Vol. 5. Issue 1, 2019 ISSN: 2453-7314

Central European Researchers Journal

CERES

A WORD OF WELCOME FROM THE EDITORS

Dear Colleagues, Readers and Authors,

This number of the journal CERES has been dedicated of the International Conference on Information and Digital Technologies (IDT 2019). This conference was held at the Faculty on Management Science and Informatics, University of Žilina traditionally. The new event as the Industrial Centre (Exhibition and Special Discussion Section) has been organised at the conference in this year. The preparation of this event has been supported by the grant of International Visegrad Fund (reg.no. 21830315).

The main objective of this event was to assemble participants from universities and industry partners from around the world and encourage for the cooperation. Young researchers and postgraduate PhD students have been involved into this event. IT companies from different countries have been involved at the exhibition and special discussion section under this Centre. It was allow to bring together researches, developers, teachers from academy as well as industry working in all areas of information technologies.

This journal content some of the papers, which have been presented at the Industrial Centre:

- Remote Sensing Satellite Virtual Constellation Optimizing with Target Recognition Probability,
- Comparison of Query Performance Between MySQL and MongoDB Database,
- Development of Foreign Economic Activity at the Regional Level: Impact Factors Modeling,
- Analysis o fData from the Social Media,
- Application of Fuzzy Filtering for Thermal Infrared Satellite Data Resolution Enhancement.

We expect that the journal will be considered as the opportunity facility to support of investigations of young researchers and PhD student in future. It is an essential journal's mission.

With best wishes

Editors

Central European Researchers Journal. Volume 5. Issue 1

Addres of the editorial office: Central European Researchers Journal - editorial, Faculty of Management Science and Informatics, University of Zilina, Univerzinta 8215/1, 01026, Zilina, Slovakia, editorial@ceres-journal.eu
Each paper was reviewed by reviewers.
Publisher: JMTM, s.r.o., Sad SNP 8, 010 01, Zilina, Slovakia, publisher@jmtm.sk
Published biannually
ISSN: 2453-7314
July 2019

CONTENTS

Ahmed S. J. Abu Hammad
Implementation of a Parallel K-Nearest Neighbor Algorithm Using MPI 1

Alexsandr M. Kondratov, Oleg V. Maslenko
Remote sensing satellite virtual constellation optimizing with target recognition probability 11

Vladislav Solovtsov, Dzmitry Adzinets
Algorithm of routes optimization for mobile robots 17

Ahmed S. J. Abu Hammad
Contour Extraction of Noisy Echocardiographic Images Based on Pre-processing 25

Wateen A. Aliady
Investigating the Role of Preprocessing and Attribute Selection Methods Towards the Performance of Classification Algorithms on News Dataset 31

Roman Ceresnak, Olga Chovancova
Comparison of Query Performance Between MySQL and MongoDB Database 45

Roman V. Fedorenko, Tamas Czegledy, Nadezda A. Zaichikova
Development of foreign economic activity at the regional level: impact factors modeling 52

Ladislav Burita, Taha Nejad Falatouri Moghaddam
Analysis of Data from the Social Media 64

Elena Zaitseva, Mykola Lubskyi, Jan Rabcan
Application of fuzzy filtering for thermal infrared satellite data resolution enhancement 73

Implementation of a Parallel K-Nearest Neighbor Algorithm Using MPI

Ahmed S. J. Abu Hammad

***Abstract*—**The K-Nearest Neighbor (K-NN) algorithm is one of the most commonly used algorithms for Classification. The traditional K-NN algorithm is; however, inefficient while working with large data sets because the computation cost is quite high as we need to compute the distance of each query instance to all training samples and sorting the distances to determine the nearest neighbors. This paper presents a parallel implementation of the K-NN algorithm using Message Passing Interface (MPI) by distributing the computations of the distance value operation to different processors. Also, we focus on improving the performance of sequential K-NN algorithm with proposed Parallel K-NN (PK-NN) algorithm. We applied our experiments on the graduate student's data set collected from the University College of Science and Technology (UCST) – Khan Younis. We compute the execution time, speedup, efficiency, parallel overhead, and parallel cost and discuss our results for serial and parallel algorithms by comparing them. Finally, we find parallel version reduce the time and more effective while applied in large data set than the sequential one.

***Keywords*—**Classification, k-NN algorithm, Parallel classifier, Parallel and distributed computing, Multicomputer cluster.

I. Introduction

Data mining is to discover interesting, meaningful and understandable patterns hidden in massive data sets. Traditional knowledge discovery systems have been found lacking in their ability to handle current data sets, due to characteristics such as their large sizes and high dimensional. Consequently, new techniques that can automatically transform these large data sets into useful information are in strong demand [2,3].

Classification is one of the important data mining techniques whereby a model is trained on a data set with class labels and then used to predict the class label of unknown objects. K-NN is one of the most widely used techniques in classification applications, where the result of the new instance query is classified based on the majority of K-NN category. The purpose of this algorithm is to classify a new object based on attributes and training samples [2].

The K-NN algorithm is very easy to implement and can produce good results, but the computation cost is quite high because we need to compute the distance of each query instance to all training samples and sorting the distances to determine the nearest neighbors. Therefore, parallel computing is an essential component of the solution to speed up K-NN [3].

In this paper, we present a parallel implementation of the K - NN algorithm using MPI and c programming language to accelerate the distance computation and sorting, and we conduct an experiment to study the performance of parallelizing K-NN and compare it with the serial K-NN version as a baseline.

The rest of this paper is organized as follows: Section 2 presents related works. Section 3 describes the PK-NN algorithm model. Section 4 presents the proposed method. Finally, a P-KNN implementation and experimental results, and conclusions are presented in Section 5 and Section 6 respectively.

Ahmed Abu Hammad, University College of Science and Technology, Khan Younis, Palestine (e-mail: asj_hammad@hotmail.com).

II. RELATED WORKS

Serial K-NN Algorithm: The k-nearest-neighbor method [2] was first described in the early 1950s. The method is labour intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition and text mining.

Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by *n* attributes. Each tuple represents a point in a *n*-dimensional space. In this way, all the training tuples are stored in a *n*-dimensional pattern space. When given an unknown tuple, a K-NN classifier searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the *k* nearest neighbors of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, $X=(x_1,x_2,\ldots,x_n)$ and $Y=(y_1,y_2,\ldots,y_n)$ is:

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i + y_i)^2} \qquad (2.1)$$

The standard K-NN algorithm can be described as follows [2]:

1. Determine a parameter *K* = number of the nearest neighbors.
2. Calculate the distance between the query-instance and all the training samples.
3. Sort the distance and determine the nearest neighbors based on the *K-th* minimum distance
4. Gather the category of the nearest neighbors
5. Use a simple majority of the category of the nearest neighbors at the prediction value of the query instance.

Few attempts have been made to parallelize K-NN to increase its speed.

Lianga et al. [3] implemented a parallel learning algorithm. The parallel algorithm is based on the k-NN algorithm. They evaluated the parallel implementation on Compute Unified Device Architecture (CUDA) enabled Graphics Processing Unit (GPU). The advantage of this method is the highly parallelizable architecture of the GPU. Recent development in GPUs has enabled inexpensive high-performance computing for general-purpose applications. Due to GPU's tremendous computing capability, it has emerged as the co-processor of the Central Processing Unit (CPU) to achieve a high overall throughput. CUDA programming model provides the programmers adequate C language like APIs to better exploit the parallel power of the GPU and manipulate it. At the hardware level, CUDA-enabled GPU is a set of Single Instruction Stream, Multiple Data Stream (SIMD) processors with 8 stream processors. They used synthetic data generated by MATLAB for the purpose of evaluation where the number of data objects is 262144 records. Their experiment showed good scalability on data objects. CUK-NN presented up to 15.2 speedup. The result shows that CUK-NN is suitable for large scale dataset. However, since SIMD processors are specially designed, they tend to be expensive and have long design cycles and the scalability of the processors is limited.

Tarhini [3] implement a parallel K-NN method on the CPU rather than a GPU where the degree of parallelism is indicated by the number of available cores. The proposed algorithm is not expected to outperform state-of-the art GPU implementations, but rather, to provide an equivalent performance on the CPU. Hence, the benefit becomes the ability of load sharing between CPU and GPU without degradation or loss of speed upon switching between any of the two processor architectures. The experiment was implemented on an Intel 8-core machine in which the cores are integrated onto a single integrated circuit die (known as a chip multiprocessor). Iris database was used to train the system. The data set contains 50,000 records. The parallel implementation greatly increased the speed of the KNN algorithm by reducing its

time complexity from *O(D)*, where *D* is the number of records, to *O(D/p)* where *p* is the number of processors.

However, our platform comprises a set of processors and their own exclusive memory (multi-computer workstation cluster), this platform is programmed using send and receive primitives, Message Passing Interface (MPI) provide such primitives.

III. P-KNN Algorithm Model

An algorithm model is a way of structuring a parallel algorithm by selecting a decomposition and mapping technique.

3.1 Decomposition Technique

The first step in developing a parallel algorithm is to decompose the problem into tasks that can be executed concurrently by identifying the data on which computations are performed, then partition this data across various tasks [1].

The task performs the computations with its part of the data. In our algorithm, the input data partitioning is the natural decomposition technique because the output (the computed distances) is not clearly known a-priori. It divides the data set equally according to the number of worker processes by sending a one data partition for each of them [1].

3.2 Mapping Technique

Once a problem has been decomposed into concurrent tasks, these must be mapped to processes (that can be executed on a parallel platform) [1].

In our algorithm, we use the static mapping technique that distributes the tasks among processes prior to the execution of the algorithm.

The scheme for this static mapping is mapped based on data partitioning because our data represented in a two-dimensional array. So, the most suitable scheme used for distributing the two-dimensional array among processes is the row-wise *1-D* block array distribution that distributes the array and assign uniform contiguous portions of the array to different processes [1].

According to the previous selected decomposition and mapping techniques, the suitable parallel algorithm model is the master-slave model in which the master process generates the work and assigns it to worker processes.

3.3 Communication Operation

In most parallel algorithms, processes need to exchange data with other processes; in our algorithm the master process divide the data set according to the number of workers and sending a one data partition for each of them with the *j* and query-instance values. Also, the worker processes send the *j-th* ordered list to the master which include the *j-th* distances and classes.

The suitable communication operation for our algorithm is the scatter operation (one-to-all personalized communication), in which the master process sends a unique part of the divided data to each of the worker processes [1].

The dual of one-to-all personalized communication or the scatter operation is the gather operation, or concatenation, in which the master process collects a unique *j-th* ordered list from each other worker processes [1].

