BULGARIAN ACADEMY OF SCIENCES INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Ivan Ivanov Blagoev

METHODS AND MEANS OF DATA ANALYSIS IN INFORMATION SYSTEMS USING TIME SERIES

A B S T R A C T

Of thesis for awarding educational and scientific degree PhD

Scientific Specialty: „Informatics“

Professional area: „4.6. Computer sciences“

Scientific supervisor: Assoc. Prof. Tatiana Atanasova, PhD

Sofia, 2021

The thesis contains 125 pages, 33 figures, 1 table and 122 bibliography sources.

Introduction

Advances in technology are so obvious that it can only be mentioned without the need for a factual description. In this respect, a significant difference from recent times is the highly expansive digital transformation. Due to COVID-19, the threat to human health, the speed of technology entering our lives has accelerated greatly, leading to a total change in many activities, and in the coming years will be even more noticeable as humanity transforms and adapts to this new way of life.

All that has been mentioned so far leads with it and very completely new to science and processes not previously researched. The collection and processing of time series and big data will be expanded by infiltrating new processes. The need for research and new discoveries will be crucial for the development of science and technology in the coming years. For this development of new methods and tools for time series research and big data processing, it is extremely important and will be a major tool for research and development of science and technology in the future.

This dissertation, through time-series research, contributes to achieving better results in methods of forecasting financial instruments, processing big data and improving cryptography and cyber security.

Purpose and tasks of the dissertation

The purpose of this dissertation is to develop new methods and means of data analysis in information systems using timelines.

For this purpose, the following tasks are defined:

1. develop a method for analyzing and predicting price movements in the financial field using time series;
2. to propose an algorithm for the training of artificial neural networks in forecasting financial time lines;
3. propose solutions to increase cryptographic protection in information systems by applying methods of analysis of time lines;
4. to conduct experimental studies to verify the proposed methods of enhancing cryptographic protection in solving cyber security tasks.

5 develop programmatic methods to overcome problems when working with big data in time series.

Structure of the dissertation

Thesis work is structured in four chapters.

Chapter 1 provides an overview of the current themes in the field of data science, especially when these data are presented as time series.

Chapter 2 presents the developed methods for researching and forecasting financial time series using different mathematical apparatus.

Chapter 3 also describes the developed solutions to provide cryptographic protection in the provision of information services by examining random number generators representing sequences of time series. The practical application of the proposed approaches to cyber security is presented. The actual results of the tests carried out, demonstrating the successful resolution of the tasks assigned, are shown.

Chapter 4 overcomes loading and processing a big data problems in limited computer resources when examining time series with the means of programming language R.

A summary of the results of the development has been presented in the Conclusion.

The dissertation work contains 125 pages, 33 figures, 1 table and 1 22 bibliography sources.

Chapter 1. Analysis of the state of the tests.

If you look at the world through the eye of technology, the first thing that would impress any specialist is how much data that is. This is a side effect of mass digital transformation and automation (Wang, 2020), leaving a digital trace of the real process. Time series in communications, technology, business come as a result of measuring characteristics from technical, natural, social, economic and other systems (Mikalef,2020), (Ciampi,2020).

1.1 Time series

The time series is a representing row of data collected at equal or uneven time intervals. A key feature of the time series is that each subsequent value is depending on

the previous values. This dependency can be both very complex and relatively simple. Currently, many forecasting methods that act as effective tools are widely accepted for evaluating and analyzing data from time line models. Of these, the most commonly used model is an integrated method of seasonal component auto-aggression (SARIMA - Seasonal ARIMA), which essentially belongs to a linear model.

1.2. Applying time series to financial instruments

Market price movements are described through time lines and are subject to analysis by finance strategists, economists and market strategists. Types of financial analyses currently used to analyze financial instruments:

- Fundamental - based on analyzing events happening around the world and concerning financial and commodity markets (Wafi, 2015);
- The technical analysis - based primarily on statistical methods with time series calculations - is based on statistical methods. Allows of forecasting to be described by statistical means and mathematical algorithms (Plummer, 1991), (Scott, 2016).

1.2.1 Neural networks

Approaches for examining time series can be divided into two categories: statistical methods and computational intelligence. Statistical methods investigate dependencies between baseline and relevant factors after studying past data, while the other group of methods mimics the human way of thinking and logical conclusion in order to gain knowledge from past experience (such as artificial neural networks) and to predict future values (Atanasova, 2017). Artificial neural networks (ANN) are used in various scientific and daily tasks. In the simplest case, it is a multilayered perceptron. Time series are attractive for research with artificial neural networks (Tomov, 2016).

1.3. Application of time lines in cryptography and cyber security

Meeting cyber security requirements is a prerequisite for the safety and security of IT infrastructures, digital resources and the protection of personal data. In its foundation is cryptography, which provides a number of processes, such as authentication, identification, encryption, data approving processes and etc. The main root of

cryptography is random numbers, and in the most frequent case for modern cryptography needs, two types of random number generators are used:

- True Random Number Generator (TRNG);
- Pseudo Random Number Generator (PRNG).

Traditional RNG measures are mainly aggregated statistics relating to deviations from mathematical chance. In order to help check the quality of a random number generator, its output may be stored in a time series and the data may be subjected to specialized mathematical analyses.

1.4 Conclusions

As a result of the conclusions drawn, it should be summarized that time series studies in different fields and applications need to develop specific methods and means to achieve the specific objectives.

Chapter 2. Methods for examining and forecasting financial time series

This chapter examines the widespread Momentum Indicator that belongs to the oscillators group. Its calculation is based on a mathematical time series processing apparatus. The dissertation aims to improve its effectiveness.

2.1.1 Momentum Oscillator

Momentum is a major oscillator that shows whether the price trend is accelerating, slowing down or moving at the same rate. The function of this oscillator is to take into account the acceleration of the price trend. As confirmation of the signals for torque differentiation, figure formations from technical analysis are included in the study to combine and confirm the current price reversal. In the specific example of Fig. 2.2. Momentum and the signal for divergence 1C confirm an upcoming price turnaround, through a multiple peak of 1D.

Fig. 2.2. EUR/USD Forex historical chart between September 2006 to August 2008 period on a daily time frame

2.1.2 Weaknesses in market trend analysis through Momentum

However, it is clear in the study that the rear incidents in which Momentum may make an exception and not take into account a divergence at the completion of an ongoing market trend, where it is depicted with 1C in Fig. 2.3.

Fig. 2.3. USD/CAD Forex historical chart between September 2006 and August 2008 period on a daily time frame

According to the example presented by the real forex market, the value of Momentum reaches a higher peak even than the previous one. But the price then makes a significant adjustment of about 60% without the oscillator's divergence. The question that excites the study is therefore, can the accuracy of the Momentum oscillator be improved?

2.1.3 Method for increasing the accuracy of Momentum

This dissertation provides a non-traditional method for signaling a market turnaround, namely the **MA Volatility Indicator** developed. And the MA is divided into two under the form:

- Simple moving average (SMA):

  To calculate the SMA, a time series is used to sum the data of the last periods*(t) where, for example, t*=10 for10 days, according to the time frame (may be *a different value,* optional). Then divide by the number of t periods. This calculation is made for the period of each bar of the chart. The SMA formula is as follows:

  $$SMA_t = \sum_{n=1}^{t} price_n / t$$

- Exponential moving average (EMA):

To reduce the lagging effect of SMA, technical analysis users often prefer Exponential Moving Average (EMA). They reduce the backlog by adding new values to

the latest prices depending on the length of the MA. The shortest EMA will be of greater value than will be applied to most MA.

$$X = K * (C - P) + P,$$

where X – current EMA, C – current price, P – EMA from the previous period (a value of SMA is used for the calculation of the first period), K – smoothing coefficient.

