	Prob.	Coef.	Prob.	Coef.	Prob.
ΔY (*βy*)	-0.090***	0.00	-0.075*	0.10	-0.212***	0.00	-0.581***	0.00
Constant (*α*)	0.236	0.23	0.436	0.28	0.820***	0.00	1.182***	0.01
Adj. R-squared	0.07		0.06		0.54		0.32	
Included observations	96		32		32		36	
Ratio of -*α/β*	2.6		5.8		3.9		2.0	

Table 1 displays the estimated Okun coefficient for the aggregate relationship in Albania. The estimated response of unemployment to output growth is statistically significant and has a negative sign as expected, hence suggesting a well-functioning labor market. The Okun coefficient, β, is, however, not constant and appears to have considerably increased in magnitude over time. The relationship is estimated to be modest in the full sample (-0.09). The result is in line with the findings from Farole et al. (2017) for middle-income countries, which are often characterized by insufficient safety nets and relatively large share of self-employed workers. However, the relatively low reaction seems to be influenced by the early transition period. Later on, the size of β coefficient increases significantly and implies a relatively significant job creation in the post-GFC period (-0.58). But, so has the trend growth in unemployment rate as captured by the intercept parameter, α; in the last period in column (4) is about 5 times higher than for the full sample, hence well above the assumption of zero changes in the "natural" rate of unemployment. Finally, the time-varying parameters are reflected in a declining ratio of model coefficients ($-\alpha/\beta$) – found to be statistically significant since 2003 – which may imply that the Albanian economy needs to grow faster in order to leave the unemployment rate unchanged.

4. Decomposing the Okun's law: do expenditure components enhance prediction ability?

Methodology

The vast empirical literature on the Okun relationship evidences that economists widely believe on the existence of a long-run relationship between GDP performance and unemployment. Yet, the short-run dynamics may disguise the long-run bond between the variables. The previous section revealed a rather time-varying Okun relationship in Albania, suggesting a lack of stability in the coefficients over the past two and a half decades.

To shed more light on the suitability of Okun's law as a "rule of thumb" for predicting unemployment rate, certain studies have resorted to analyses that decompose the Okun relationship into expenditure or production components. We follow them by decomposing here the aggregate demand into private consumption (con), public expenditure (gov), private investment (inv), exports (exp) and imports (imp).

Furthermore, the mainstream studies have commonly relied on traditional econometric models to estimate and test the stability of the Okun coefficient. In contrast, we follow a novel approach to test for the Okun relationship by relying on unemployment predictions generated by nonlinear nonparametric Machine Learning (ML) methods. These innovative techniques could be suitable in resolving possible issues of nonlinearity and time-varying relations in our variables, due to changing domestic and external economic conditions. Among the wide class of these techniques, we have selected to apply in our analysis the long short-term memory (LSTM) model. This method belongs to the artificial recurrent neural network family, which have gained popularity in the area of artificial intelligence and deep learning. Moreover, machine learning methods are nowadays being seen as a new exciting 'car' on the street by many central bank economists, who are increasingly experimenting with ML techniques as additional supporting tools to help improve economic forecasts and monetary policy decisions. Unlike the feedforward neural network that we have experimented before (e.g., Vika & Vika, 2021), the recurrent neural network LSTM has feedback connections and is capable of learning long-term dependencies. Once we train and test the LSTM network structure, we use the out-of-sample forecasts computed by it to assess the relevance of economic growth in the Okun relationship by relying on the information embedded in GDP and its expenditure components to improve upon the benchmark univariate forecasts of the unemployment rate.

Data and forecasting procedure

All variables are seasonally adjusted by using the X-13 ARIMA-SEATS procedure in EViews software. The whole sample period covers quarterly data from 1996Q1 to 2019Q4. Each model includes three lags as inputs of each variable, as suggested by Akaike and HQ info criteria in the estimated least squared regressions. Forecast evaluation is undertaken for out-of-sample forecasts, using the RMSE measure for a forecast horizon of 4 quarters. In our analysis, the 1996Q1:2019Q4 period has been divided into the so-called training period 1996Q1:2014Q4 (76 quarters) and the forecast evaluation period stretching over the 2015Q1:2019Q4 sample (20 quarters).

Consequently, the model starts for the period 1996Q1:2014Q4. For each specification the forecast and its corresponding RMSE is noted down for the 4 quarters ahead. The training and testing period is recursively extended by one quarter (1996Q1:2015Q1), and similarly calculating and retaining the RMSE of forecasts for the desired one-year ahead horizon. The evaluation process is repeated 16 times for the 4 quarters horizon until we predict the last quarter of 2019.

Forecast evaluation results

Table 2 displays the forecast gains/losses in using GDP growth and its expenditure components to predict changes in the unemployment rate. A ratio above 1 indicates underperformance of the bivariate LSTM networks. It turns out that including GDP in the unemployment LSTM model does not outperform the latter's forecasts that are based on its own past values. The ratio of RMSE is just above 1, casting doubt on the predictive content of output growth on unemployment rate. However, the evaluation of loss differentials for individual expenditure components reveals some interesting results on the effectiveness of this approach with the aggregate relationship. The RMSE ratios for government spending (0.84) and private investment (0.81) suggest us that these expenditure components could be the key factors in delivering jobs and relieving unemployment in Albania. On the other hand, private consumption (1.15) and external trade (1.07) seem to have no predictive power on the rate of unemployment.