IV. The Proposed Method

Since the computation of the distance between the input sample and any single training sample is independent of the distance computation to any other sample, that allows for partitioning the computation work with the least synchronization effort. In fact, no intercommunication or message passing is required at all during the time each processor is computing the distance between samples in its local storage and the input sample. When all processors terminate the distance computation procedure, the final step is to select a master processor to collect the results from all processors, sort the distances in ascending order, and then use the first *j* measures to determine the class of the input sample.

The proposed algorithm is described in the following steps:

1. Reading the training data set, *j* value and the test object attribute values. Then, determining one process as a master process and the remaining processes as workers.
2. The master process divides the data set equally according to the number of workers by sending a one data partition for each of them with the *j* and query-instance values.
3. Each worker process receives its data partition and the other parameters then:
    a. Calculate the distance between the query-instance and all the training samples.
    b. Sort the distance and determine nearest neighbors based on the *j*-th minimum distance locally.
    c. Send the *j*-th ordered list to the master which include the *j*-th distances and classes.
4. The master process receives from each worker the *j*-th ordered list and combining them in a *j*-th master list.
5. The master now:
    a. Sort the *j*-th master list elements in ascending order.
    b. Select the *j*-th top element.
    c. Compute the majority of classes in the top *j*-th element.
    d. Define the query-instance according to the major class.

Figure 1 exhibits the proposed solution:

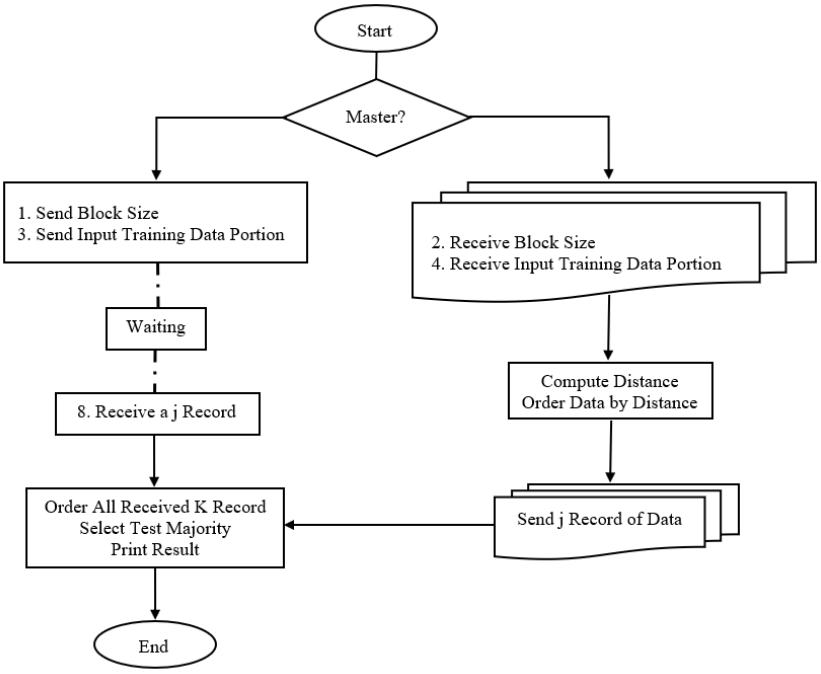


Fig. 1 The Flow chart of the proposed solution.

V. P-KNN Implementation and Experimental Results

We implemented the sequential and parallel versions of the K-NN algorithm and apply our experiments on graduate student's dataset collected from the University College of Science and Technology in Khan Younis. Also, we compare the results between the sequential and parallel versions of the K-NN algorithm, then determine the time in the two cases, and we calculate the speed up, efficiency, parallel overhead, and parallel cost.

For our experiment, the tests are done in a cluster which has consisted of 8 workstations, all workstations have the same specification; Intel(R) Core (TM)2 Quad CPU Q8300 @ 2.50 GHz, 4.00 GB RAM, 320 GB hard disk drive and Windows 7 operating system installed, using the parallel message passing software MPICH2

5.1 Data Collection

The dataset used in this paper contains graduate student's information collected from the University College of Science and Technology – Khan Younis for a period of fifteen years in the period from 1993 to 2011. The graduate student's dataset consists of 17000 records and 18 attributes. Each record belongs to 1 of 5 categories (Excellent – Very Good – Good – Acceptable – Fail). Table I presents the attributes and their description that exist in the data set as taken from the source database.

TABLE I
THE GRADUATE STUDENT'S DATASET DESCRIPTION

Attribute	Description	Data Type	Selected
ID	An identifier for the record	Number	
Name	Student's named	String	
DOB	Date of birth	String	
Gender	Student's gender	String	√
Nationality	Student's nationality	String	
City	Student's address details	String	√
GS Source	Source general secondary	String	
GS Year	Year general secondary	Number	
GS Avg	Average general secondary	Number	√
GS Sec	Section general secondary	Number	√
YI Join	Year institution join	Number	
YI Term	Year institution term	Number	
Std Level	Student's level	Number	
College	Student's college	String	
Specialization	Student's specialization	String	√
HC Num	Hours Completed number	Number	
GPA	Student's a cumulative grade point average	Number	
Grade	Student's performance	String	√

As part of the preparation and pre-processing of the data set, irrelevant and weakly relevant attributes should be removed. The attributes marked as selected as seen in Table 1 are processed via the rapid miner software to apply the data mining methods on them.

5.2 Experimental Results and Discussion

We execute the K-NN (sequential/parallel) program on the training data set with varying number of processes and problem size to evaluate the performance. We run our experiment on 3 to 8 processors, and a problem size with 5000, 13000, and 17000 records, then we compared them with the sequential version. For sequential method, we compute the time using a clock () method from time.h library. For the parallel method we use MPI_Wtime() from mpi.h library that returns an elapsed time on the calling processor in seconds. The execution time in seconds is recorded in Table II.

TABLE II
THE EXECUTION TIME OF THE SEQUENTIAL AND PARALLEL CLASSIFIERS.

Problem Size / PNum		5000	13000	17000
Sequential K-NN		0.07800	0.37500	0.56200
PKNN with No. of processes	*3-process*	0.04610	0.16076	0.22337
	4-process	0.03740	0.15242	0.15924
	5-process	0.03285	0.14024	0.13518
	6-process	0.03093	0.12872	0.11934
	7-process	0.03008	0.11769	0.10394
	8-process	0.02814	0.10967	0.09854

As we note from Table II, the sequential version takes more time than the parallel version. In the parallel version; the execution time decreases when the number of processes increases. However, the parallel implementation achieves a good execution time compared to the sequential one. Figure 2 illustrates the execution time. The time curve decreases from 3 processors until use 8 processors

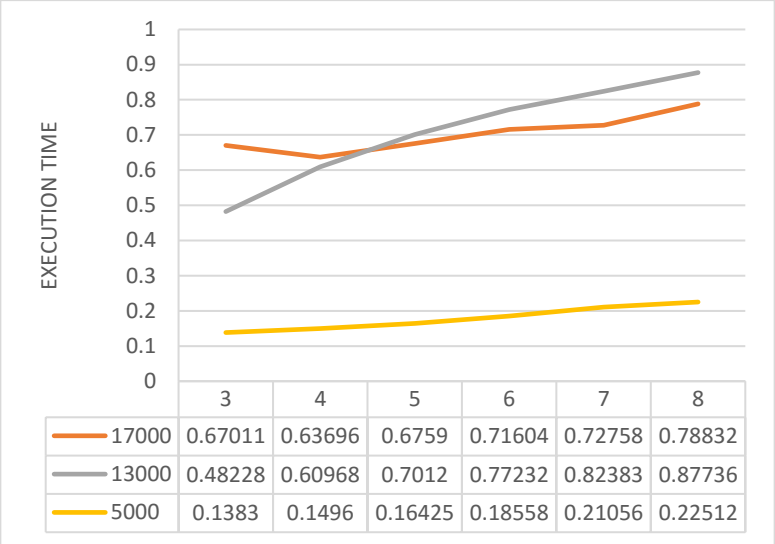


Fig. 2 The Curves of Execution Time for the Classifiers.

Also, we compute Speedup (S) which gained from this parallelization by using equation 1. Speedup values are recorded in Table III.

$$S = T_S / T_P \qquad (5.1),$$

where T_S is serial time and T_P is parallel time.

TABLE III
THE RELATIVE SPEEDUP OF THE PROPOSED PARALLEL CLASSIFIER.

Problem Size / PNum	5000	13000	17000
3-process	1.69197	2.33267	2.51600
4-process	2.08556	2.46031	3.52926
5-process	2.37443	2.67399	4.15742
6-process	2.52182	2.91330	4.70923
7-process	2.59309	3.18634	5.40697
8-process	2.77186	3.41935	5.70327

As we note from Table III, when increasing the number of processors, the Speedup values tend to be saturated, also we note that the Speedup values less than the number of processing elements. Figure 3 illustrates the speed up curve.

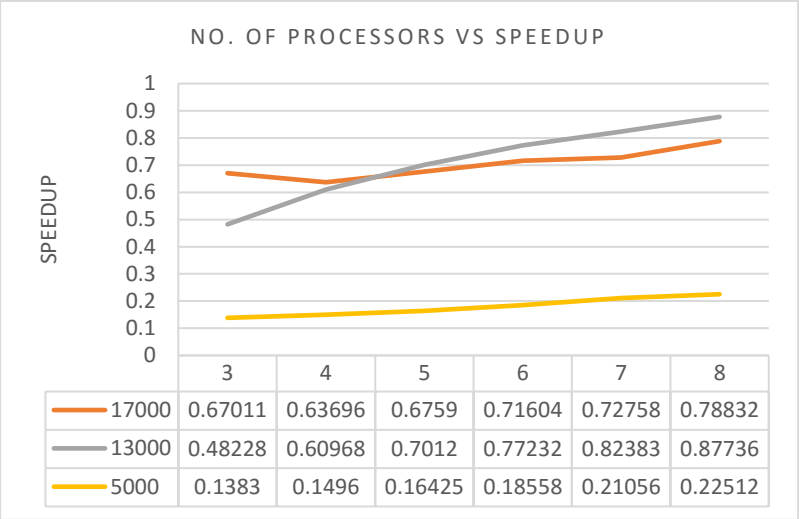


Fig. 3 The Relative Speedup Curves of the Proposed Parallel Classifier.

From the Speedup computation, we can compute the efficiency (E) using equation 2. Efficiency values are recorded in Table IV.

$$E = S/P \qquad (5.2),$$

where S is the Speedup and P is the number of processing elements.

TABLE IV
THE EFFICIENCY OF THE PROPOSED PARALLEL CLASSIFIER.