The adjustment coefficient applies an appropriate coefficient to the newer prices, which are related to the previous EMA prices. Formula for smoothing coefficient:

$$K = 2 / (1+ N),$$

where N – number of previous EMA prices.

A conventional approach to MA trading is a higher time frame for the price not to cross MA, as in case of market adjustment, reaching MA from the market price is considered a strong support for the current trend. In case of a break in the MA price, it is taken as a signal for a reversal and, in the case of a bounce, as a signal to confirm the current market trend. The other method is an analysis with more than one MA, all of which are at different speeds.

The MA Volatility Indicator method developed in the dissertation relies on determining extreme values for moving the price away from MA, On the basis of which to determine the sentiment of market participants at the present time.

Fig.2.8 Combining Momentum with the proposed method - MA Volatility Indicator (historical Forex USD/CAD chart, on a daily time frame, period is between 1997 and 1999)

On the Figure 2.8, the MA Volatility Indicator method is applied and combined with the values of Momentum, and the simulation data is historically real from the Forex market. Figure 2.8 makes it clear that Momentum, depicted with a green line, does not take into account lower values at the last market peak. This is highlighted by the straight red line of its trend.

In conclusion, it can be said that Momentum is an effective oscillator, which has become part of many automated systems and trading strategies in financial markets. But the MA Volatility Indicator method manages to improve forecasting accuracy. Therefore, it could be applied both to automated systems and to analysis of market trends by man.

2.2 Forecasting financial time series via neural networks

Multilayer perceptron is the most commonly used type of artificial neural network that can be presented as an oriented weighted graph. In this study, the basic idea is that instead of the number of hidden layers, the number of neurons at the entrance increases and the hidden layers expand during neural network training. The extension of the input layer is related to the fact that each time line grows with the appearance of a new measurement. The purpose of the training is to make the size of the input layer as large as the size of the full time line.

2.2.1 Modeling prerequisites

The proposed model uses a set of artificial neural subnets and these subnets are joined into a common artificial neural network. The smallest artificial neural subnet has a 1-1-1 topology (Fig. 2.2.6 - left). The network is trained with examples that have only one value. The target in the model is a forecast of only one value ahead of time. Therefore, all sub-modes have only one output. All input values are provided as examples of elastic learning to reverse the error. The training stops at a certain *level of epsilon* for complete change of neural network errors.

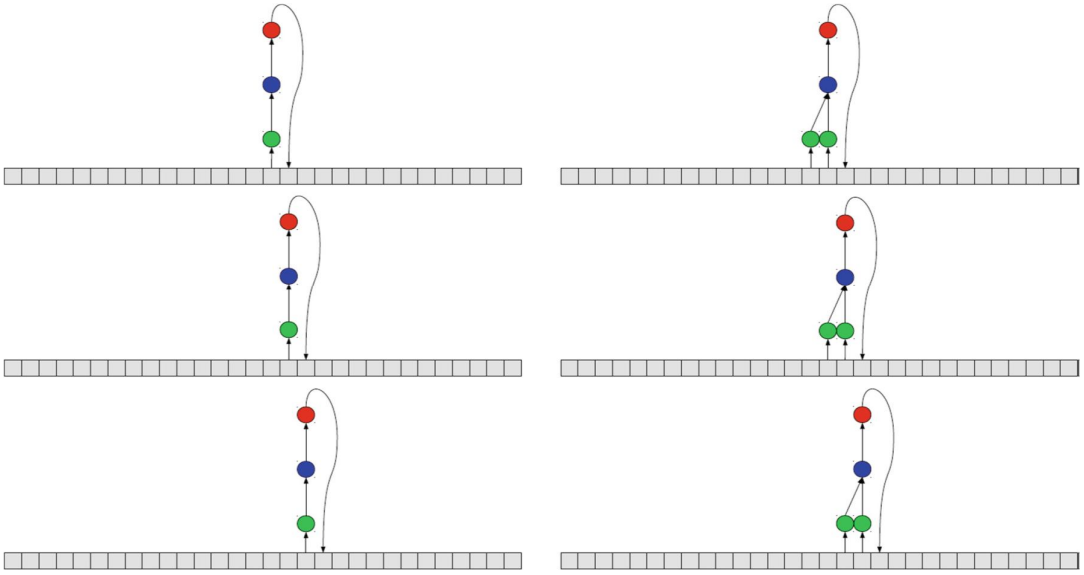

Fig. 2.2.6. Training of artificial neural subnets with 1-1-1 topology (left) and 2-1-1 topology (right).

After training the 1-1-1 topology, the weight values of the first subnet are loaded in the second subnet with a 2-1-1 topology (Fig. 2.2.6-right). A third subnet has a 3-2-1 topology. The size of the hidden layer is selected automatically by a gradual pruning algorithm implemented in the Encog Machine Learning Framework (http://www.heatonresearch.com/encog/). The topologies of the subnets are formed by adding one neuron to the input layer and adjusting the size of the hidden layer with an incremental trimming algorithm. The ultimate goal is to reach n-m-1 topology (Fig. 2.8), which covers all known values of time lines.

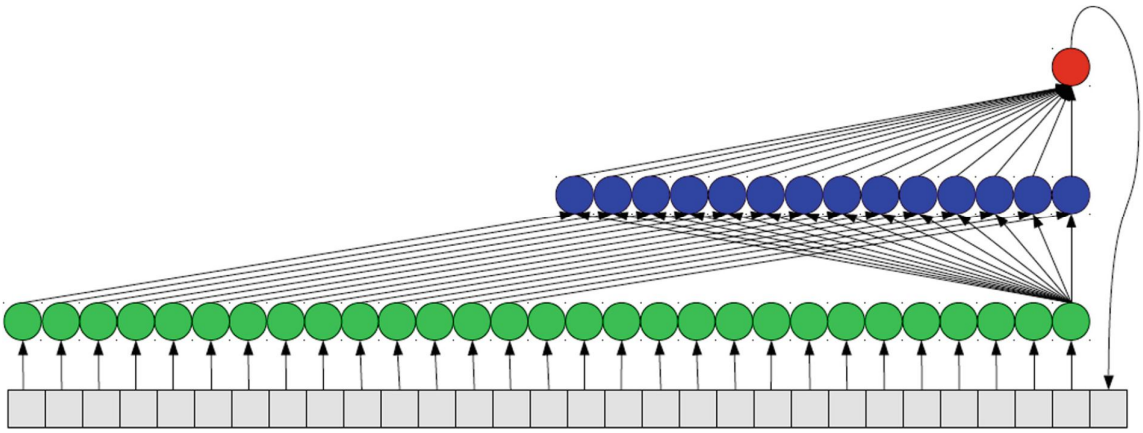

Fig. 2.2.8. Training of artificial neural subnet with n-m-1 topology. Some of the links between the input layer and the hidden layer are not displayed for a better look.

The general idea behind the proposed model is the gradual training of racing in size artificial neural networks. The common problem in the training of artificial neural networks is the size of the network. The proposed model has a higher degree of self-adaptation, since when a new value appears in the time line, the size of the artificial neural network increases, which means that the training phase and the work phase are simultaneous.