Table 2. The loss differential between forecasts, 2015q1:2019q4, h=4 quarters

Table 2. The loss differential between forecasts, 2015q1:2019q4, h=4 quarters

Bivariate Models:	Δu-Δy	Δu-Δcon	Δu-Δgov	Δu-Δinv	Δu-Δexp	ΔU-Δimp
Ratio of average RMSE: (Bivariate/Univariate f'casts)	1.01	1.15	0.84	0.81	1.07	1.07

5. Concluding remarks

The results from OLS estimations seem to be in favor of the Okun relationship in Albania. The coefficient is found to be rather low for the past two and a half decades, in line with other studies for low and middle income countries. Estimating different sub-samples shows considerable lack of constancy in the coefficients that measure the sensitivity of unemployment to output as well as the trend growth of unemployment. Yet, faster domestic growth may be needed to keep unemployment rate from rising. Moreover, LSTM forecasts indicate that unemployment could be more related to private investment and government spending, while private consumption and external trade developments do not seem to improve upon the univariate forecast accuracy.

While social demographic changes, informality and emigration behavior may limit the ability to increase the precision of measuring unemployment, inadequate safety nets and the relatively high level of self-employed persons in the country might contribute to the underestimation of output-unemployment relationship.

Future research may need to focus on alternative approaches to assess the Okun relationship with various model specifications and lag length selections, along with variable transformation such as other ratios of labor market indicators or the "gaps" version. The latter, for instance, could be invaluable in taking into account possible changes in the "natural" rate of unemployment and potential output growth, as the Albanian economy strives to catch-up with other higher per capita income countries.

References

An, Z., Ghazi, T. & Prieto, N.G. (2017) "Growth and jobs in developing economies: trends and cycles" IMF Working Paper WP/17/257.

Ball, L., Furceri, D., Leigh, D., & Loungani, P. (2019) "Does one law fit all? Cross-country evidence on Okun's law" Open Economies Review 30, 841–874 (2019).

Bartolucci, F., Choudhry, M.T., Marelli, E. & Signorelli, M. (2018) "GDP dynamics and unemployment changes in developed and developing countries" Applied Economics, 50 (31), 3338-3356.

Blinder, Alan S. (1997) "Is there a core of practical macroeconomics that we should all believe?" The American Economic Review, vol. 87, no. 2, 1997, pp. 240-243.

Cazes, S., Verick, S. & Hussami, F.A. (2013) "Why did unemployment respond so differently to the global financial crisis across countries? Insights from Okun's law" IZA Journal of Labor Policy, Springer; Forschungsinstitut zur Zukunft der Arbeit GmbH, vol. 2 (1), pp 1-18, December.

Farole, T., Ferro, E. & Gutierrez, V.M. (2017) "Job creation in the private sector: an exploratory assessment of patterns and determinants at the macro, sector, and firm levels" Jobs Working Paper, no. 5. World Bank, Washington D.C.

Garo, O. (2020) "Inquiring into the relationship between the unemployment rate and output growth", Paper presented at Bank of Albania, 14th South-Eastern European Economic Research Workshop, Tirana, 10-11 December 2020.

IMF (2010) "World Economic Outlook. Chapter 3"

Lee, S., Schmidt-Klau, D., Weiss, J., Chacaltana, J. (2020) "Does economic growth deliver jobs? Revisiting Okun's Law" ILO Working Paper 17, November / 2020.

Okun, A. (1962) "Potential GDP: its measurement and significance" Proceedings of the Business and Economic Statistics Section of the American Statistical Association.

Pizzo, A. (2019) "Literature review of empirical studies on Okun's law in Latin America and the Caribbean" ILO Employment Policy Department Working Paper No. 252, 2019.

Prachowny, M.J.F. (1993) "Okun's law: Theoretical foundations and revised estimates" The Review of Economics and Statistics, vol. 75, no. 2, 1993, pp. 331-336.

Vika, B., Vika, I. (2021) "Forecasting Albanian time series with linear and nonlinear univariate models" Academic Journal of Interdisciplinary Studies, Richtmann Publishing, vol. 10, no. 5, September 2021.

Using polynomials over the GF(2) field for detecting and correcting errors in cyclic codes

Diellza Berisha[1*], Bukurie Imeri Jusufi[2], Mirlinda Reqica[3], Blinera Zekaj[4] and Azir Jusufi[5]

[1]UBT Higher Education, Prishtinë, Kosovë

{diellza.berisha, bukuri.imeri, mirlinda.reqica, blinera.zekaj, azir.jusufi}@ubt-uni.net

Abstract. The developments of the last decades in the field of digital communication have created a close connection between mathematics and computer engineering fields. The Galois field GF(2)={0,1} is of great use in Computer Science, along with the polynomials with coefficients from the field GF(2). If we denote by V(n,q) the vector space over the field GF(q), then the linear binary code C[n,k] is nothing more than a subspace of the vector space V(n,q). During the transfer of word codes through channels with obstacles of various natures, errors may also occur, which must be detected and corrected. Cyclic codes are a n important group of linear binary codes. They are widely used in the theory of codes, as they are easily applied and in particular their polynomial form. In this paper we will give the algorithm for detecting and correcting errors that may occur in the cyclic code.

Keywords: Vector space, linear code, cyclic code, word code, detection, polynomials.