PNum \ Problem Size	5000	13000	17000
3-process	0.56399	0.77756	0.83867
4-process	0.52139	0.61508	0.88232
5-process	0.47489	0.53480	0.83148
6-process	0.42030	0.48555	0.78487
7-process	0.37044	0.45519	0.77242
8-process	0.34648	0.42742	0.71291

As we note from Table IV, the values of the efficiency are between 0 and 1 and we note that the efficiency decreases as the number of processing elements is increased for a problem size to 5000, and 13000 records and this is common to all parallel programs, but the efficiency where the problem size is 17000 increased when we use 4 processing elements, and after that is decreased, so we can induce that the suitable number of processing elements of our problem is 4. Also, we note that the efficiency increases if the problem size is increased while keeping the number of processing elements constant. Figure 4 illustrates the efficiency curve.

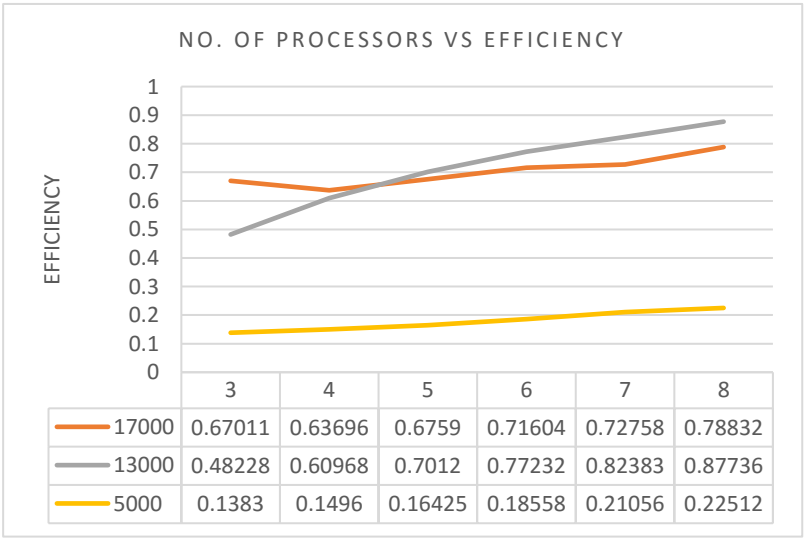


Fig. 4 The Efficiency Curves of the Proposed Parallel Classifier.

Also, we compute parallel Overhead (TO) which is the total time to spend by all processors combined in non-useful work by using equation 3. Overhead values are recorded in Table V.

$$\text{parallel Overhead} = PT_P - T_S \quad (5.3),$$

where T_S is serial time, T_P is parallel time and P is the number of processing elements

TABLE VI
THE SERIAL AND PARALLEL COST.

	Problem Size / PNum	5000	13000	17000
Serial cost		0.07800	0.37500	0.56200
Parallel cost	*3-process*	0.13830	0.48228	0.67011
	4-process	0.14960	0.60968	0.63696
	5-process	0.16425	0.70120	0.67590
	6-process	0.18558	0.77232	0.71604
	7-process	0.21056	0.82383	0.72758
	8-process	0.22512	0.87736	0.78832

As we note from Table V, when increasing the number of processors, the values of the parallel overhead increases for a problem size with 5000, and 13000 records, but the parallel overhead where the problem size is 17000 decreased when we use 4 processing elements, and after that is increased. Figure 5 illustrates the parallel overhead curve.

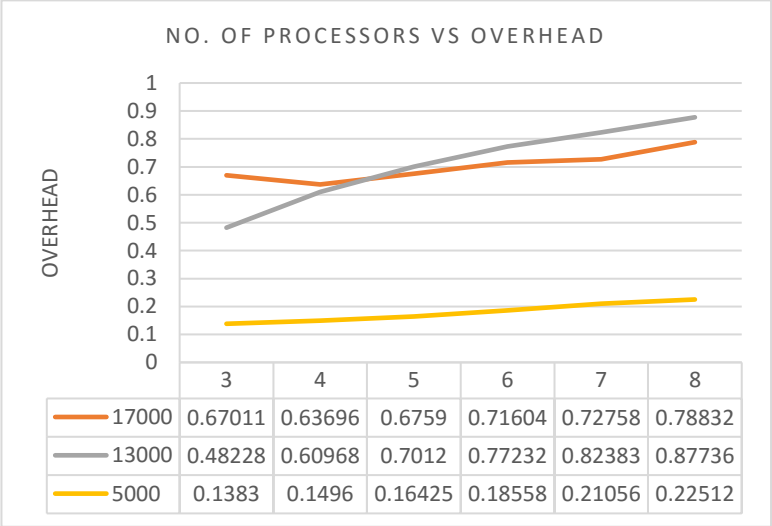


Fig. 5 The Parallel Overhead of the Proposed Parallel Classifier.

Finally, we compute the parallel cost which is the total time collectively spent by all the processing elements using equation 4. Serial & parallel cost values are recorded in Table VI.

$$\text{parallel Cost} = PT_P \quad (5.4),$$

where T_P is parallel time and P is the number of processing elements.

TABLE V
THE PARALLEL OVERHEAD OF THE PROPOSED PARALLEL CLASSIFIER.

Problem Size / PNum	5000	13000	17000
3-process	0.06030	0.10728	0.10811
4-process	0.07160	0.23465	0.07496
5-process	0.08625	0.32620	0.11390
6-process	0.10758	0.39732	0.15404
7-process	0.13256	0.44883	0.16558
8-process	0.14712	0.50236	0.22632

As we note from Table VI, when increasing the number of processors, the values of the

parallel cost increases for a problem size to 5000, and 13000 records, but the parallel cost where the problem size is 17000 decreased when we use 4 processing elements, and after that is increased. Figure 6 illustrates the parallel cost curve.

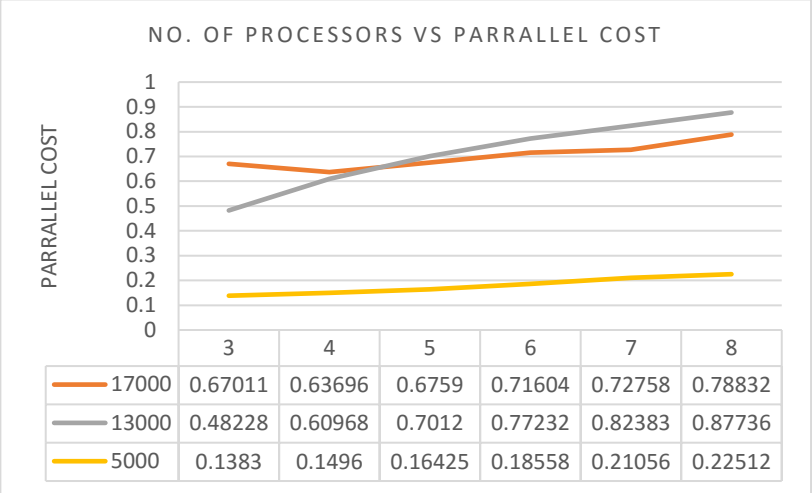


Fig. 6 The serial and parallel cost curve.

To ensure that the classifier works well with the tested records, we also examined the quality of the classification. We split the dataset into two parts (90% of the dataset for training and the remaining 10% to test).

For the purpose of evaluating the classification results, we use confusion matrices that are the primary source of performance measurement for the classification problem. Each column of the confusion matrix represents the instances in an actual class, while each row represents the instances in a predicted class. The K-NN classifier obtains 54.37% accuracy as the best results the number of neighbors is 5.

VI. Conclusion

In this paper, we parallelize the K-NN classification algorithm using message passing routines. We applied our experiments on (5000, 13000, and 17000) × 6 data matrices, by use 3 to 8 processors, and compared them with the sequential version by calculating the execution time for them, we observe in parallel version that the execution time decreases when the number of processes increases. However, the parallel implementation achieves a good execution time compared to the sequential one. We note that when compute Speedup from serial time and parallel time, its values tends to be saturated, also we note that the values less than the number of processing elements. We note that when compute the efficiency, its values are between 0 and 1 and we note that the efficiency decreases as the number of processing elements is increased for a problem size to 5000, and 13000 records and this is common to all parallel programs, but the efficiency where the problem size is 17000 increased when we use 4 processing elements, and after that is decreased, so we can induce that the suitable number of processing elements of our problem is 4. Also, we note that the efficiency increases if the problem size is increased while keeping the number of processing elements constant. Also, we note when increased the number of processors, the values of the parallel overhead increased for a problem size with 5000, and 13000 records, but the parallel overhead where the problem size is 17000 decreased when we use 4 processing elements, and after that is increased.

Finally, we note when increasing the number of processors, the values of the parallel cost increases for a problem size with 5000, and 13000 records, but the parallel cost where the problem size is 17000 decreased when we use 4 processing elements, and after that is increased,

also, we note that from [3] and [4] MPI is not so good as CUDA or using a computer with a multicore microprocessor.

References

[1] Grama A., Gupta A., Karypis G. and Kumar V., "*Introduction to Parallel Computing*", 2nd edition, Addison Wesley, 2003.

[2] Han, J. and Kamber, M. "*Data Mining: Concepts and Techniques*", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, 2006.

[3] Lianga, S., Liua, Y., Wangb, C., and Jiana, L.CUKNN: "*A parallel Implementation of k-Nearest Neighbor on Cuda-Enabled GPU*", The 2009 IEEE Youth Conference on Information, Computing and Telecommunication (ICT2009) – Conference Proceedings, pp. 415-418, 2009.

[4] Tarhini, A. "*Parallel k-Nearest Neighbor*",[Online], Available: http://alitarhini.wordpress.com/2011/02/26/parallel-k-nearest-neighbor, 2011.

[5] Yiqun, C. , "*Parallel and Distributed Techniques in Biomedical Engineering*", M.Sc. Dissertation, Department of Electrical Computer Engineering, National University of Singapore, 2005.

[6] Zufrin, R. , "*Decision trees on parallel processors*", The International Joint Conference on Artificial Intelligence (IJCAI 1995) – Conference Proceedings, Montreal, Canada, August,1995.

Remote sensing satellite virtual constellation optimizing with target recognition probability

Alexsandr M. Kondratov, Oleg V. Maslenko

***Abstract*—**The improved method of satellite systems selection for the electro-optical observation of the specified objects and regions of the Earth is described. The method provides the probability of targets correct detection (recognition). Other indicators of system selected efficiency, which reflect the controlled area, the timeliness of space imagery and the cost should not be less than a predetermined threshold.

***Keywords*—**Earth observation, remote sensing satellite system, virtual constellation, target recognition probability.

I. INTRODUCTION

Currently, the market of information services in the field of remote sensing of the Earth is developing rapidly. Applications for satellite imagery of the necessary regions of the Earth are realized by the electro-optic imaging satellite systems. Satellite images are distributed by dealers that provide their services in Internet. They offer both new-ordered and archive images for anywhere Earth territory. When ordering the satellite images (both new and archive), a great number of parameters that affect the efficiency of the mission accomplishment should be taken into account.