2.2.2 Experiments on the study

Experiments are done through a JAVA program where artificial neural networks are implemented through the API provided by the Encog Machine Learning Framework. As input data for experiments, financial time lines are used on the FOREX market. Data are taken from daily two-month trading for EUR/USD and USD/JPY currency pairs. Time line values are scaled in the range from -0.99 to +0.99 with the Min Max zoom rule. The results of the experiments are still within the range of statistical error that comes from the complexity of financial processes and the high-frequency noise inside the data.

The proposed model for self-build three-layer MLP for predicting time series is a promising approach to speeding up the training of artificial neural networks.

2.3 Conclusions

This chapter offers new methods for analyzing and forecasting market price movements through time series and neural networks.

As a result, the following conclusions have been drawn:

1 The study so far covers the main aspects of the analysis process – from defining the problem and placing the tasks, to presenting methods for solving them. In each of the stages, real evidence is presented to identify weaknesses or the need to find a more rational approach in the area under question.

2 The methods allow them to be integrated into automated processing and decision-making systems. The developed method (MA Volatility Indicator) improves oscillator precision (Momentum) and works in the combination of two EMA or

SMA instruments, offering a new methodology for interpreting results in market analyses and helping to reduce the risk of losses and increase success in automated trading.

3 The proposed self-build training algorithm in three-layer MLP accelerates ANN's training in forecasting financial time lines.

The methods presented so far can be applied by specialists in different fields in systems of an estimated nature, for decision-making, analyzing events and processes based on time lines.

Chapter 3. Solutions for providing cryptographic protection by applying time series in cryptography and cyber security

The dissertation proposes time series approach be applied to the quality analysis of a random number generation system (RNG) to ensure cryptographic protection in information systems. For the current RNG survey, a numeric array is retrieved to analyze the values from random numbers in time rows. The results are displayed graphically, where the vulnerable random numbers produced by the generator become more prominent.

3.1 Application of time line techniques for analyzing a random number generator in the field of cyber security

RSA is an asymmetric encryption algorithm that allows anyone to send encrypted messages that only the private key holder can decode. The principle of operation can be explained briefly by generating a very large random number *p*, then generating another such number *q and* calculating their *work x=p*q*, in *fact x is* known as a public key.

3.1.1 The researchers of the (almost) secret algorithm – weaknesses due to insufficient RNG entropy

On the surface, RSA encryption seems invincible. But according to the study presented, the problem lies in the random number generators that provide the algorithm. The vulnerability is fundamental and comes from the fact that RSA needs very large numbers to create encryption keys, and generators in mass computer systems have significantly less capacity. By producing from the number generator, a starting value called seed, entered into the pseudo generator and after computing-intensive computing, cryptographic RSA keys are generated. The problem is that of devices such as phones, IoT, small routers, etc. small systems are even more pronounced because they often do not have sufficient resources for this laborious work. This greatly speeds up the process of generating RSA keys as needed, but opens up a major vulnerability in cryptography security. And considering these disturbing observations, they are reason enough to do research on the subject.

The modern criteria for a reliable RSA key is a minimum of 2048 bits, the recommended length being even 4096 bits. Other research has also found that between 4096, 8192 and 16384 bits of RSA key, greater security of larger keys is minimal. The

reason also comes from the limitations of random number generators. For larger RSA keys, extremely large real random numbers are required. Which in a computer system is extremely difficult to obtain.

And if weaknesses in cryptographic functions are not illuminated, we run the risk of being discovered and exploited by malicious individuals without it being known to others. In conclusion, it can be said that the weaknesses do not proceed from an error in RSA arithmetic. They come from the technological weakness with which RSA is applied.

3.2 Method for assessing the vulnerability of random number generators for cryptographic protection in information systems

The subject of the study includes the technology of the widespread PHP programming language. For the needs of systems developed with this technology to ensure the need for random numbers, PHP has the following means:

1. Lyne congressional generator (LCG), e.g. lcg_value()
2. The Marsenne-Twister algorithm, e.g. mt_rand()
3. Locally supported function C, i.e. rand()

They are also reused for functions such asarray_rand() and uniqid(), and the downside of entropy and random number generators of the above functions consists in easily predicting future PRNG values. This is because the initial internal states or PRNG SEED are limited and the output of values is in an insufficient range, and this is predictable from readily available modern computational resources. Often, to get a SEED value in PHP, developers use mt_rand() or the following script to use automatically:

```
<? php

mt_srand(3231153718);

for ($i=1; $i < 15; $i++) {

   echo mt_rand(), PHP_EOL;

}
```

Which, due to the weak entropy of the tools on offer, risks the recovery of SEED by an attacker. For this purpose, a simulation of a real information system is created in the study, which uses the following source code to generate a token for the different purposes of the application:

$newtoken = hash('sha512', mt_rand());

Generating a token in the presented way is a nice example, such as a single conversion to mt_rand(), which is the hash with SHA512. The fact is that, in fact, if a programmer assumes that the functions of the random PHP values are "random enough", he will be much more inclined to embed a simple usage model. But the method used above to generate mark erssuffers from one flaw - random values are limited to numbers (i.e. its uncertainty or entropy is close to negligible). If you check the output of mt_getrandmax(), it will be found that the maximum random number mt_rand() can generate only 2.147 billion. This limited number of options makes it vulnerable to a violent attack. In the presence of a modern good video card (GPU) and with the help of specialized brute force attack software such as hash cat, such a calculation can be completed in just a few minutes. Therefore, the use of hash to hide the output of mt_rand() is useless.

To protect this type of system, random values of higher quality must be generated. For use in non-trivial tasks, PHP requires sources of high-end entropy that can be provided by the operating system. In Linux is usually used with /dev/urandom, unless devices with even high erentropy are installed. In Linux, with the correct setting, a regular random number generator that is of the PRNG type (which is a pseudo random number generator), is often loaded from a source of high entropy /dev/random, which here it makes it resistant to attacks. Therefore, any software system developed with PHP inorder to be well protected should be redirected to the reuse functions of the mcrypt_create_iv openssl_pseudo_random_bytes OpenSSL external library. They are optimized to use a cryptographically protected pseudo-random generator. Which is tailored and integrated with the operating system.

3.2.2. Understanding RNG Entropy in Linux

In the Linux operating system, the random numbering architecture has the following type:

1. /dev/random is a real random number generator if the entropy ends.
2. /dev/urandom is a pseudo random number generator (PRNG) and it is not blocked due to entropy depletion.

3. /dev/hwrng is an additional hardware for true random numbers that is specialized and not installed in the default computer systems. It provides entropy noise to maintain random numbers;

The accumulated entropy in Linux system can be verified by the following command:

$ cat /proc/sys/kernel/random/poolsize

4096

$ cat /proc/sys/kernel/random/entropy_avail

3868

Where:

/proc/sys/kernel/random/poolsize is used to declare the size (in bits) of the Entropy Pool buffer, for example: How many random numbers should we store before we stop "pumping" for more.

/proc/sys/kernel/random/entropy_availshows the quantity (in bits) of currently stored random numbers in the pool.