1. Introduction

In order to develop the theory for general linear codes, we need some definitions from abstract and linear algebra.

The binary alphabet GF(2)={0,1}, along with modulo 2 addition and multiplication, is the smallest example of a field.

Defnition 1.1. Let *F* be a field. A set *V* of elements called vectors is a *vector space* if for any u,v,w*V* and for any *c,d F*:

1. u + v*V*
2. *c*v *V*
3. u + v = v + u.
4. u + (v + w) = (u + v) + w
5. 0*V* such that u + 0 = u.
6. u*V* , -uV, u + (-u) = 0.
7. *c*(u + v) = *c*u = *c*v.
8. (*c* + *d*)u = *c*u + *d*u
9. *c*(*d*u) = (*cd*)u.
10. 1u = u

We often say that *F* is the field of *scalars*. An important example in coding theory is the vector space of all binary *n*-tuples. Here, the vectors are the binary *n*-tuples and the field is *GF*(2)={0,1}. We say that 0 and 1 are the scalars

Note V (n,q) denotes the vectorial space of vectors with length n on the field GF {q}.

Example 1.1. Let's mark it V(3,2)={000,100,010,001,110,101,011,111} a vector space over GF {2}.

We will use the vector subspaces more often than the above formal definition of a vector space.

Defnition 1.2. Let V be a vector space. A subset U of V is called a subspace of V if U is itself a vector space under the same operations as V.

Theorem 1.1 ([1].p.148) Let V be a vector space and F be a field. A subset U of V is a subspace if:

1. ∀ u,v ∈U, u+v ∈U; 2. c ∈F, u∈U, cu∈U

The vectors $u_1, u_2, \ldots, u_n$ of a vector space *V* are said to be *linearly dependent*, if there exist a *finite* number scalars c1, c2, …,cn, not all zero, such that $c_1u_1 + c_2u_2 + \ldots + c_nu_n = 0$. where 0 denotes the zero vector.

The vectors $u_1, u_2, \ldots, u_n$ of a vector space *V* are said to be *linearly independent* if the equation $c_1u_1 + c_2u_2 + \ldots + c_nu_n = 0$ can only be satisfied ci=0 for *i*=1,2,…,n.

Definition 1.3. A *basis* is a (minimal cardinality) set of linearly independent vectors that span the vector space. The *dimension* of a vector space *V* is number of vectors in any basis for *V*.

Definition 1.4. Let *V* be a vector space over a field *F* and let *W* be a subset of *V*. If *W* is a vector space over *F*, then *W* is a *vector subspace* of *V*.

Theorem 1.2. ([2], p.196) If *W* is a subset of a vector space *V* over a field *F*, then *W* is a vector subspace of *V* if ***au* + *bv*** *W* for all *a,b F* and all *u,v W*.

2. Binary Linear Codes

Definition 2.1. A *binary linear code C* of length *n* is a set of binary *n*-tuples such that the component wise modulo 2 sum of any two wordcodes is contained in *C*

Example 2.1.

Let'sV(3,2)={000,100,010,001,110,101,011,111} a vector space over GF {2}.

The set C(3,2)={000,100,001,101} is a subset of the V(3,2), we see that 000+000=000 C, 000+100=100 C, 000+001=001 C, 000+101=101 C,

100+100=000 C, 100+001=101 C, 100+101=001 C, 001+001=000 C, 001+101=100 C, 101+101=000 C.

so C is the binary linear code. Also C completes the conditions of the subspaces.

Assertion 2.1. The linear space V(n, 2) over the field GF(2) is a linear binary code of length n

Proof. From Example 2.1. V(n, 2) is the set of vectors (codewords) that are variations of class n of the set GF(2) ={0; 1}, which forms the Galois field

(GF(2); +; ·), where the extension + is defined by Table 5.1. and the multiplication · is defined by Table 5.2. That example highlights the fact that (V(n, 2), +, ·), where + is the addition (mod 2) of the vectors in V(n, 2) and · is the multiplication (mod 2) of the vectors with the elements 0, 1 of GF(2), is a linear space over the field GF(2).

But adding (mod 2) to V(n, 2) as an internal operation means that the sum (mod 2) of any two vectors from V(n, 2) is again a vector in V(n, 2). This shows, by Definition 2.1, that the linear space V(n, 2) over the field GF(2) is a linear binary code with its codeword vectors.

Assertion 2.2([8], pg.265). Every linear binary code of length n is a linear subspace of the linear space V(n, 2) over the Galois field GF(2).

Flow 2.1. ([8], p. 265). Every linear binary code is a linear space over the Galois field GF(2).

Definition 2.3. The *Hamming weight w*(c) of a word code c is the number of nonzero components in the word code.

Example 2.2. *w*(0000) = 0, *w*(111) = 3, *w*(1011001) = 4

Definition 2.4. The *Hamming distance* between two wordcodes *d*(x*;* y) is the number of places in which the word codes x and y difference. In other words, *d*(x*;* y) is the Hamming weight of the vector x -y, representing the component-wise difference of the vectors x and y.

Definition 2.5. The *minimum (Hamming) distance* of a code *C* is the minimum distance between any two wordcodes in the code:

d(C) = min {d(x; y) | x y; x, y C}.