The main elements of any satellite system for electro-optical surveillance are spacecraft and imaging equipment mounted on them. The satellite system selection depends on the geographical, geometrical and physical characteristics of areas and objects of interest; imaging conditions (time of year, time of day, state of the atmosphere, cloudiness level); technical specifications of the imaging equipment; spacecraft position in space and time in relation to the areas of interest and data reception points; state affiliation of the satellite system owner; information requirements (accuracy, reliability, timeliness of acquisition, cost). Therefore, special algorithms are required for optimal solving the problem of a suitable satellite system selection.

The most common way to forecast the efficiency of the remote sensing task solution is the service simulation of satellite system application [1]. During the simulation the swath of Earth's surface and their positioning on the area of interest are computed. The time intervals when imaging is possible are calculated. The orbits by which the spacecraft passes over the area of interest at a specified time are selected. The illuminance conditions of the area of interest, the viewing angle inclination, imaging repeat cycle, the cost of a satellite imagery scene and other parameters are determined [2, 3]. Thereafter, the problem of the most suitable satellite systems selection is solved. By comparison, the following indicators of the efficiency of satellite observation systems are analyzed: the time of data obtaining and the information renewal rate; productivity of a satellite system that is specified as the total area imaged per day; spatial resolution of the satellite image; positional accuracy of the captured images in geodetical coordinate system, the geometric distortions of image, etc. [4]. Most of the available techniques for satellite observation planning rely on sophisticated and accurate models of the spacecraft orbital motion, however, as a rule, without consideration the probabilistic characteristics of the target detection and recognition [5].

II. METHOD

To determine the expected time of satellite imaging, it is necessary to set the geographical coordinates of the area of interest and to define the moments when the path of viewing axis

crossing the area boundaries [6, 7]. Initial data are the Kepler elements of solar synchronous orbits. The inclination of the viewing axis for the roll η and the current moment of the spacecraft flying time t_j, $j = 0, 1, 2, \ldots$ are considered to be known. Required values are the geographic coordinates of the observation point: geographic latitude $\varphi_j = \varphi(t_j)$ and longitude $\lambda_j = \lambda(t_j)$ at specific time interval t_j within the boundaries of the orbit selected. At first, the coordinates of the viewing point are calculated in accordance with a spacecraft coordinate system. Then, the coordinate system is converted to the other one, in which the geographic coordinates of the viewing point are calculated. Further, the path of the viewing axis is designed as a set of the viewing points with allowance for the flying time. Current coordinates of the viewing points in a geocentric spherical coordinate system at a time point t_j [6, 8] are following:

$$\varphi_j = \arcsin\ (\sin u_j \sin i \cos\psi_\eta + \cos i \sin\psi_\eta) \tag{1}$$

$$\lambda_j = \pm\arccos\left(\frac{(a_{11}\cos\psi_\eta + a_{31}\sin\psi_\eta)\cos\theta_ç + (a_{12}\cos\psi_\eta + a_{32}\sin\psi_\eta)\sin\theta_ç}{\sqrt{(a_{11}\cos\psi_\eta + a_{31}\sin\psi_\eta)^2 + (a_{12}\cos\psi_\eta + a_{32}\sin\psi_\eta)^2}}\right) \tag{2}$$

where ψ_η is the geocentric angle between the radius-vector of a spacecraft and the radius-vector of a viewing point, a_{11}, a_{12}, a_{31}, a_{32} are the elements of transition matrix from the inertial to the Greenwich geocentric coordinate system.

The calculation of a latitude argument at a time point t_j is carried out according to the formula:

$$u(t_j) = \sqrt{\frac{\mu_0}{a^3}}\Delta t_j \tag{3}$$

where $\Delta t_j = t_j - t_\Omega$ is the time interval since the moment the spacecraft was located in the ascending orbit.

An optical axis path makes it possible to set the time when satellite survey begins as a time point when the path coordinates match the geographical coordinates of the area of interest at a certain orbit.

Inasmuch as the area of interest is defined as a trapezoid mostly, the start of satellite survey can be set as [9]:

$$t_m^{\text{sa}}(n) = \begin{cases} t_m^S(n), \text{ if } (\varphi_j = \Phi_m^S) \wedge (\Lambda_m^W \le \lambda_j \le \Lambda_m^E) \wedge (S \to N) = 1 \\ t_m^W(n), \text{ if } (\lambda_j = \Lambda_m^W) \wedge (\Phi_m^S \le \varphi_j \le \Phi_m^N) \wedge (S \to N) = 1 \\ t_m^N(n), \text{ if } (\varphi_j = \Phi_m^N) \wedge (\Lambda_m^W \le \lambda_j \le \Lambda_m^E) \wedge (N \to S) = 1 \\ t_m^E(n), \text{ if } (\lambda_j = \Lambda_m^E) \wedge (\Phi_m^S \le \varphi_j \le \Phi_m^N) \wedge (N \to S) = 1 \end{cases} \tag{4}$$

where Φ_m^S and Φ_m^N are the southern and northern latitudes of the spherical trapezium sides of the *m*-th region, Λ_m^W and Λ_m^E are the western and eastern longitude of the spherical trapezium sides of the *m*-th region, $t_m^S(n)$, $t_m^N(n)$ are the time points when the sighting axis path crosses the southern and northern boundaries of the *m*-th region at the *n*-th orbit of the spacecraft, $t_m^W(n)$, $t_m^E(n)$ are the time points when the sighting axis path crosses the western and eastern boundaries of the *m*-th region at the *n*-th orbit of the spacecraft, $S \to N$ is the spacecraft orbital motion from south to north, and $N \to S$ from north to south.

The region area *S* defined as the spherical trapezoid can be calculated [9] by formula:

$$S(\Phi, \Lambda) = R^2\,(\Lambda_m^E - \Lambda_m^W) \cdot (\sin\Phi_m^N - \sin\Phi_m^S) \tag{5}$$

where *R* is the Earth's radius.

The instantaneous projection area of the imaging coverage in case of the viewing axis

inclination from the nadir in relation to the flat Earth's surface can be calculated using a formula:

$$S(\eta) = H^2 \operatorname{tg}\beta\,[tg(\alpha+\eta)+\operatorname{tg}(\alpha-\eta)]\cdot[\sec(\alpha+\eta)+\sec(\alpha-\eta)] \quad (6)$$

The presence or absence of visibility conditions of the μ-th spacecraft for the *m*-th region during the satellite observation can be determined [6] using a logic function

$$\Phi_m^F(n_\mu)=\begin{cases}1,\text{if } (K_m^S(\mu)\ge\overline{K}_m^S)\wedge(K_m^T(\mu)\ge\overline{K}_m^T)\wedge(\beta_m^c(\mu)\ge\overline{\beta}_m^c)\wedge(Q_m^\xi(\mu)\le\overline{Q}_m^\xi)=1\\ 0,\text{if } (K_m^S(\mu)\ge\overline{K}_m^S)\wedge(K_m^T(\mu)\ge\overline{K}_m^T)\wedge(\beta_m^c(\mu)\ge\overline{\beta}_m^c)\wedge(Q_m^\xi(\eta)\le\overline{Q}_m^\xi)=0\end{cases} \quad (7)$$

where $K_m^S(\mu)$ is the spatial coefficient of the *m*-th region coverage by the swath of the μ-th spacecraft, which should not be less than the pre-defined valid value K_m^S. It can be found as a ratio of the expected area of the imaging of the *m*-th region by μ-th the spacecraft to the total area of the region:

$$K_m^S(\mu) = S_m(\mu)/S_m \;,\quad K_m^S(\mu)\to\max \quad (8)$$

Secondly, $K_m^T(\mu)$ is the time coefficient of the *m*-th region coverage by the swath of the μ-th spacecraft, which should not be less than the pre-defined valid value $\overline{K}_m^T$.

The conditions of the temporal cover of the *m*-th region by the swath of the μ-th spacecraft can be submitted [6] as

$$[T_m(\mu)\in\overline{T}_m]\wedge[T_m(\mu)\ge\overline{T}_m]=1 \quad (9)$$

where $T_m(\mu)=t_m^{end}(\mu)-t_m^{start}(\mu)$ is the duration of an expected imaging interval, $\overline{T}_m=\bar{t}_m^{end}-\bar{t}_m^{start}$ is the duration of a specified imaging interval, $t_m^{end}(\mu)$, $\bar{t}_m^{end}$ are the expected and specified imaging termination time, $t_m^{start}(\mu)$, $\bar{t}_\grave{o}^{start}$ are the expected and specified imaging start time.

The temporal coefficient of coverage can be found as the ratio of the expected duration of the *m*-th spacecraft over the μ-th area $T_m(\mu)$ to the specified time of visibility of the area:

$$K_m^T(\mu) = T_m(\mu)/\overline{T}_m \;,\quad K_m^T(\mu)\to\max \quad (10)$$

Thirdly, $\beta_m(\mu)$ is the current of the Sun's elevation angle for the satellite survey of the *m*-th region by the μ-th spacecraft, which needs to meet the requirement $\beta_m(\mu)\ge\overline{\beta}_m$, where $\overline{\beta}_m$ are the permissible minimum angle of the Sun's elevation [6]:

$$\beta_m(\mu)=\begin{cases}\dfrac{(t_{local}-t_{sunset})\,\beta_{\max}}{12-t_{sunset}} & \text{if } t_{local}<12^h\\ \dfrac{(24-t_{local}-t_{sunset})\,\beta_{\max}}{12-t_{sunset}} & \text{if } t_{local}\ge12^h\end{cases} \quad (11)$$

where t_{local} is the current local time, t_{sunset} is local time of sunset; $\beta_{\max}$ is the maximum angle of the Sun's elevation within the area of interest.

In the fourth place, $Q_m^\xi(\mu)$ is the forecasted cloud-covered area of the *m*-th region for the μ-th spacecraft, taking into account the coefficient of transparency of the atmosphere ξ, which should be no less than the valid value $\overline{Q}_m^\xi$. Information about clouds over certain areas of the Earth can be obtained from the world meteorological services or other relevant institutions.

The expected cost of the image of the observed part of the *m*-th region, acquired by the μ-th spacecraft $C_m(\mu)$, can be estimated using the following data: the area of the *m*-th region observed by the μ-th spacecraft $S_m(\mu)$; commercial offers of the Earth observing systems

operators concerning the minimum scene size in order $S_{\min}$; cost of an image C_1 per 1 sq. km; threshold area S_n, which exceeds the operator's ability to reduce the cost C_1; the cost of image acquisition C_2 per 1 sq. km with a discount ($C_2 \le C_1$).