Through the user activity and operation of the computer system, such as network, disks, memory state, central processor, peripherals, etc. With the penal functions in the Linux kernel, they have functions for continuously procuring random numbers. Which is designed to compensate for the constant need for them when the computer system is working. For the purposes of the study, such a situation can be easily triggered in order for this process to be observed. By next command, just discard everything that is in /dev/random random generator of random numbers and displayed on the screen:

$ hexdump /dev/random

0000000 d5c4 ff0a b8ef 9bdc ad95 480b e853 f0ef

0000010 e0cb 7c08 4bc4 daef 2b21 ea62 0eac 2c6c

0000020 d6bd 70e6 5d6f a7e3 0874 d52f 77df 6a2b

0000030 1909 efe8 9964 acee 2aad 2522 4ddb 1d0b

At the same time, the entropy buffer status may be displayed in a parallel open command terminal, with the content refreshed every second. To do this, it is necessary to run the following combination of commands:

```
$ watch  -n 1  cat  /proc/sys/kernel/random/entropy_avail
```

As a result, the presence of entropy will begin to decline, and its condition will reach critical values, even to zero. By pressing Ctrl-C, this pointless waste is stopped. Perhaps this should never be done in practice - especially on a real server system- except for research purposes of course. But often the systems have problems with the accumulation of entropy in the buffer, and the result seems disturbing:

```
$ cat  /proc/sys/kernel/random/entropy_avail

96
```

From the example presented, the machine produced an entropy result of 96 bits and the increase in this value is too slow and insufficient. The reasons for this can be heterogeneous. For example, from a lack of specific hardware, incorrect settings, virtualization, too much activity with random system numbers, and inability to compensate for the contagion of random values, etc. One possible solution is to launch specialized software to help collect random numbers. This is a daemon that is designed to use any events that can be considered relatively random when the machine is working to produce more and better random numbers. For example, the cpu "flicker", the change in memory status, input output operations, network traffic can add more entropy to the system buffer. Installing this solution and the basic setting in the system are as follows:

```
# apt  install  haveged

# systemctl  start  rngd

# update-rc.d  haveged  defaults

# rngd  -r /dev/urandom
```

On a system with relatively moderate traffic:

```
# pv  /dev/random  > /dev/null
```

```
 40 B 0:00:15 [ 0 B/s] [ <=> ]

 52 B 0:00:23 [ 0 B/s] [ <=> ]

 58 B 0:00:25 [5.81 B/s] [ <=> ]

 64 B 0:00:30 [6.05 B/s] [ <=> ]

^C

# systemctl  start  haveged

# pv  /dev/random  > /dev/null

7.12MiB 0:00:05 [1.43MiB/s] [ <=> ]

15.7MiB 0:00:11 [1.44MiB/s] [ <=> ]

27.2MiB 0:00:19 [1.46MiB/s] [ <=> ]

 43MiB 0:00:30 [1.47MiB/s] [ <=> ]

^C
```

Using the pv command, you can see how much data is transmitted for this purpose. It is clear from the data flow shown that 2.1 bits per second (B/s) were obtained before, while then ~ 1.5 MB / sec was obtained.

3.2.3. Time series for random number generators

The specifics of RNG and PRNG allow them to be analyzed using time serires analysis and forecasting techniques, such as capturing the flow of output numerical values as a sequence and in itself arranged sequentially over time. Such a flow of numerical values can be described as follows:

$$N = T * V$$

where: N - the length of the number row, T - time (duration) of the generation of numbers, V - *number* of generated numbers per unit of time.

So through the time lines it is possible to determine the quality of entropy over time. If a random number generator is not very reliable, then its weaknesses could be found for a shorter time series of data, for which fewer resources will be needed for processing and analysis. For the needs of the current study, a number array will be used,

which will not be created by a high-quality random number generator, but by a mediocre one. The idea is to apply the approach and analyze the time series of random values of a mid-range computer system, which everyone usually has.

3.2.4. Study of random number generators with time series

Time series as stochastic process can also be used for analysis of RNG/PRNG. For this purpose, an algorithm has been developed to detect repetitive patterns (patterns) of data in the time lines generated by RNG. A specially written program is used to collect data from random number generators in a time series for the needs of the survey. Using a re-written program, the random number data collected is presented graphically, helping to make it easier to spot important elements of the time series (Fig. 3.2 and Fig. 3.3). At first glance, with the results of the data from System.Random in Fig. 3.2, everything is fine and it is possible to think that they have a good quality of entropy. But let me suggest another way in another graphical view to make sure of our judgment.

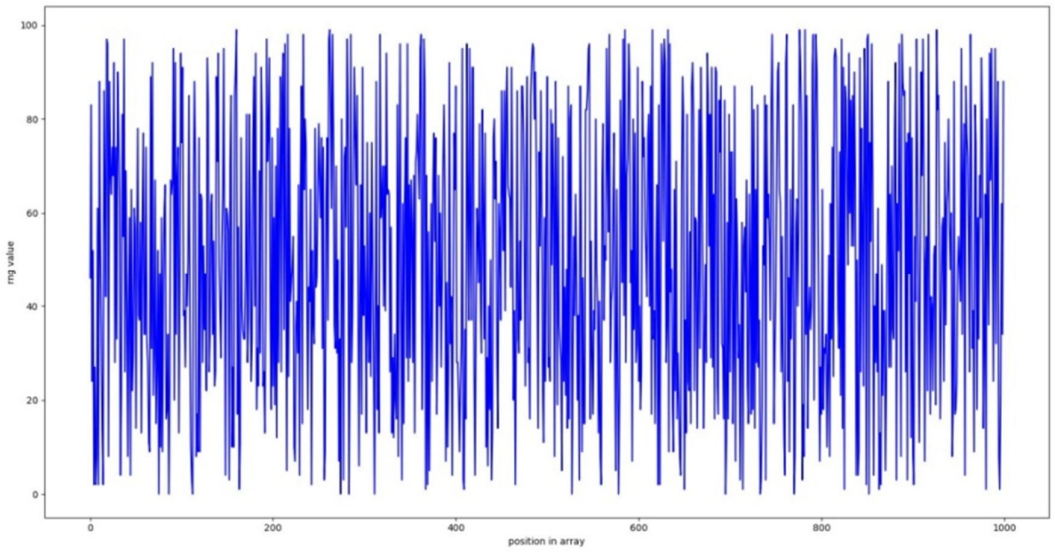

Fig. 3.2. Visualization of data obtained by System.Random as a noise diagram.

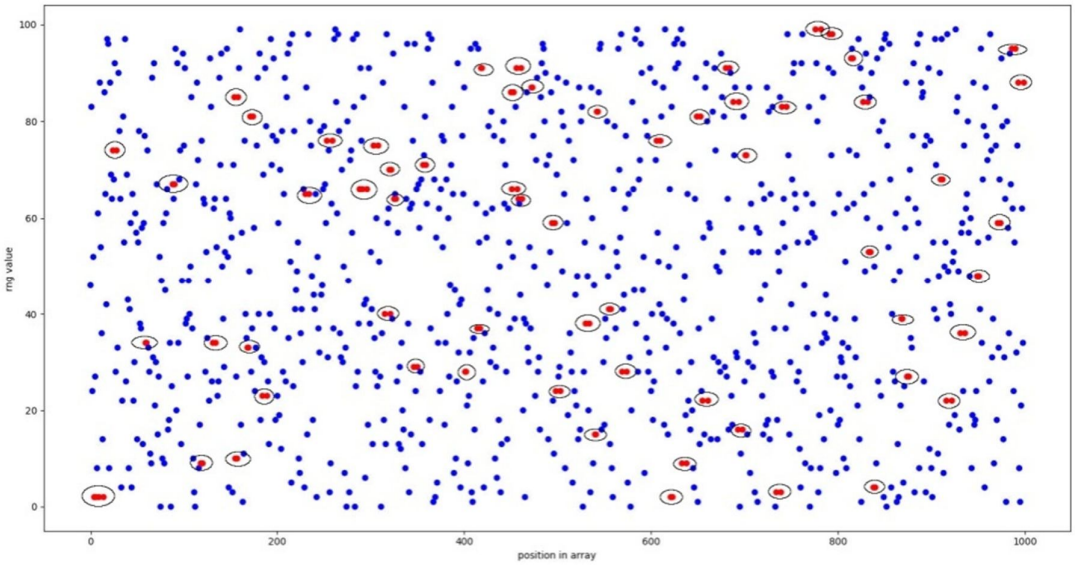

Fig. 3.3. Presentation of System.Random data in a field of dots preview

Presenting the same data with a different graphical interpretation can help to reveal some quality problems with the values studied. In Fig. 3.2 and Fig. 3.3. They are recognized and stained red and circled for better visibility.