To easily calculate the distance of code C, we get

Theorem 2.1.([6], p.6) The minimum distance, *d*(*C*), of a linear code *C* is equal to *w**(*c*), the weight of the lowest-weight nonzero wordcode.
Proof. There exist wordcodes x and y in *C* such that d(C) = d(x y). By the definition of Hamming distance, we can rewrite this as d(C) = w(x-y). Note that x-y is a word code in *C* by the linearity of *C*. Therefore, since *w**(*c*) is the weight of the lowest weight wordcode, we have w*(c) w(x-y) =d(C).
On the other hand, there exists some wordcode c *C* such that *w**(*c*) = *w*(c). By the definition of the weight of a word code, we can write *w*(c) = *d*(c,0). Since c and 0 are both wordcodes, the distance between them must be greater than or equal to the minimum distance of the code: *d*(c, 0) *d*(*C*). Stringing together these (in)equalities, shows that *w**(*C*) *d*(*C*). Since we have now shown that both *d*(*C*) *w**(*c*) and *d*(*C*) *w**(*c*), we conclude that

$$d(C) = w^*(c).$$

Example 2.4. The minimum distance, *d*(*C*), of a linear code *C*={000,100,011,111} is
d(*C*) =*w**(*100*) =1

3. CYCLIC CODES. POLYNOMIAL REPRESENTATION

Let it be **c=c0 c1 … cn-1** a code word of length n. Switching from it to a different codeword c(i) = **cn-i cn-i+1 … cn-1 c0 c1 … cn-i-1** formed by the same coordinates is called right cyclic shift with i **(0<i<n)** positions of code word c or left cyclic shift with its n-i positions

Definition 3.1. A linear [*n, k*]-code C_n is called cyclic if for each of its codewords $c = c_0 c_1 ... c_{n-2} c_{n-1}$ its right shift by one position $c^{(1)} = c_{n-1} c_0 c_1 ... c_{n-2}$ is also in C_n .

For example the binary linear code C_4 = {000, 110, 011, 101} is easily seen to be a cyclic linear code.

Since *V*(*n*, *q*) is a linear code and contains all codewords of length *n*, this turns out to be true

Flow 3.1. For every n>0 the linear space *V*(*n*, *q*) over the field *F*=*GF*(*q*) is a cyclic code.

The convenience of using cyclic codes comes from their connection to the ring of one-variable polynomials *F[x]* over the Galois field *F=GF(q)*. It is known that the set *F[x]* of polynomials with one variable, where F is a field, forms a commutative-associative, complete (without zero divisor) and unitary ring, i.e an integral domain. Then the terminology and results applicable to such a ring can be used for it.

As in the ring Z of integers we introduce in the ring *F[x]* the meaning of congruent polynomials according to a modulus polynomial. This is why we examine the community

I_n={q(x)(x^n+1)q(x) F[x]}, where n>0.

Let it be c=c0 c1 … cn-1 a codeword of the [*n, k*]-cyclic code C_n. The expression

c(*x*)= c0+ c1x+ …+ cn-1xn-1 ,

which is a polynomial *c*(*x*) ∈ *F*[*x*] with multiple degree *n*-1, we call it the *polynomial representation of the code word c*.

We denote by C_n [*x*] the set of polynomial representations C_n [*x*] of the codewords

c ∈ C_n

C_n[x]={c(x) cC_n}.

The polynomial representation of the cyclic shift to the right with *i* positions of the codeword *c*

$c^{(i)}$ =cn-i cn-i+1 … cn-1 c0 c1 … cn-i-1 , (1)

we denote it as $c^{(i)}$(*x*):

cix=cn-i+cn-i+1x+. . . +cn-i-1xn-1. (2)

Example 3.1. The [3, 2]-binary linear code C_3={000,110,011,101} is cyclic, because the right shift by one position of codeword 000 is 000, of codeword 110 is 011, of codeword 011 is 110, and of codeword 101 is 110

In this cyclic code, the polynomial $x+x^2$ ∈ *F*[*x*] is the polynomial representation of the codeword 011 C_3, because $x+x^2 = 0+x+x^2$, while 1+*x* ∈ *F*[*x*] is the polynomial representation of the codeword 110 C_3, because $1+x=1+x+0x^2$.

Even from the above example, it appears that a polynomial of degree *n*-1 with coefficients coordinates of a codeword from the cyclic code C_n is a polynomial representation of that codeword, while not every polynomial representation *c*(*x*) is a polynomial of degree *n*-1 from the ring *F*[*x*], as it follows from this one

Assertion 3.1. If a polynomial

p(*x*) = p0+ p1x+…+ pr-1xr-1+prxr ϵ *F*[*x*]

is of degree r<n-1, then it can be written in the form

px=p0+ p1x+…+ pr-1xr-1+prxr+pr+1xr+1+…+pn-1xn-1 , (3)

where pr+1,…., pn-1=0

***Proof**:* The form *p*(*x*) = p0+ p1x+…+ pr-1xr-1+prxr is its usual form, if the coefficient of pr the oldest limit x^r is not zero, indicating that the scale is *r*. Considering the condition that r<n-1, we add to the polynomial *p*(*x*)*F*[*x*] the limits 0xr+1·····0xn-1 which give it the form (30):

px=p0+ p1x+…+ pr-1xr-1+prxr+pr+1xr+1+…+pn-1xn-1, where pr+1+…+pn-1=0

From what was said above, view (30), when the degree of *p*(*x*) is less than n-1, is not the usual form of the polynomial *p*(*x*) *F*[*x*]. We call the view (30) its *completed representation*

Flow 3.1. If the polynomial p(x) F[x] is the remainder of division by a polynomial g(x) F[x] with degree n, then its completed representation has the form

px=p0+ p1x+p1x2+…+ pn-1xn-1 . (4)

Proof. Under these conditions, the degree of the residual p(x) is multiple n-1. When it is exactly n-1, the usual form of the polynomial p(x) is the completed representation (4). When its degree is less than n-1, according to Assertion, it is given the completed representation (3), so (4).