Then the total image cost can be calculated by the following formula:

$$C_m(\mu) = \begin{cases} C_1 S_{\min} \text{ if } S_m(\mu) < S_{\min} \\ C_1 S_m(\mu) \text{ if } S_n > S_m(\mu) \ge S_{\min} \\ C_2 S_m(\mu) \text{ if } S_m(\mu) \ge S_n \end{cases} \quad (12)$$

Then the normalized cost factor of the satellite image should be used in the form

$$K_m(\mu) = \begin{cases} C_m(\mu) / C_m^{\max} \text{ if } C_m(\mu) < C_m^{\max} \\ 1 \text{ if } C_m(\mu) \ge C_m^{\max} \end{cases} \quad (13)$$

where $C_m^{\max}$ is the maximum cost of the image of the *m*-th area acquired by available spacecraft.

The coefficient (13) can take a range of values (14)

$$0 \le K_m(\mu) \le 1 \quad (14)$$

At the same time, the following term should be compiled with

$$K_m(\mu) \to \min \quad (15)$$

and the selection of suitable satellite system should be carried in accordance with smallest values of the coefficient (15) from the ordered set

$$K_m(1) < K_m(2) < \ldots < K_m(\mu) \quad , \quad \mu = \overline{1, \mathrm{M}} \quad (16)$$

The probability of correct detection (recognition) of the target $P(x, \theta)$ can be described by a generalized relation

$$P(x, \theta) \cong 1 - \prod_i \varepsilon(x_i, \theta_i) \quad (17)$$

where x is the input vector of optical signal, θ is the set of parameters, $\varepsilon(x, \theta)$ is the probability of error. The probability of error can be written as [10]

$$\varepsilon(x, \theta) \cong 1 - \Phi\left(\frac{\Delta x \sqrt{n}}{2\sigma} \right) \quad (18)$$

where $\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}} du$ is the probability integral [11], Δx is the difference between target and background optical signals, σ is the standard deviation of the optical signals, n is the number of resolution elements within the target image.

Equation (17) describes the process of object detection. In passing to recognition, the dependence $\varepsilon(x, \theta)$ becomes more complicated. This fact can be taken into account using the Johnson criteria [12] or any other model that describes image recognition to the required information level.

III. RESULT

In this research, the model of satellite systems selection for the electro-optical imaging has been applied. For this purpose, the satellite survey of a specified area has been simulated. The area of interest was specified as shown in Fig. 1.

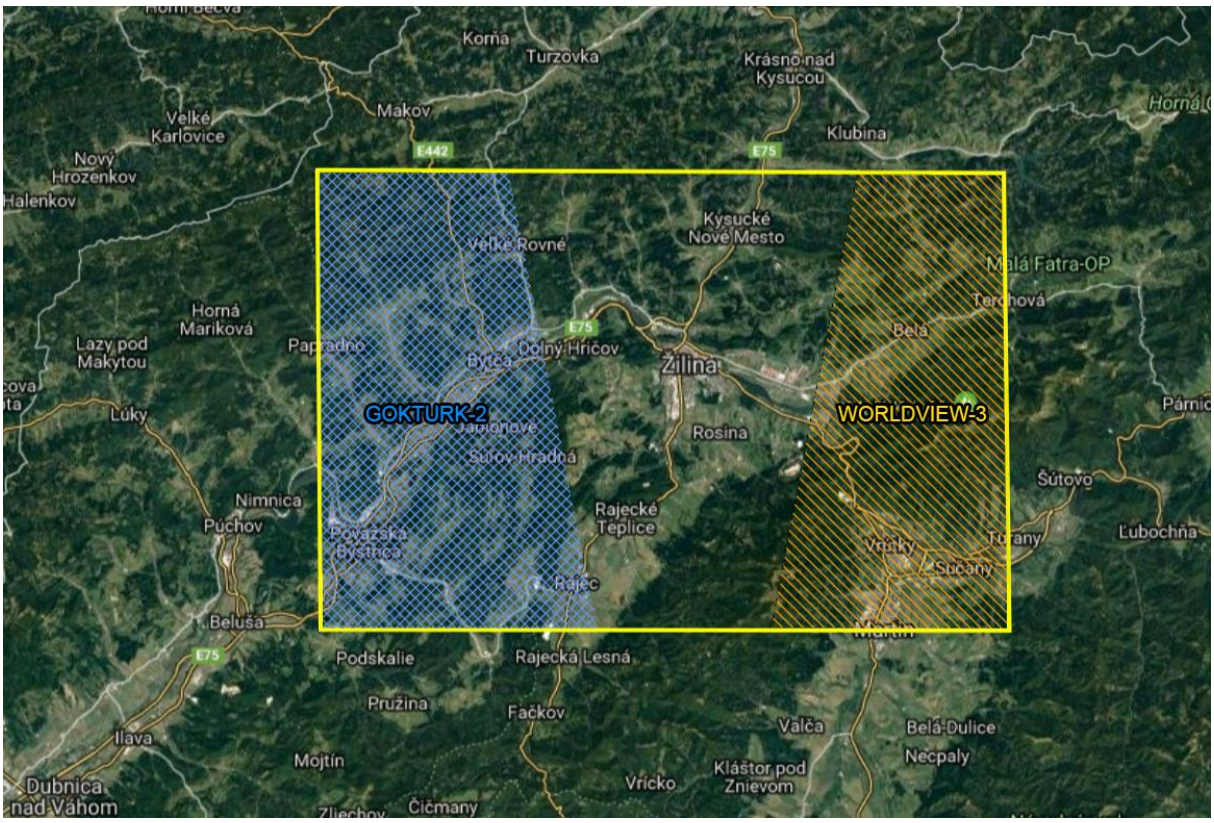


Fig. 1. The area of interest for satellite system selection

Also the some satellite system's swaths are plotted within the area of interest. The simulation's results are presented in the Table I.

TABLE I
THE RESULTS OF THE SATELLITE SYSTEMS EVALUATION

Time frame: form 6:00 a.m., 3 January 2019 till 6:00 a.m., 13 January, 2019			Time required to cover the region completely by one spacecraft swath	
Satellite system	Detection probability	Recognition probability	Hours	Days
WorldVie-1	0,9978	0,9912	122,4	5,1
Geoeye 1	0,9985	0,9941	123,912	5,163
Worldview-2	0,9981	0,9925	123,888	5,162
SPOT 6	0,9651	0,8681	3,24	0,135
GokTurk 2	0,9461	0,8017	74,208	3,092
KompSat 2	0,9912	0,9653	122,88	5,12
KompSat 3	0,9957	0,9828	77,592	3,233
Pleiades 1A	0,9978	0,9912	52,296	2,179
WorldView-3	0,9991	0,9966	148,344	6,181
KazEOSat 1	0,9912	0,9653	51,816	2,159
KazEOSat 2	0,6874	0,2244	27	1,125
Pleiades 1B	0,9978	0,9912	76,176	3,174
SPOT 7	0,9651	0,8681	27,144	1,131
NigeriaSat 2	0,9461	0,8017	98,832	4,118
Cartosat 2A	0,9963	0,9852	26,928	1,122
Cartosat 2B	0,9963	0,9852	74,904	3,121
DubaiSat 2	0,9912	0,9653	51,072	2,128
Deimos 2	0,9912	0,9653	123,192	5,133
WorldView-4	0,9991	0,9966	76,32	3,18
Sentinel 2A	0,4118	0,0291	28,272	1,178
Sentinel 2B	0,4118	0,0291	3,936	0,164

As it follows from the analysis of Table 1 data, the optimal satellite system for target detection is SPOT 6, and for target recognition is Cartosat 2A. Other satellite systems either do not meet the probability of detection (recognition), or consume more time to capture the entire area of interest.

IV. CONCLUSIONS

The improved method of satellite systems selection for the electro-optical observation of the specified targets and regions of the Earth is presented. The method provides the selection of electro-optical observation satellite systems with allowance for area of interest characteristics, conditions and timeliness of satellite survey, and also the cost of satellite imagery. The fundamental advantage of the method is the providing the probability required for targets correct recognition by the satellite imagery.

REFERENCES

[1] I.A. Glazkova, V.V. Malyshev, and V.V. Darnopykh, "Estimation of perspective micro-satellite Earth observation system efficiency on the base of imitative modeling (in Russian)", *Computer Science and Control*, vol. 16, no. 6, pp. 125-134, June 2009.

[2] *STK User's Guide*. Exton, PA: Analytical Graphics, Inc., 2004, 536 p.

[3] V.M. Vishnyakov, "Optimization of orbital constellation parameters of the satellite system for emergency monitoring (in Russian)", *Current Problems in Remote Sensing of the Earth from Space*, vol. 1, no. 2, pp. 222-237, June 2005.

[4] O.D. Fedorovskyi, M.V. Artiushenko, and Z.V. Kozlov, "Parametric synthesis of space systems for remote sensing of the Earth on the basis of the genetic method (in Russian)", *Space Science & Technology*, vol. 10, no. 1, pp. 54-60, March 2004.

[5] V. Malyshev, and V. Bobronnikov, "Mission planning for remote sensing satellite constellation", in: *Mission Design & Implementation of Satellite Constellations*, ed. by Jozef C. van der Ha, Dordrecht: Kluwer Academic Publishers, 1998, pp. 431-437.

[6] P.V. Friz, *The Spacecraft Orbital Movement Fundamentals* (in Ukrainian), Zhytomyr: Korolev Zhytomyr Military Institute, 2012, 348 p.

[7] *Spacecraft Flight Theory Fundamentals* (in Russian), ed. by G.S. Narimanov, Moscow: Machine Building, 1972, 608 p.

[8] B.S. Skrebushevsky, *Spacecraft Orbits Formation* (in Russian), Moscow: Machine Building, 1990, 256 p.

[9] P.V. Friz, "Improved mathematical apparatus for determining the observed area of the landed earth region in space monitoring problems" (in Ukrainian), *The Journal of Zhytomyr State Technological University. Series: Engineering*, vol. 1, no. 2, pp. 126-134, Feb. 2017.

[10] S.A. Stankevich, "Estimating the linear resolution of digital aerospace images" (in Russian), *Space Science & Technology*, vol. 8, no. 2-3, pp. 103-106, May 2002.

[11] T.T. Soong, *Fundamentals of Probability and Statistics for Engineers*, Hoboken, NJ: Wiley, 2005, 498 p.

[12] T.A. Sjaardema, C.S. Smith, and G.C. Birch, *History and Evolution of the Johnson Criteria*, Oak Ridge, TN: Sandia National Laboratories, 2015, 40 p.