The visualization of Fig. 3.3 shows the weaknesses of the processed results. Patterns of recurrence occur periodically over time. These cases are colored red by a program developed using predefined forecasting models, as mentioned earlier. If such a random number generator is used in cryptography, the SEED values produced by it can be successfully attacked by predicting the next SEED value or by monitoring encrypted data transmitted, the values underlying the encryption system can be adopted at a certain point.

3.3 Neglected cyber security risks in public internet hosting service providers

So far, RNG quality analytics research and problem areas have been able to affect cryptographic algorithms, programming languages and operating systems. Now the focus is shifting to massively offered public hosting services. This research uses a web hosting provider that is one of the most popular in the industry. The web application service is installed on a mass-marketed shared hosting. Web certificate added and SSL access is enabled, all running on standard communication ports. On the first line of client-server security, the critical ciphers supported by the hosting server come out. If they are up to

date and there are no vulnerable and already outdated and time-compromised ones, it can be considered that the communication protocol is sufficiently secured.

A test was carried out by scanning the cryptographic protocols that provide the connection between client and server (hosting service). Another protocol that the server supports is TLS v1.2, which is still up to date and approved for use, It contains cryptographic ciphers that need to be removed, but the server offers them for communication, which is also a significant security vulnerability of the service provided. The analysis of protocols and ciphers also identified another significant flaw. The TLS v1.3 protocol is not supported at all, this is currently the most up-to-date and secure protocol of the TLS family for tunnel connectivity.

After checking the cryptographic protocols and ciphers supported for communication, and the follow-up moved on to the more sensitive theme – random number generators. To perform this analysis, a computer program was created that establishes server connectivity in the available cryptographic security protocols between client and server. In this case, TLS v1.2 was used, and in the connection phase, the program takes the generated random numbers from the server and saves them to a file as a time line. The program in question runs in a loop until it collects a sufficient amount of data for analysis.

The data collected from random numbers is analyzed using the specialized open source software to analyze random numbers used in Robert G. Brown's cryptography Dieharder (Brown,2021). A simulation of 114 tests, as well as a check of the quality of numbers and the cyber security standard of FIPS-140 random number generators have been performed. In summary, the data from the random number simulation test are:

• Only 25 tests have passed successfully;

• Failed, which have compromised /predictable/ value and therefore detectable cryptography are 76;

• Vulnerable where cryptography can be revealed with relatively good computer hardware are 13;

From the results presented, it can be concluded that the that due to weaknesses in random numbers and the identified violation of cryptographic protection, the risk of success in cyber-attacks for compromising cryptography is critically high.

The solutions that are allowed in this case are to use private hosting on its own infrastructure, which will not allow excessive load of the type described. However, if it is not possible to provide continuity for a hardware configuration and a suitable location, such as a server room, it is better to rent a VPS server that will only be under the control of one client and also avoid the problem. However, action can also be taken on the part of the hosting provider to increase the capacity of cyber security. The techniques for configuring proper functioning and enhancing the capacity of Linux entropy described in section 3.2.2 "Understanding RNG Entropy in Linux" of this thesis should be applied.

After the correct system setup, you can resort to another unconventional approach, noting the principle of working on collecting entropy in its buffers from the Linux operating system. A program can be written that generates a series of events that will not particularly harass the system, but will create numerous processes supporting the collection of entropy:

```
#!/bin/sh

## list of sites using round-robin DNS

ROUND_ROBINS="www.yahoo.com google.com twitter.com outlook.com"

## Entropy stard and end value limits

STOP_LIMIT="3800"

START_LIMIT="3000"


until [ "$(cat /proc/sys/kernel/random/entropy_avail)" -gt "$STOP_LIMIT" ]

do while [ "$(cat /proc/sys/kernel/random/entropy_avail)" -lt "$START_LIMIT" ]

do for thing in "/tmp/loyeyoung" "/tmp/sueellen" "/tmp/rootdev" "/tmp/files"

do echo $thing ====================================

        touch /tmp/toss

for robins in $ROUND_ROBINS

do nslookup "$robins" 8.8.8.8 > /tmp/toss
```

```
                nslookup "$robins" 9.9.9.9 >> /tmp/toss
                nslookup "$robins" 192.168.2.3 >> /tmp/toss
                nslookup "$robins" >> /tmp/toss
                cat /tmp/toss
mkdir $thing  -p
cp /tmp/toss  $thing/toss
                cat $thing/toss
                rm -f /tmp/toss
                rm -f $thing/toss
              done
            done
          done
        done
```

The presented program script is a quite basic and could be upgraded and compiled in other programming or scripting languages. The rate of entropy build-up has improved. Which contributes to the system in question to bear greater loads on the generation of RNG values. The mode of action is as currently specified is that the additional operations in memory, processor, disk and network will be activated when a value in the entropy buffer reaches less than 3000. Also, the solution provided could be used in combination with hardware solutions supporting cryptographic algorithms and random number entropy, which Intel also provides in its processors.

The name of the random number generation module is Intel Secure Key, its previous code name is Bull Mountain Technology. Therefore, it must be verified whether the current system has such processors and its configuration could be upgraded. In the presence of a computer system with a Linux operating system, verification may be done except through the technical documentation of the chips from the manufacturer and through the following combination of commands:

```
$ cat  /proc/cpuinfo  | grep  -i  rdrand  | echo  $?
```

0

As a result, 0 means that an RDRAND flag is available and the processor can be turned on to improve the cryptographic functions of the system as follows:

```
# apt install rng-tools-debian
# /etc/init.d/rng-tools-debian start
# /etc/init.d/rng-tools-debian status
* rng-tools-debian.service - LSB: rng-tools (Debian variant)
Loaded: loaded (/etc/init.d/rng-tools-debian; generated)
Active: active (running) since Fri 2020-11-28 17:30:54 EET; 3min 10s ago
Docs: man:systemd-sysv-generator(8)
  Tasks: 4 (limit: 4915)
 Memory: 1.3M
 CGroup: /system.slice/rng-tools-debian.service
'-3597 /usr/sbin/rngd -r /dev/hwrng
$ cat /proc/sys/kernel/random/entropy_avail
4096
```

The results show that the rate of entropy collection for our case exceeds the rate of its consumption.