Flow 3.2. A polynomial *p*(*x*) = p0+ p1x+…+ pr-1xr-1+prxr ϵ *F*[*x*] when the degree r=n-1, it is a polynomial representation in C_n[*x*], while when the degree r<n-1, it returns to the polynomial representation

p(*x*) = p0+ p1x+…+ pr-1xr-1+prxr+pr+1xr+1+…+pn-1xn-1

in C_n[x], if its coefficients p0,p1…. pr-1 pr,…,pr+1,…, pn-1, where pr+1,…., pn-1=0 , are coordinates of a codeword of the cyclic code C_n.

Proof: The case when r=n-1 is evident. In the case when the polynomial has degree r<n-1, according to the Assertion, we give it the complete representation (30).If the coefficients

p0, p1…. pr-1, pr,…,pr+1,…, pn-1, the coordinates of the code word

p0,p1…. pr-1 pr,…,pr+1,…, pn-1, of the cyclic code C_n, then it is the polynomial representation of this code word.

Remarks. Under the conditions of this statement, it is noted that the polynomial representation of the cyclic shift to the right with *i* positions of the code word

p0, p1…. pr-1 pr,…,pr+1,…, pn-1, ϵ Cn will have the form

p(i)(x)= pn-i,pn-i+1 x++pn-i-1 x-1, but as a polynomial from *F*[*x*] its degree *r* does not change, although its limits have a new setting.

Theorem 3.1 ([8], page 287). If C_n is a cyclic binary code, then ∀ cxCn[x],

xicx≡ $c^{(i)}$(x) (0<i<n).

Theorem 3.2([8], pg.287). A [*n, k*]-linear binary code C_n s cyclic if and only if ∀ cxCn[x] and ∀ rxF[x] of degree i<n , the congruent of *r*(*x*)*c*(*x*) to be in C_n[*x*].

Theorem 3.3. If a monic polynomial of the ring $F[x]$ with degree $r < n$ is transformed into the polynomial representation of a code word of the cyclic binary code C_n, then its free term is equal to 1.

Proof. Let $g(x)=g_0 + g_1x+...+ g_{r-1}x^{r-1}+x^r$ be such a polynomial from $F[x]$, which, being monic, has the coefficient next to x^r as 1. As in the remark of above, it is transformed into the polynomial representation

$$g(x)=g_0+g_1x+...+g_{r-1}x^{r-1}+x^r+g_{r+1}x^{r+1}+...+g_{n-1}x^{n-1}$$

of the word code $g = g_0 ,g_1 ..., g_{r-1},g_{r+1}\ ...\ g_{n-1}\ C_n$, where $g_{r+1}, ..., g_{n-1}=0$, with coordinates in $F=\{0, 1\}$.

According to that remark, the polynomial representation of the n-1 position cyclic right shift of this codeword will have the form

$$g^{(n-1)}(x)=g_1+g_2x+...+g_{r-1}x^{r-2}+x^{r-1}+g_{r+1}x^r+...+g_{n-1}x^{n-2}+g_0x^{n-1},$$

or as a polynomial from $F[x]$, its degree r does not change. From the fact that $g_{r+1}, ..., g_{n-1}=0$, this polynomial form returns to the polynomial

$$g1+g2x+...+g_{r-1}x^{r-2}+x^{r-1}+ g_0x^{n-1} = g(x).$$

Assume now that g0=0 Then:

gx=g1x+. . . + gr-1xr-1+xr

=x(g1+g2x+. . . + gr-1xr-2+xr-1) = $x\ g(x)$.

This is a wrong equation because on its sides we have two polynomials from F[x] with different degrees. It remains that g0=1.

Definition 3.2. A monic polynomial g(x)∈ F[x] with degree r<n, which transforms into a polynomial representation of a codeword of [n, k]-cyclic binary code C_n is called its generating polynomial.

Usually the generating polynomial g(x) of a [n, k]-cyclic binary code is taken of degree r=n-k, where the coefficients $g_0=g_r=1$.

4. STANDARD PROCEDURE OF CODING IN CYCLIC CODES

The procedure of encoding a message into a $[n, k]$-cyclic code C_n is based on the so-called message polynomial and a generating polynomial of C_n. We are stopping at the so-called *standard coding procedure.*
Let the message $m = m_0\ m_1\ ...m_{k-1}$ be a codeword in the code $V(k, 2)$ and $g(x)$ of degree n-k a generating polynomial of the $[n, k]$-cyclic binary code C_n. We construct the polynomial $m(x)=m_0+m_1x+...+m_sx^s \in F$[x] with its complete representation

$m(x)=m_0+m_1x+...+m_{k-1}x^{k-1}$,   (5) when degree $s<k$-1.