Algorithm of routes optimization for mobile robots

Vladislav Solovtsov, Dzmitry Adzinets

***Abstract* —** The article presents a method of route optimization on resources which should be spent for route passage. The proposed algorithm is based on the idea that rise require more resources than the descent for passage. For flat locality, the result of the algorithm coincides with the result obtained from Google Maps. For hilly and mountainous terrain, the algorithm selects routes that are differ from Google Maps with a lower content of the number of lifts, which accordingly makes routes less energy consuming.

***Keywords* —** energy functional, navigation solution, optimal route.

I. Introduction

Nowadays mobile robots can be used for carrying out explosion works and works in dangerous areas, carrying out works during liquidation of consequences of worst-case situations. Also, robots are widely used in delivery services. Before sending robots to work on the ground, the routes they will follow should be optimized. Existing navigation solutions offer the following route search optimizations: the fastest route and the shortest route [1], [2]. In this article we will propose an algorithm of route optimization on resources which should be spent for route passage.

The target audience of this algorithm is pedestrians, cyclists and drivers.

II. Task statement

It is required to find a route between two points (the route can also include a number of waypoints), the passage of which requires the least amount of resources.

We will proceed from the fact that more resources are spent on a rise than on the descent, then *the least cost route* is the route with the least quantity of lifts, as the rise takes more petrol for drivers as well physical effort for pedestrians and cyclists.

Primary data. Suppose the route includes 4 points, where point #1 is the starting point, point #4 is the end point, points #2 and #3 are waypoints of the route (Fig. 1). There are 3 routes (can be more or less) between each current and next points, in the figure they are marked with green, red and yellow colors.

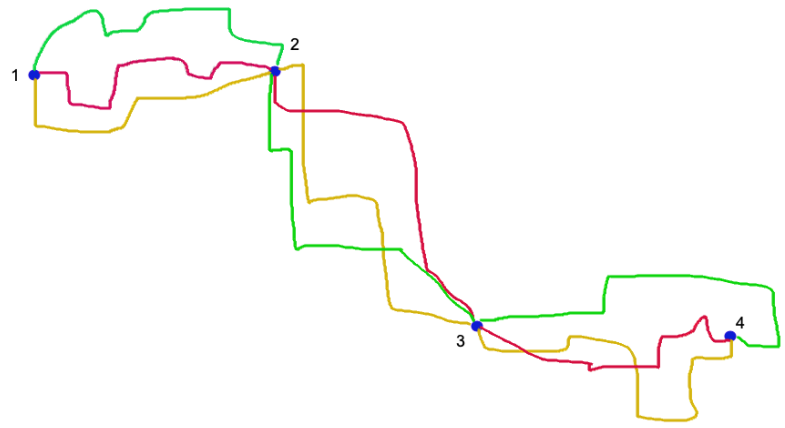


Fig. 1 Route and waypoints, top view
(the figure does not carry any information about the elevation on the ground)

V Solovtsov, Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus (e-mail: solovtsov.vladislav@gmail.com).
D. Adzinets, Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus (e-mail: adzinets@bsuir.by).

Routes between points shown in the figure are obtained using navigation services such as Google Maps/Here Maps/Sygic. In this paper, the service Google Maps will be used as an electronic map data provider. Google Maps Directions API will be used to get routes [3], to do this, HTTP GET request should be sent to the following address:

https://maps.googleapis.com/maps/api/directions/json?parameters

where parameters are query parameters, among them:

— origin, required parameter, departure address;

— destination, required parameter, destination address;

— key, required parameter, API Authorization Key, obtained in Google Developers Console;

— mode, travel mode for route building, one of the DRIVING, WALKING, BICYCLING;

— alternatives, a flag, specifies that the Directions service may provide more than one route alternative in the response.

The value of the alternatives flag must be set to true, because then the least cost route will be selected from the returned main and alternative routes.

Elevation value of any location can be obtained with Google Maps Elevation API [4]. To do this, HTTP GET request should be sent to the following address:

https://maps.googleapis.com/maps/api/elevation/json?parameters

where parameters are query parameters, among them:

— key, required parameter, API Authorization Key, obtained in Google Developers Console;

— locations, the location(s) on the earth (latitude and longitude) from which to return elevation data.

Thus for any point of any route it is possible to get the elevation value. For example, for the green route section between points 1-2 of Figure 1, the chart of the elevation differences may look like in Fig. 2:

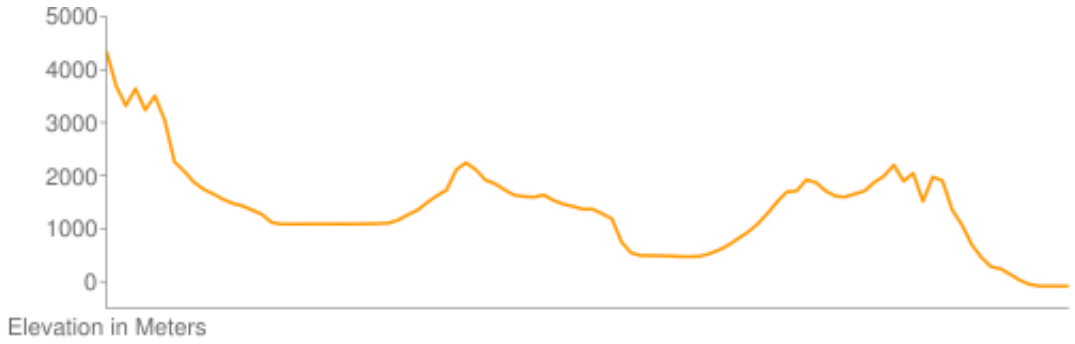


Fig. 2 Chart of the elevation differences for the green route section 1-2

III. Description

The main idea is to assign a weight to each route between each adjacent two points. For this, the route is divided into many sections, for each section the length L, the elevation of the starting point Hstart and the elevation of the end point of the section Hend are known.

Let's introduce the concept of *energy functional*:

$$E = f(a), \tag{1}$$

where α — slope angle of the route. The angle can be calculated by the following formula:

$$\alpha = sin\frac{H_{end} - H_{start}}{L} \tag{2}$$

Let's introduce the concept of *cost functional*:

$$b = f(E, L), \tag{3}$$

where E — previously introduced energy functional, L — section length.

The weight value for the entire route will be calculated as the sum of the cost functional values for each section:

$$B = \sum_{i=1}^{n} b_i \tag{4}$$

Weights are calculated for each route. Thus, each route will correspond to its total weight, proportional to the cost of its passage:

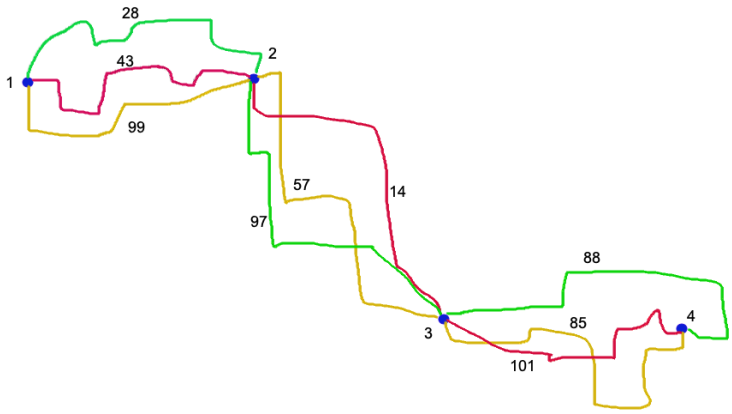


Fig. 3 Routes with calculated weights

After the weights are calculated, we can move from the map (Fig. 3) to the graph where the vertices are the points of the route, and the routes between the points are the edges of the graph. The weight of the route is equal to the edge weight (Fig. 4).

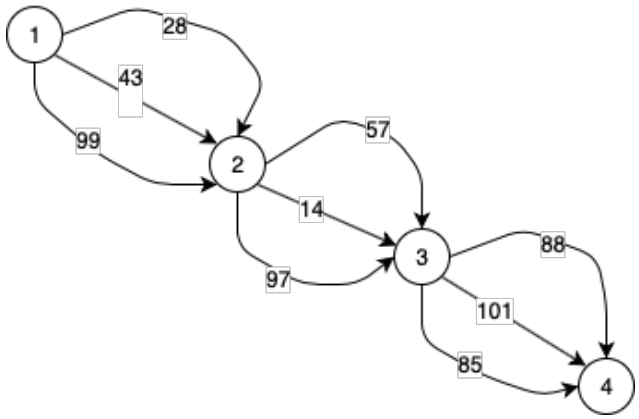


Fig. 4 Route graph

Here are examples of calculations of energy functional and cost functional (Fig. 5, 6):

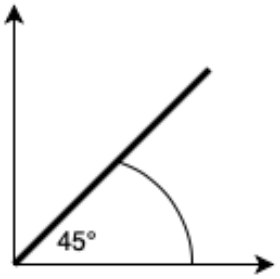


Fig. 5 Route section — rise — has a slope α = 45

For this section the energy functional will be calculated according to the following formula:

$$E = 1 + \frac{a}{100} = 1.45 \tag{5}$$

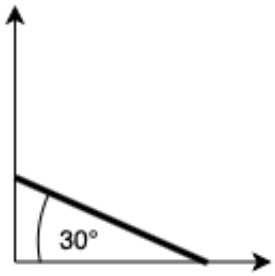


Fig. 6 Route section — declivity — has a slope α = 30°

For this section the energy functional will be calculated according to the following formula:

$$E = 1 - \frac{a}{100} = 0.70 \tag{6}$$

Thus for the rises the functional energy will be calculated by the formula:

$$E = 1 + \frac{a}{100} \tag{7}$$

for descents:

$$E = 1 - \frac{a}{100} \tag{8}$$

In turn, the cost functional will be calculated as:

$$b = f(E, L) = E \cdot L \tag{9}$$

IV. Test Results

An example of the algorithm work is the route searching in Lisbon, Portugal, the beginning of the route — R. Pinheiro Chagas 1050, the end of the route — R. Damasceno Monteiro 1170 (Fig. 7).

Alternative routes and corresponding weights calculated according to the algorithm are presented in table 1.