3.4 Results in real technological infrastructure

The proposed approach to improving cyber security in cryptography and random number generators in busy server systems with public services has been applied in the technological infrastructure of the IICT-BAS. The hardware configuration used is mid-range, taking into account the complexity of the task performed. The server is equipped with one six-core Xeon(R) E-2236 V6, 32GB RAM and two hard drives in Raid1 configuration. The public service operating server is operating on Linux and all services are entirely open source software. The services as running from the virtual machine are:

- server, currently with 242 user accounts. Available through SMTP, POP3, IMAP, all of which are protected by cryptographic communication protocol TLS v1.2 and TLS v1.3. Certified with a server certificate to establish TLS sessions with an asymmetric algorithm of the type elliptical curve secp384r1. Connecting to the service cannot take place without encryption of communication;
- Web mail that allows all 242 users to operate their mail and through a web browser. The communication is protected by the cryptographic communication protocol TLS v1.2 and TLS v1.3. Certified by a server certificate for establishing TLS sessions with an asymmetric algorithm of the type elliptical curve secp384r1. Connecting to the service cannot be done without encryption of communication;
- Web portal of the Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences, which is the main web space of the institute. It contains activity information, two scientific journals, and structural information. Communication is protected by cryptographic communication protocol TLS v1.2 and TLS v1.3 and server certificate for establishing sessions with asymmetric algorithm with elliptical curve secp384r1.
- SSH remote administration service with the highest degree of cryptographic protection currently offered by the protocol.
- FTP Web Content Remote Management Service. Communication is protected by cryptographic communication protocol TLS v1.2 and TLS v1.3 and server certificate for establishing sessions with asymmetric algorithm with elliptical curve secp384r1.

For all services, the priority protocol for encrypted connectivity is the latest and most secure protocol TLS v1.3, but if it turns out that the client does not support it passes the protocol TLS v1.2. The latter is left only for compatibility, removing all cryptographic algorithms in which vulnerabilities are detected.

It has been found that at present a server is used significantly intensively by IICT users and external Internet users. The level of cryptographic protection is at the highest at the moment according to established standards and no compromises have been made with cryptographic protocols or ciphers. As evidence of the quality of cyber security by cryptographic means, the result of an SSL Labs scanner test was applied to the level of encryption in the TLS protocols offered on the available services. The results of this test

are derived on the basis of current cryptographic protection requirements at present, which have been validated by the international cryptographic protection laboratories FIPS and NIST for the US and Common criteria for Europe. The results show that the protocols and encryption tools are of the current level. The assessment of all tests is the highest possible A+. The level of protection of the HTTP protocol that communicates with the browser via the TLS tunnel is also the maximum level of protection A+.

Tests have also been carried out on the quality of entropy, using the methods presented in the dissertation. Two well-established methods have been applied by the command shell server, the first checking the quality of FIPS entropy with RNG test, and the second with the Dieharder analysis tool. All tests of the entropy of random numbers shall pass with the highest possible result according to the criteria of the programs in question.

Despite the good results, another study has been carried out which shows whether, at times of high user activity and intensive cryptography load, the capacity of random numbers is exhausted. This statistic is collected within half a month. To make it clear whether there are moments that lead to the depletion of the entropy buffer faster than its charging by the system. As a result, after half a month of data collection in a time series, 2137 values were obtained. These values are displayed in graphical form and are depicted in Fig. 3.10:

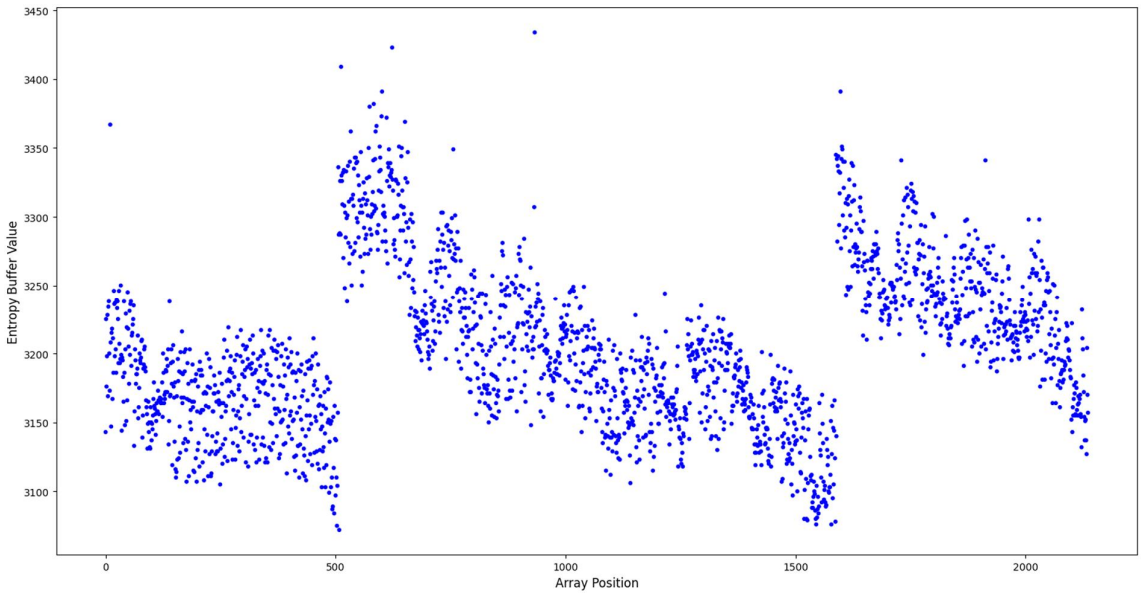

Fig. 3.10 Level of entropy at different moments in time

Fig. 3.10 shows that the system had peaks of more intense activity, which resulted in a strong drawing from the entropy accumulated in the buffer. However, the values at times fell sharply. However, the approach proposed in the study was able to compensate for the high consumption of entropy. While the graph amplitudes between maximum and minimum values appear wide - the values are in a narrow range, with an entropy level between 3000 and just under 3450. The lack of values below 3000 indicates that the system is in very good health and is able to absorb greater loads because the values are far from critical. Taking into account all these results from the actual working environment of the server, the proof of effectiveness of the proposed approach is available. It can therefore be considered that the proposed approach can be beneficial and support the different Internet systems and solutions.

3.4 Conclusions

The research of the services presented and their level of cyber security is key to a more secure transition to modern digital transformation. The rapid transfer of all social and economic activities to digital platforms proves that but in terms of cyber security, many of the current IT services are still lagging behind. Increasing the success rate of cybercrime can lead to a loss of trust in technology and obstruction of these processes, which will also affect scientific and technical progress.

Applying mathematical and statistical analyses with time series to solve cyber security problems is effective. The approaches proposed here can also be combined with other cyber security analysis techniques and methods to be more complex and effective. In this dissertation work, a method for examining the quality of RNG and PRNG in an information system has been developed by applying time series. The method allows to increase the quality of entropy in the use of cryptography, providing various Internet services. The algorithm for detecting repetitive patterns of data generated by RNG has been developed. A study of cryptographic tests and the quality of entropy on real-world busy server systems with public Internet services has been conducted.

Chapter 4. Software approaches to working with big data sets and limited computer resources with programming language R

4.1 The programming language R

Programming language R is a product with powerful tools for statistical calculation and analysis. R is both a programming language and a software environment (Borcard, 2011),(The R, 2017). Language R offers a wide variety of statistical techniques such as linear and non-linear modeling, classical statistical tests, analysis of time lines, classification, grouping, etc., as well as graphic techniques and is extremely expandable (Long, 2015).

Despite the many advantages of R, the richness of its statistical models and data processing tools, as well as powerful visualization capabilities, problems arise when working with large amounts of data. R's limitations stem from the fact that it is designed to operate in single process calculation mode only (single CPU core) and data loaded at once in RAM.