We call the completed representation (5) the *polynomial of the message m*. In the ring *F*[x] there is a single pair of polynomials *q*(*x*), *p*(*x*), which are the quotient and the remainder of the division of the polynomial
$x^{n\text{-}k}m(x)$= m0xn-k+m1xn-k+1+. . . + msxn-k+s (6)
with the generating polynomial *g*(*x*), that is $x^{n\text{-}k}m(x)=q(x)\ g(x)+\ p(x)$, where the remainder
$p(x)=+\ x+...+\ x^r$ has the degree $r < n$-k. We write the remainder in the completed representation (4):
$p(x)=+x+...+x^{r\text{-}1}+x^r+...+x^{n\text{-}k\text{-}1}$, ku ,...,=0. (7)

Theorem 4.1. ([8], pg. 289). For each codeword $m_0\ m_1\ ...m_{k\text{-}1}$ V(k, 2), under the above conditions,
cx=p0+p1x+. . . + pn-k-1xn-k-1+m0xn-k+xn-k+1+. . . + mk-1xn-1, (8)
where, ...,=0, is the polynomial representation of the code word
c=...... C_n. (9)

The code (9) is called the standard code related to the message code m, while the polynomial representation (8) is called its standard representation.

The standard password c associated with the message m = $m_0\ m_1\ ...m_{k\text{-}1}$ V(k, 2) is obtained by this **algorithm**, which is called the standard procedure of encoding a message in a cyclic code:
The polynomial xn-km(x) is divided by the generating polynomial gx and we get
xn-kmx=qxgx+p(x) px+ xn-kmx=qxgx,
where p(x), when the scale is less than n-k-1, is completed with additional limits with 0 coefficients to obtain the completed representation (7).
From the left side of the last equation, we move to the polynomial representation
cx=p0+p1x+. . . + pn-k-1xn-k-1+m0xn-k+ m1xn-k+1+. . . + mk-1xn-1,

of the standard codeword c =p0+p1+…+pn-k-1 m0 m1… mk-1 ∈ Cn, related to the message m.
From the Theorem this follows immediately
Flow 4.1. If c= c0 c1cn-k-1 cn-k cn-k+1 ..cn-1 is a codeword of the [*n*, *k*]-cyclic code C_n, generated by the polynomial *g*(*x*), then it serves as a standard codeword, associated with the message *m*= cn-k cn-k+1 ..cn-1 ... ∈ *V*(*k*, 2), while c0 c1cn-k-1 is the vector of the coefficients of the completed representation of the remainder of the division of the polynomial $x^{n\text{-}k}m(x)$ by the generating polynomial *g*(*x*).

Example 4.1. For the cyclic code [7,4], generated by the polynomial 1+x+x3, find the standard codeword related to the message m=1010 .
Solution. The message polynomial is mx=1+x2 and n-k=7-4=3, so xn-k*m*(*x*)= x3mx=x3+x5.

- We divide the polynomial x3mx=x3+x5 by the generating polynomial 1+x+x3:

(x5+x3) : (x3+x+1)= x2

x5+x3+x2
x2

We take
x3+x5=x2 x3+x+1+x2x2+x3+x5=x2 x3+x+1.
x2+x3+x5
cx=x2 +x3+x5=0+0x+x2 +x3+0x4+x5
c=0011010.

Below is a table, where in the first column there are all the message codewords from *V*(4, 2), dhe and in the second column all the standard codewords related to them are placed in the
[7, 4]-cyclic code, generated from the polynomial *g*(*x*)= 1+x+x3 .

5. Generation and control matrix of a cyclic code.

We will now show how a cyclic[*n*, *k*]-code can be constructed, when its generating polynomial of degree*n-k* is known.
Theorem 5.1. ([8], pg. 291). If $g(x)= g_0+g_1x+ g_2x^2+...+g_{r-1}x^{n-k}F[x]$ is the generating polynomial of [*n*, *k*]-cyclic binary code C_n, then the matrix

G =

is the generating matrix of the code C_n.

The corresponding control matrix for the C_n code has the form

=

5.1 Algorithm of the standard generating matrix of [*n, k*]-cyclic code

Perform the divisions xn-k+j:gx and, after finding the remainders pjx,*j*=0,1,...,*k*-1, write them in their completed representation pjx=pj0+pj1x+. . . +pj,n-k-1xn-k-1.

- We find the code words cj=pj0 pj1 . . . pj,n-k-1 0…010….0 ∈ Cn, *j*=0, 1, ..., *k*-1.

The standard generating matrix G_{kxn} of the [*n, k*]-cyclic code is constructed.

Construct the control matrix H(n-k)xn for the [*n, k*]-cyclic code, if required.

Definition 5.1.1. Control polynomial for the [*n, k*]-cyclic code C_n, we call the polynomialh(x), such that gxhx=xn-1, where g(x) is the generating polynomial of degree r=n-k of the code C_n.
Since xn-1 is monic, then the generating polynomial $g(x)$ of the cyclic code and the control polynomial hx for that code are monic. When the generator has the scale $r = n - k$, the control one has the scale k=n-r.

Example 5.1.1. It is directly proved that the polynomial x7+1**,** factorizes as follows
x7+1=1+x1+x+x3(1+x2+x3),
where each of the factors and the products of each two of them are monics with degree less than 7, therefore they serve as generating polynomials of cyclic codes with length n=7.