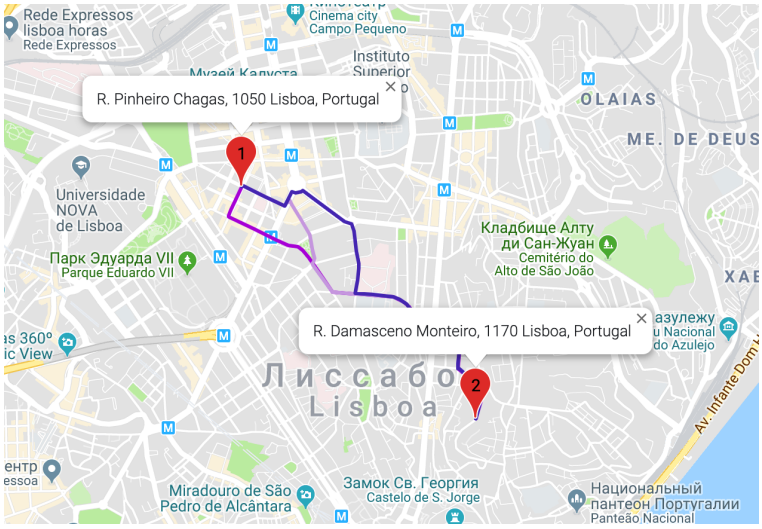


Fig. 7 Routes between waypoints, Lisbon, Portugal

TABLE I
Routes weights, Lisbon, Portugal

#	Start	End	Color	Weight
1	R. Pinheiro Chagas 1050	R. Damasceno Monteiro 1170		2603.28827
2	R. Pinheiro Chagas 1050	R. Damasceno Monteiro 1170		2478.29080
3	R. Pinheiro Chagas 1050	R. Damasceno Monteiro 1170		2492.28877

Thus the algorithm has selected the route with the lowest weight value, namely the route #2 from table 1 (Fig. 8).

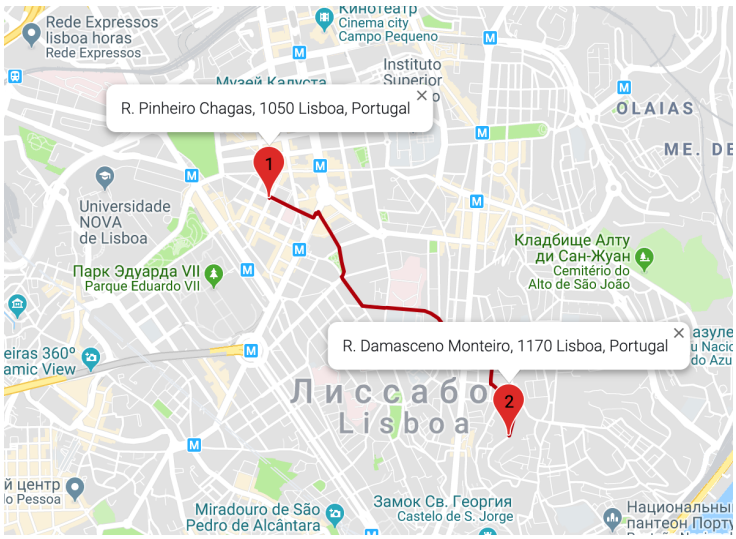


Fig. 8 Route chosen by the algorithm, Lisbon, Portugal

The Route offered by Google Maps navigation service corresponds to route #3 from table 1 and is shown in Fig. 9.

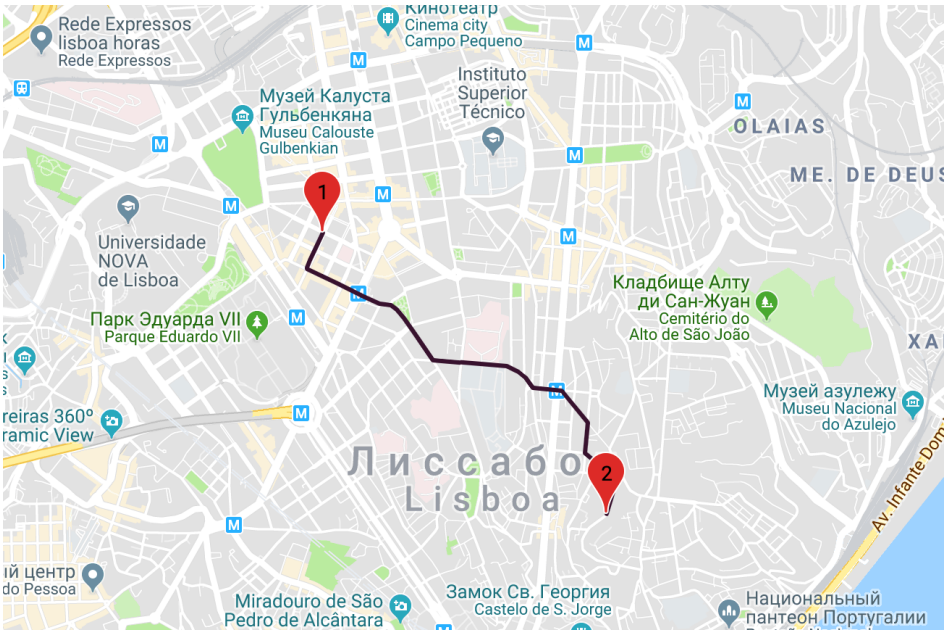


Fig. 9 Route offered by Google Maps Service

Example for a route that includes waypoints is route searching in Tokyo, Japan. The Point of departure is 2-chōme−11−4, 105-0012 Tōkyō-to, Minato City, Shibadaimon, the destination point is 2-chōme-4-12 Kyōbashi, Chuo City, Tōkyō-to 104-0031, and the waypoint is 2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013 (Fig. 10).

Fig. 10 Routes between waypoints, Tokyo, Japan

Alternative routes and corresponding weights calculated according to the algorithm are presented in table 2.

Based on the data in table 2, the search route will consist of two parts, route #1 and route #5, as these routes have the smallest weights for their sections of the path. Searching route is shown in Fig. 11.

The route offered by Google Maps navigation service corresponds to route #3 and route #4 from table 2 and is presented in Fig. 12.

TABLE II
Routes weights, Tokyo, Japan

#	Start	End	Color	Weight
1	2-chōme−11−4, Shibadaimon, Minato City, Tōkyō-to 105-0012	2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013		5025.43739
2	2-chōme−11−4, Shibadaimon, Minato City, Tōkyō-to 105-0012	2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013		5071.41868
3	2-chōme−11−4, Shibadaimon, Minato City, Tōkyō-to 105-0012	2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013		5042.46492
4	2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013	2-chōme-4-12 Kyōbashi, Chuo City, Tōkyō-to 104-0031		2306.16193
5	2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013	2-chōme-4-12 Kyōbashi, Chuo City, Tōkyō-to 104-0031		2235.16115
6	2-chōme-2 Kasumigaseki, Chiyoda City, Tōkyō-to 100-0013	2-chōme-4-12 Kyōbashi, Chuo City, Tōkyō-to 104-0031		2280.21866

Fig. 11 The route chosen by algorithm

Fig. 12 The route proposed by Google Maps service

V. Conclusion

By analyzing finished research, it can be concluded that the result of the algorithm is better than the similar solution from Google Maps for hilly and mountainous terrain. For flat locality, the result of the algorithm coincides with the results obtained from Google Maps. For hilly and mountainous terrain, the described algorithm selects routes that are differ from Google Maps with a lower content of the number of lifts, which accordingly makes routes less energy consuming.

References

[1]. Tarasyan V., Polushkin A. (2017) OPTIMIZATION OF THE PATH IN THE INHOMOGENEOUS ENVIRONMENT // Fundamental research. Retrieved from http://www.fundamental-research.ru/ru/article/view?id=41828

[2]. Improving Operations with Route Optimization – Towards Data Science (2018). Retrieved from https://towardsdatascience.com/improving-operations-with-route-optimization-4b8a3701ca39

[3]. Developer Guide | Directions API | Google Developers (2019). Retrieved from https://developers.google.com/maps/documentation/directions/intro

[4]. Developer Guide | Elevation API | Google Developers (2019). Retrieved from https://developers.google.com/maps/documentation/elevation/intro

Contour Extraction of Noisy Echocardiographic Images Based on Pre-processing

Ahmed S. J. Abu Hammad

***Abstract*—** Contour extraction from two-dimensional echocardiographic images has been a challenge in digital image processing. This is essentially by reason of the heavy noise, poor quality of these images and some artifacts like papillary muscles, intra-cavity structures as chordate, and valves that can interfere with the endocardial border tracking. In this paper, we will introduce a technique to extract the contours of heart boundaries from a sequence of noisy echocardiographic images, where it started with pre-processing to reduce noise and produce better image quality. In order to do this, we combine many pre-processing techniques (filtering, morphological operations, and contrast adjustment) to avoid unclear edges and enhance low contrast of echocardiograph images, after applying these techniques we can obtain legible detection for heart boundaries and valves' movement by traditional edge detection methods.

***Keywords*—**Echocardiography images, Noise reduction, Edge detection.

I. Introduction

Echocardiography is a valuable tool for imaging the heart and reflects the limits of anatomy and heart movement in two-dimensional cardiac sections. It becomes one of the most common ways used to diagnose heart diseases. Automatic boundary extraction from echocardiography images appears as a clinical important need to produce the most effective and reliable results. However, its inherent poor image quality and it have heavy noise.

The major edge detection algorithms fail due to the presence of noise and the low contrast in the heart echocardiograph image and the improvement of echocardiograph images is very important for the accurate detection of both heart boundary and movement of the heart valves. Therefore, noise reduction must be applied before edge detection.

In this paper, we will present a technique for extracting the contours of the heart in different echocardiography images. This technique based on high levels of pre-processing to produce a clear detection for the heart anatomy in echocardiograph images. This technique consists of two main stages of our method, specifying the method in terms of the flowchart. Main stages contain pre-processing stage, which consists of three operations: median filtering, morphological opening, and contrast enhancement to reduce the noise, the second stage applies edge detection and combines two images to get distinct detection. The results provided by our method are shown in section three with a simple description. Some discussion and conclusion are also presented concerning the usefulness of preprocessing to get better detection.

The paper is organized as follows: section 2 presents related works. Section 3 describes our methodology and the two main stages of the method. Section 4 describes the experimental results of the method. Section 5 contains the conclusion.

II. Related Works

In recent years several techniques proposed to reduce the noise without distorting the relevant clinical details. E Boonchieng et al. [1] proposed a method contained three steps: firstly, image improvement algorithms of noise suppression, histogram, brightness adjustment threshold, and median filtering. Second, edge detection with the Sobel algorithm and the third used segmentation and computer graphics algorithms to generate contour lines of echocardiogram border.

Ahmed Abu Hammad, University College of Science and Technology, Khan Younis, Palestine (e-mail: asj_hammad@hotmail.com).

Lacerda et. al. [2] combines classic image processing techniques with Radial search to extract the left ventricular borders from echocardiograph images. High boost filtering and thresholding were used as a pre-processing step, followed by watershed and radial search for segmentation. Then the final contour smoothed by morphological closing.

However, Ries et al [3] presented two methods, where both of them apply a pre-processing filter to reduce noise and increase contrast by using mathematical morphology, high boost filtering, image segmentation, and motion estimation.

Santos et. al [5] analyzed the importance of the pre-processing procedure for border extraction in echocardiograph images, explained image filtering, histogram modification and pre-processing effect in the left ventricular boundary extraction.