4.2 Overcoming problems with big data using a multi-core microprocessor

Parallel programming calculation of more than one CPU core is possible by recompiling and adding some program components to R. This is possible due to the fact that R is an open source system and this is one of the advantages that this concept brings.

4.3 Methods for optimizing data volumes

One of the well-known features of the R language is that it loads all the data it operates in the RAM of the computer system, which would be critical even on powerful systems with a large resource. As a way to solve this problem, the dissertation looks at ways to load data into memory by excluding data with incorrect content at the time of loading it.

In some statistical surveys, it is not necessary to load all the data, but only certain time frames in order to make an approximate statistical analysis in a time frame. In this case, positioned reading methods suggested in the dissertation work can be applied. This makes it possible to process only a certain snippet of the data located in a big data file. Another common problem is when loading large data is that after loading in memory and processing, some of the data is no longer needed, but it continues to occupy significant

amounts of memory. The dissertation presents a way to reduce memory data by removing excess data and releasing memory.

4.4 Conclusions

The author's contribution is that this material, with the means of programming language R, helps solve problems when working with large data sets with limited computer resources.

In conclusion, it can be said that with the examples presented so far, the topic of optimized data loading cannot be exhausted when working with programming language R. Working with real data is always a challenge (Baumer, 2017). But the techniques presented so far are between good practices and are often used, they could also be combined with other approaches to problem solving in this area.

Conclusion - summary of the results obtained

The dissertation examined in detail methods and means of using time lines in solving various tasks arising in modern applications of information technology and systems.

A method titled MA Volatility Indicator has been proposed to improve precision in oscillator (Momentum). MA Volatility Indicator works in the combination of two EMA or SMA tools and offers a new methodology for interpreting results, which contributes to the detection of levels of over-purchase and over-selling in the market trend. All EMA, SMA and Momentum tools used in the study, as well as MA Volatility Indicator use time series.

The applicability of the neural networks apparatus for forecasting time lines in the financial field has been examined. It has been shown that a new model of presentation of input data characteristic of financial indicators results in a higher degree of self-adaptation in neural network training. Experiments conducted confirm the complexity of the financial processes and the presence of high-frequency noise in the data.

A method for examining the quality of RNG and PRNG in an information system has been developed by applying time lines to increase entropy in the use of cryptography providing various Internet services. This contributes to better cyber security of it infrastructure for digital resources and data protection. In the dissertation, the topic of cryptography received special attention due to its critical importance.

The practical results of the actual experiment showed that the golden ratio between mass services and actual cyber security requirements was found.

In view of the work carried out in this dissertation and the results obtained in the course of the studies and set out above, the following scientific and applied results may be formulated:

1. A method entitled MA Volatility Indicator has been developed to combine indicators for detecting price movements with new approaches when using time lines of financial data.

2. The apparatus of artificial neural networks shall be applied for the purpose of examining financial time lines. An algorithm has been developed to train the neural network by increasing the size of the neural network input and creating a hybrid structure, and a model for self-build three-layer MLP has been proposed.

3. A method has been developed to increase cryptographic protection in information systems based on studies on the quality of random number generators.

4. Experimental research has been carried out to solve cyber security problems in public widespread hosting services. The results obtained confirm the validity of the proposed method of enhancing cyber security.

5. Programming methods have been developed for efficient operation with large data with means in the R language.

6. The developed methods for increasing cryptographic protection are implemented in the technological infrastructure of IICT-BAS. A study of cryptographic tests and the quality of entropy on real-world busy server systems with public Internet services was conducted.

Guidelines for future research

The guidelines for future research on the subject of the dissertation include:

- Implementation of the MA Volatility Indicator method and its application in combination with other methods of analysis and forecasting of market price trends;
- Application of the MA Volatility Indicator method to automated systems for analyzing market trends and extracting decision-making signals;
- Conducting more research in the field of training algorithms and neural network systems for analyzing and forecasting time lines;

- The development of new methods to increase cryptographic protection in information systems;
- Research of a combination of developed method with other RNG methods and systems for analysis in cryptography and other technological fields to help create and improve RNG, as well as to more accurately determine the range of tasks that the generator can perform well;
- Find more approaches to loading and filtering big data to make it more efficient.

Publications on the subject of the dissertation work

1 **Ivan Blagoev**, Nikolay Dokev, Combining Momentum with one method for predicting market price movements for more accurate results (Combination of Momentum with One Method for Forecasting Market Trends to Improve the Results),International Scientific Conference"UNITH'17" – Gabrovo, 2017 Selected papers,ISSN 2603-378X, pp. II-265-II-270

2 **Ivan Blagoev**, Methods for Optimized Use of Computer Memory during Data Loads with R Programming Languages, International Conference "Automatics and Informatics'2017", 4-6 October 2017, Sofia, Bulgaria, ISSN:1313-1850, pp.213-215.

3 **Blagoev I.,** Improving the Momentum Oscillator Accuracy by a Method for Forecasting of Market Price Movements, Collection of Reports from An International Conference, Vasil Levski National University, 14-15 June 2018, vol. 9, p. 1. 177-185. (ceeol.com)

4 **Blagoev I.,** Method for More Reliable Users' Authentication on the Internet, Collection of Reports from International Conference, Vasil LevskiNational University, 14-15 June 2018, Vol. 9, p. 1. 167-176. (ceeol.com)

5 **Blagoev, I.,** Using R Programming Language for Processing of Large Data Sets, Proc. Int. Conf. Big Data, Knowledge and Control Systems Engineering – BdKCSE'2018, 21-22 November 2018, Sofia, Bulgaria ISSN 2367-6450, pp. 91-98.

6 **Ivan Blagoev**, Application of Time Series Techniques for Random Number Generator Analysis, Proceedings of XXII Int. Conference DCCN 2019, September 23-27, 2019, Moscow, Russia, pp.437-446. ISBN 978-5-209-09683-2, 2019 (RINZ).

7 **Blagoev I.,** Neglected Cybersecurity Risks in the Public Internet Hosting Service Providers. *Information&Security International Journal* - ISIJ, 47, no. 1, pp. 62-76 (2020)

8 Balabanov T.D., **Blagoev I.I.,** Dineva K.I. (2018) Self Rising Tri Layers MLP for Time Series Forecasting. In: Vishnevskiy V., Kozyrev D. (eds) Distributed Computer and Communication Networks. DCCN 2018. Communications in Computer and Information Science, vol 919. Springer, Cham. https://doi.org/10.1007/978-3-319-99447-5_50, pp. 577-584, **SJR:0.188**

9 **Blagoev I.** (2020) Method for Evaluating the Vulnerability of Random Number Generators for Cryptographic Protection in Information Systems. In: Dimov I., Fidanova S. (eds) Advances in High Performance Computing. HPC 2019. Studies in Computational Intelligence, vol 902. Springer, Cham. https://doi.org/10.1007/978-3-030-55347-0_33 **SJR:0.215**

Citations noted

I **Blagoev, I.,** 2018. Using R Programming Language for Processing of Large Data Sets, Proc. Int. Conf. Big Data, Knowledge and Control Systems Engineering – BdKCSE'2018, pp. 91-98.

It is quoted in:

1 Dineva, K., Atanasova, T.: Regression Analysis on Data Received from Modular IoT System. ESM'2019, EUROSIS-ETI, ISBN: 978-9492859-09-9, EAN: 9789492859099, pp.114-118, 2019

2 Ivaylo Blagoev, G. Vassileva and V. Monov, "Methodology for content preparation of online courses," 2020 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 2020, pp. 1-4, doi: 10.1109/ICAI50593.2020.9311364.