1. Find the size *k* of each [7, *k*]-code with the generating polynomial of each of them.
2. Find the standard generating matrix of that cyclic code of length 7, which has the generating polynomial polynomial 1+x.
3. Find the standard generating matrix of that cyclic code of length 7, which has the generating polynomial polynomial $1+x^2+x^3$.
4. Find the generating matrix of that cyclic code of length 7, which has the generating polynomial polynomial $1+x+x^2+x^3+x^4+x^5+x^6$.

Solution. 1. The products of any two of the factors from the given decomposition
in the factor of the polynomial x7+1 are:

1+x1+x+x3 = $1+x+x^3+x+x^2+x^4=1+x^2+x^3+x^4$;
1+x(1+x2+x3) =$1+x^2+x^3+x+x^3+x^4=1+x+x^2+x^4$;
1+x+x31+x2+x3 = $1+x^2+x^3+x+x^3+x^4+x^3+x^5+x^6$
= $1+x+x^2+x^3+x^4+x^5+x^6$.

Consequently, from that polynomial, as its factors, these polynomials with degrees smaller than7: 1+x; 1+x+x3; 1+x2+x3; $1+x^2+x^3+x^4$; $1+x+x^2+x^4$; $1+x+x^2+x^3+x^4+x^5+x^6$.

Polynomial 1+x , where r=1, is the generating polynomial of [7, *k*]-code with size k=7-1=6; the polynomials 1+x+x3 and 1+x2+x3, where r=3, are generating polynomials of [7, *k*]- code with size k=7-3=4; the polynomials $1+x^2+x^3+x^4$ and $1+x+x^2+x^4$, where r=4, are generating polynomials of [7, k]- code with size k=7-4=3 and polynomial $1+x+x^2+x^3+x^4+x^5+x^6$, where r=6, is the generating polynomial of the [7, k]- code with size k=7-6=1.

2. The cyclic code generated by the polynomial g1x=1+x, according to Theorem 10.12.1, has the generating matrix*:*

G_{6x7}=.

The standard generating matrix of this cyclic code, considering that it is a linear space, can be found, avoiding the algorithm, by summing the rows of the G_{6x7} matrix.

This is justified by the fact that its lines are codewords of that code, therefore the corresponding sums are also codewords in it. Adding the first row to the second, the sum obtained to the third and so on, we get this generating matrix of standard form:

$G1_{6 \text{ x } 7}$ = ,

from which it can be seen that in each line the sum of bits 1 is an even number.

3.The polynomial g2x=1+x2+x3 is the generating polynomial of a cyclic code of size 4. To find its standard generating matrix, we apply the algorithm, where n=7, k=4, n-k=3, j=0. 1, 2, 3.

We perform the divisions:

We find the word codes from the completed representations of the residues

=1 0 1 1 0 0 0 ; =1 1 1 0 1 0 0 ; =1 1 0 0 0 1 0 ; =0 1 1 0 0 0 1 .

We construct the standard generating matrix $G_{4\text{x}7}$ of [7, 4]-cyclic code:

$G_{4\text{x}7}$=

We note that in each row of the $G_{4\text{x}7}$ matrix there are at least 3 units. Since $n=2^3-1$ and n-k=3, so r=3, then this [7, 4]-cyclic code is an H_3 Hamming code.

If the problem also required the control matrix for this Hamming code, it would be:

=**.**

4.Finally, the generating polynomial $g_3(x)=1+x+x^2+x^3+x^4+x^5+x^6$ is of [7, 1]-cyclic code,which has generating matrix of type , because k=7-6=1. Its coefficients are$g_0=g_1=g_2=...=g_6=1$, therefore, according to Theorem 5.1, this code has only one generating matrix: =.

6. COMBINATION OF THE POLYNOMIAL SYNDROME. POLYNOMIAL DETECTION. RESULTS

1. Calculating the polynomial syndrome from the generating polynomial

Let $f = f_0 \ f_1 , ..., f_{n-1} \ V(n, 2)$ be the code word expected from the transmission of the message code word c by the $[n, k]$-cyclic code C_n with generating polynomial $g(x)$. Then its polynomial representation will be
$f(x) = f_0 + f_1 x + ... + f_{n-1} x^{n-1}$,
where the corresponding polynomial $f(x)F[x]$ with branch $f(x)<n$. We divide the polynomial $f(x)$ by the generating polynomial $g(x)$ with branch $g(x)=n-k$. We take

$$f(x) = q(x)g(x)+s(x), \qquad (10)$$

where branch $q(x)<k$ and branch $s(x)<n-k$. According to Theorem 3.2, if the remainder $s(x)=0$, then $f(x) = q(x)g(x) \ C_n[x]$, which brings $f \ C_n$, so f is acceptable, therefore no errors occurred; if $s(x) \ 0$, then $f(x) \ C_n[x]$, which yields fC_n, so f is not acceptable, therefore errors occurred in waiting for the vector f . In this way, an indicator of the presence of the code word and the transmission error is found, which is exactly the residue $s(x)$, therefore, the syndrome $S(f)=s(x)$.