II **Blagoev I**., Neglected Cybersecurity Risks in the Public Internet Hosting Service Providers. *Information&Security International Journal* - ISIJ, 47, no. 1, pp. 62-76 (2020)

It is quoted in:

3 M Terzieva, D Karastoyanov, ICT for Innovation in Advanced Banking, PROBLEMS OF ENGINEERING CYBERNETICS AND ROBOTICS • 2020 • Vol. 73, pp. 47-54 p-ISSN: 2738-7356; e-ISSN: 2738-7364, doi: 10.7546/PECR.73.20.05

III **Blagoev I.**, Method for more reliable users' authentication on the Internet, Collection of reports from international conference, Vasil LevskiNational University, 14-15 June 2018, vol. 9, p. 1. 167-176.

It is quoted in:

4 Ivaylo Blagoev, Gergana Vassileva and Vladimir Monov, "Methodology for content preparation of online courses," 2020 International Conference Automatics and Informatics (ICAI), IEEE, Varna, Bulgaria, 2020, pp. 1-4, doi: 10.1109/ICAI50593.2020.9311364.

5 Dineva, K., Atanasova, T.: Security in IoT Systems. Proceedings 19th International Multidisciplinary Scientific Geoconference SGEM 2019, 19, 2.1, ISBN:978-619-7408-79-9, ISSN:1314-2704, DOI:10.5593/sgem2019/2.1, 576-577. SJR (Scopus):0.232 Q4

Participation in projects

1 National Scientific Program "Information and Communication Technologies for a Digital Single Market in Science, Education and Security" (ICT in the IA) -2018-2021.

2 Project Zora on Order No 147/14.06.2019 "Digital and Cyber Sustainable IICT"

Awards

1. Award of IICT-BAS for excellence in 2019 in the category "PhD students".

Bibliography

1 Atanasova, T., Barova, M.: Exploratory analysis of Time Series for hypothesizes feature values. In: International Scientific Conference UniTech 2017, vol. II, pp. 399-403, University publishing house V. Aprilov, Gabrovo (2017)

2 Balabanov T.D., Blagoev I.I., Dineva K.I. Self Rising Tri Layers MLP for Time Series Forecasting. In: Vishnevskiy V., Kozyrev D. (eds) Distributed Computer and

Communication Networks. DCCN 2018. Communications in Computer and Information Science, vol 919. Springer, Cham. https://doi.org/10.1007/978-3-319-99447-5_50 (2018)

3 Balabanov, T., Atanasova, T., Blagoev, I., Activation Function Permutation for Multilayer Perceptron Training, International Conference on Big Data, Knowledge and Control Systems Engineering BdKCSE'2018, Sofia, Bulgaria, ISSN 2367-6450, pp. 9-14 (2018)

4 Blagoev I., Dokev N.: A Method for Investigating the Alterations in the Price Trends of the Currency Markets and Forecasting of Probable Future Alterations, *Problems of Engineering Cybernetics and Robotics*, vol.65, pp.39-48 (2012)

5 Blagoev I., Neglected Cybersecurity Risks in the Public Internet Hosting Service Providers. *Information&Security International Journal - ISIJ*, 47, no. 1, pp. 62-76 https://doi.org/10.11610/isij (2020)

6 Blagoev I.: Method for Evaluating the Vulnerability of Random Number Generators for Cryptographic Protection in Information Systems. In: Dimov I., Fidanova S. (eds) Advances in High Performance Computing. HPC 2019. Studies in Computational Intelligence, vol 902. Springer, Cham. https://doi.org/10.1007/978-3-030-55347-0_33. (2021)

7 Blagoev, I., Using R Programming Language for Processing of Large Data Sets, Proc. Int. Conf. Big Data, Knowledge and Control Systems Engineering – BdKCSE'2018, 21-22 November 2018, Sofia, Bulgaria ISSN 2367-6450, pp. 91-98.

8 Camara C., Martín H., Peris-Lopez P., Aldalaien M., Design and Analysis of a True Random Number Generator Based on GSR Signals for Body Sensor Networks, *Sensors* 19, 2033; doi:10.3390/s19092033 (2019)

9 Plummer T., Forecasting Financial Markets: The Psychology of Successful Investing, January (2010)

10 Pseudo-Random Number Generators, https://crypto.stanford.edu/pbc/notes/crypto/prng.html

11 Zhao P., R with Parallel Computing from User Perspectives, https://www.r-bloggers.com/r-with-parallel-computing-from-user-perspectives/ (2016)

12 Brown R. G.: Dieharder: A Random Number Test Suite, https://webhome.phy.duke.edu/~rgb/General/dieharder.php (2021)

13 Ciampi F., G. Marzi, S. Demi, M. Faraoni, The big data-business strategy interconnection: a grand challenge for knowledge management. A review and future perspectives, *Journal of Knowledge Management,* Vol. 24, Issue 5 (2020).

14 Koeune F. Pseudo-random number generator. In: van Tilborg H.C.A. (eds) Encyclopedia of Cryptography and Security. Springer, Boston, MA . https://doi.org/10.1007/0-387-23483-7_330 (2005)
15 Borcard, D., Gillet, F., Legendre, P. Numerical Ecology with R, Springer, pp. 9 – 30 (2011)
16 Baumer B. S., Kaplan D. T., Nicholas J., Modern Data Science with R, Horton Chapman & Hall/CRC, Boca Raton, (2017)
17 Long C. (Ed.) Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley & Sons, Inc., (2015)
18 Martínez-Acosta L., Medrano-Barboza J.-P., López-Ramos Á., López J., López-Lambraño Á., SARIMA Approach to Generating Synthetic Monthly Rainfall in the Sinú River Watershed in Colombia, *Atmosphere,* 11, 602; doi:10.3390/atmos11060602 (2020)
19 Mikalef P., Krogstie J., Examining the interplay between big data analytics and contextual factors in driving process innovation capabilities, *European Journal of Information Systems,* Volume 29, - Issue 3: Business Process Management and Digital Innovation https://doi.org/10.1080/0960085X.2020.1740618 (2020)
20 Scott G., Carr M., Cremonie M. , Technical Analysis: Modern Perspectives, e CFA Institute Research Foundation (2016)
21 The R Journal, ISSN: 2073-4859, https://journal.r-project.org/ (2017)
22 Tomov, P., Monov, V., Artificial Neural Networks and Differential Evolution Used for Time Series Forecasting in Distributed Environment, Proc. of Int. conference Automatics and Informatics, ISSN 1313-1850, pp.129-132, Sofia, Bulgaria, (2016)
23 Wafi A.S., Hassan H., Mabrouk A., Fundamental Analysis Models in Financial Markets – Review Study, *Procedia Economics and Finance*, Vol. 30, 939 – 947. Elsevier (2015)
24 Wang, W., Y. Wang, Analytics in the era of big data: The digital transformations and value creation in industrial marketing, *Industrial Marketing Management,* Vol. 86, pp. 12-15, ISSN 0019-8501, https://doi.org/10.1016/j.indmarman.2020.01.005 (2020)
25 Li C., Zhang J., Sang L., Gong L., Wang L., Wang A., Wang Y., Deep Learning-Based Security Verification for a Random Number Generator Using White Chaos, *Entropy,* 22, 1134; doi:10.3390/e22101134 (2020)