2. Polynomial Detection of acceptable codewords

When $s(x)=0$, there is no correction, so it is assumed that f is decoded to obtain the message. For this we take into account Flow 4.1. When $s(x)0$, the expected vector f is corrected, removing (in addition to having the field $F=\{0, 1\}$) $F=\{0, 1\}$) the error e, such that $f+e=c \ \ C_n$. To find the error we are based on (10):
$f(x) = q(x)g(x) + s(x) \qquad f(x) + s(x) = q(x)g(x) \ C_n[x]$, where $f(x)$ and $s(x)$ are the completed representations of the polynomials $f(x)$, $s(x) \ F[x]$, so not only $f(x) = f_0 + f_1 x + ... + f_{n-1} x^{n-1}$, but also $s(x) = s_0 + s_1 x + ... + s_{n-1} x^{n-1}$. Then $e = s_0 \ s_1 \ ... \ s_{n-1}$. In this way, the corrected $f+e \ \ C_n$. codeword is accepted for decoding

The polynomial decoding algorithm

We form the polynomial representation $f(x) = f_0 + f_1 x + ... + f_{n-1} x^{n-1}$ of the expected word code $f= f_0 \ f_1 , ..., f_{n-1} \ V(n, 2)$.

We divide the polynomial $f(x)$ by the generating polynomial $g(x)$ with branch $g(x)=n-k$: $f(x) = q(x)g(x)+s(x)$. From here we find the syndrome $S(f)=s(x)$.
If $S(f)= s(x)=0$, it is accepted to decode f;
if $S(f)=s(x)$ 0, its complete representation $s(x)=s_0 + s_1 x+...+ s_{n-1} x^{n-1}$, is formed, from which we find the error vector $e = s_0\ s_1\ ...\ s_{n-1}$. Corrected codeword f +e $\in C_n$ is accepted for decoding..

Example 6.2.1 ([9].pg.260) From [7,4]-cyclic code C_7 with generating polynomial $g(x)=1+x+x^3$ the message code word c was started, which was received by the code word
$f = 1\ 0\ 1\ 1\ 0\ 1\ 1V(7, 2)$.

1. Find acceptable code words in C_7.
2. Determine the message that comes out of its decoding.

*Solution.*1. We use the polynomial decoding algorithm.
The polynomial representation of the expected code word is
$f(x)=1+0x+x^2+x^3+0x^4+x^5+x^6$.
We do the division

$$f(x)=(x^3+x^2+x+1)(x^3+x+1)+ x^2S(f)= s(x)=x^2.$$

$S(f) = x^2$ 0 $s(x)=0+0x+1x^2+0x^3+0x^4+0x^5+0x^6$
$e=0\ 0\ 1\ 0\ 0\ 0\ 0f +e=1\ 0\ 1\ 1\ 0\ 1\ 1+0\ 0\ 1\ 0\ 0\ 0\ 0=1\ 0\ 0\ 1\ 0\ 1\ 1C_7$. Then the latest is the codeword acceptable for decoding.
2.According to Flow 4.1, acceptable password $c=1\ 0\ 0\ 1\ 0\ 1\ 1$ is started from C_7 as a codeword containing the message $m = m_0\ m_1\ m_2\ m_3$ $V(4, 2)$, therefore it has the form $c=p_0\ p_1\ p_2\ m_0\ m_1\ m_2\ m_3$ $m= 1\ 0\ 1\ 1$.

References

1. Bashkim Gazidede, “Algjebra 2”, SH.B Ilar, Tiranë, 2006May, P., Ehrlich, H.-C., Steinke,
2. David Cherney, Tom Denton, Rohit Thomas and Andrew Waldron, “Linear Algebra” First Edition. Davis California, 2013.
3. R. Hill, A flrst course in coding theory, Oxford University Press, New York, flrst ed., 1986
4. Steven J. Leon, “Linear Algebra with Applications”, Eighth Edition, University of Massachusetts, Dartmouth
5. Sarah Spence Adams, “Introduction to Algebraic Coding Theory” , 2008
6. Yehuda Lindell, “Introduction to Coding Theory Lecture Notes”, Department of Computer Science Bar-Ilan University, Israel January 25, 2010
7. E. R.: Algebraic Coding Theory. New York: McGraw-Hill, 1968
8. A. Jusufi, K.Filipi, Matematike Diskrete dhe Aplikime, UBT, Prishtine, 2022

9. A. Jusufi, D.Berisha,M.Reqica, Permbledhje detyrash nga matematika diskrete, UBT, Prishtine, 2022

Katalogimi në botim – (CIP)
Biblioteka Kombëtare e Kosovës "Pjetër Bogdani"

519.68 (062)
061.3(496.51)

Proceedings International Conference : proceedings of the 11th UBT Annual International Conference : international Conference on Computer Science and Communication Engineering / edited by Edmond Hajrizi. - Prishtinë : UBT, 2022. - 92 f. : ilustr. ; 30 cm.

1**.** Hajrizi, Edmond

ISBN 978-9951-550-78-9

INTERNATIONAL CONFERENCE ON BUSINESS,TECHNOLOGY AND INNOVATION

CHAPTERS:

- Computer Science and Communication Engineering
- Management, Business and Economics
- Mechatronics, System Engineering and Robotics
- Energy Efficiency Engineering
- Information Systems and Security
- Architecture – Spatial Planning
- Civil Engineering, Infrastructure and Environment
- Law
- Political Science
- Journalism, Media and Communication
- Food Science and Technology
- Pharmaceutical and Natural Sciences
- Design
- Psychology
- Education and Development
- Fashion
- Music
- Art and Digital Media
- Dentistry
- Medicine & Nursing

Lagjja Kalabria p.n KS - 10000, Prishtinë
+383 38 541 400
+383 38 542 138

www.ubt-uni.net
conferences@ubt-uni.net

ISBN 978-9951-550-78-9
9 789951 550789