Ling et. al. [6] in the meantime modified the combination of morphological operations to avoid the unclear edges of images and presented a comprehensive modified edge detection algorithm of morphological processing as erosion and dilation, also the combination with multi-structure elements to achieve the edge detection of echocardiograph images, wherein this way it could eliminate noise and reduce the fuzzy edge contours effectively.

Almost all approaches for contour extraction use common image processing procedures, the most important ones being pre-processing and edge detection. The main differences between these approaches are at the level of the pre-processing procedures to achieve an improved image and get an effective boundary tracking.

III. Methodology

The proposed method consists of two main stages:

1. The image enhancement part which utilizes the median filtering followed by morphological opening and contrast improvement.
2. The detection part which is used to detect the boundaries of the heart boundaries and the movement of the heart valves.

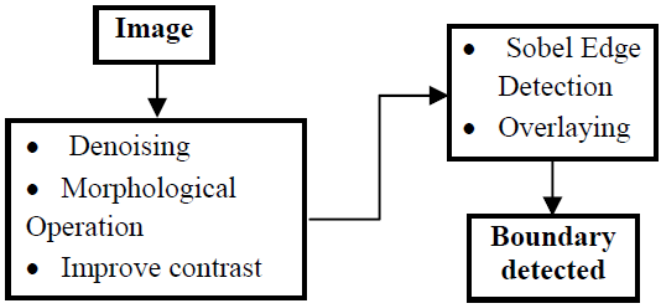


Fig. 1 Block diagram of the proposed algorithm.

3.1 Image Enhancement

The major disadvantage in echocardiograph image is the presence of noise, which perturbs features locations and creates artifacts, thus, we need a method to suppress this heavy noise without presenting additional artifacts or losing image features.

In our method, the first step is applying a median filter, where it is a non-linear technique widely used as a smoother. It is calculated simply by first sorting all the pixel values from the surrounding neighborhood in numerical order and then replacing the pixel being considered with the middle pixel value if the neighborhood under consideration contains an even number of pixels, the average of the two middle pixel values is used.

The smallest size of the neighborhood is 3 pixels, in the medical images often they use 5 pixels because they have the problem of noise and poor quality of the image and when

increasing the size of the neighborhood, they gain a better result because anything smaller than the radius of the neighborhood cannot contribute the median value will be eliminated. So, in our method, we propose a simple and effective enhancement, by increasing the size of the neighborhood, which used to define the size of details to 9 pixels. This step is required because we care about the boundaries of the heart and the movement of the valves, and we assume that all small details that are defined as noise can be ignored. After testing different sizes of the neighborhood, we conclude that the size proposed to give a better smoothing performance while sustaining the edge preserving characteristic of the conventional median filter.

After smoothing implementation, a morphological operation seems to be an effective way for more improvement in echocardiographic image. It offers a unified and powerful approach to numerous image processing problems because it could generate a certain amount of smoothing. In our method, we apply opening operation to improve filtering. The opening operation performs an erosion operation followed by dilation operation using predefined structure elements, in the method we use flat and small structure element [4]. The last step of enhancement part will be contrasted adjustment by linearly scaling pixel values between upper and lower limits, the pixels that are above or below the limits will be saturated to the upper or lower limit value.

3.2 Boundary detection

After the image enhancement from an earlier stage, the edge detection algorithm was applied. Edge detection is an essential tool, which is commonly used in many illustration techniques. In our method, we will use Sobel edge detection [1, 6] to detect the boundaries between the contours and the background in the image. The Sobel detection algorithm uses simple convolution kernel to create a series of gradient magnitudes, two convolution kernels, one to detect changes in vertical contrast GX and another to detect horizontal contrast Gy.

A (source image)
$*$ (Convolution operation)

$$Gy = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A \quad \text{..... equ. 1,}$$

and

$$Gx = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A \quad \text{..... equ. 2,}$$

The gradient magnitude at each point can be calculated by using:

$$G = \sqrt{Gx^2 + Gy^2} \quad \text{..... equ. 3,}$$

The direction of gradient calculated by:

$$\emptyset = \arctan\left(\frac{Gy}{Gx}\right) \quad \text{..... equ. 4,}$$

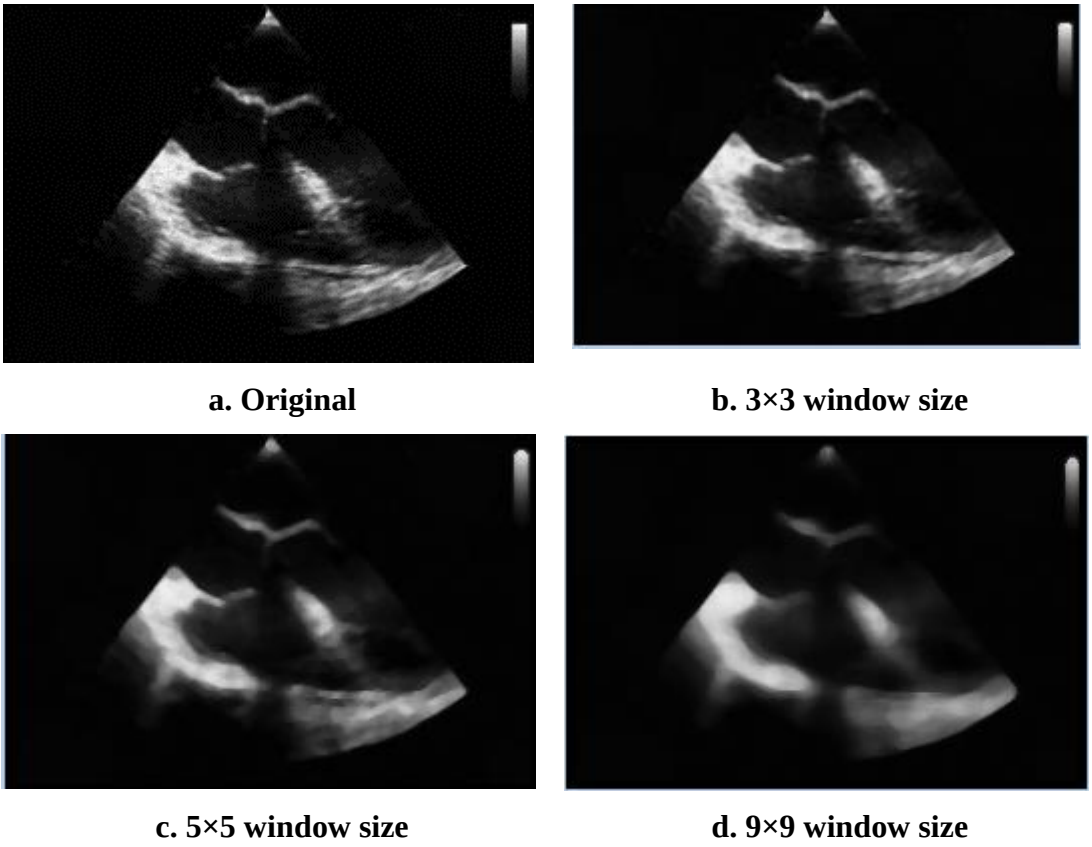
Finally, we can add one creative step to get a legible illustrative view. This step is adding operation when adding the pixel values of two images (two different images: one from opening step and the other from edge detection step) we could gain a better view for the heart contours and the movement of the valves. This combination helps in clarifying the contours of the heart and give rise to the final illustration view.

IV. EXPERIMENTAL RESULTS

Our method coded in Java. Then it applied to several echocardiograph images to demonstrate the effectiveness of the method. Echocardiograph image gathered from the Internet. The dataset

contains samples for echocardiograph image with JPEG format.

Figure 2 illustrates the effect of changing the size of the neighborhood to get a better reduction of noise. Increasing the neighborhood size gives better smoothing while the edges of the heart still distinct and clear. In figure 2 (d) we see the median filtering applied to higher window size, this neighborhood size gives rise to a more effective noise reduction and less blurring effect.

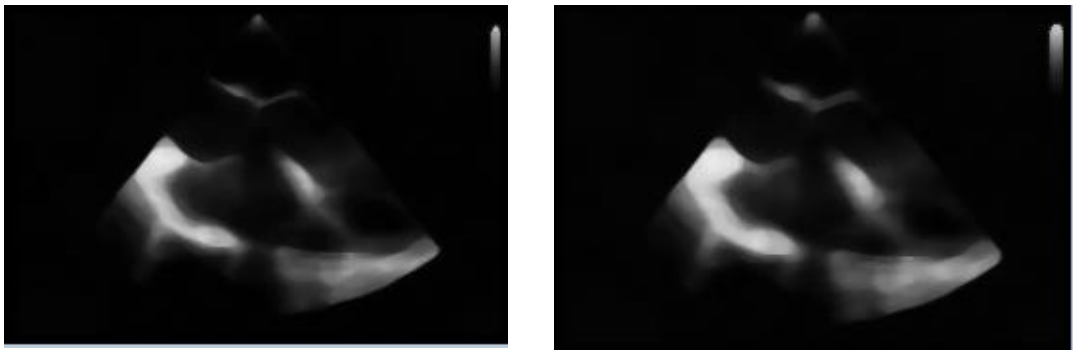


a. Original **b. 3×3 window size**

c. 5×5 window size **d. 9×9 window size**

Fig. 2 Shows Median filtering with different window sizes (window size is the size of surrounding neighborhood. (a) 2D echocardiographic image. (b) Same image after Median filtering with window size (3×3). (c) Median filtering with window size (5×5). (d)Median filtering applied with window size (9×9).

Figure 3 illustrates the effect of opening to get smooth image, whittle the narrow part and eliminate bright details.



a. Erosion **b. Dilation**

Fig. 3 Shows Morphological Opening. (a) Apply erosion operation. (b) Apply dilation operation.

Figure 4 illustrates the effect of contrast adjustment to increase the contrast by adjusting image intensity values.

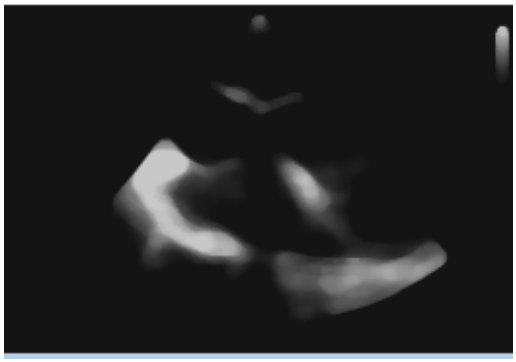


Fig. 4 Apply contrast adjustment with Lower limits = 20 & Upper Limits =200

Figure 5 presents method results after applying traditional Sobel edge detection to get contour extraction.

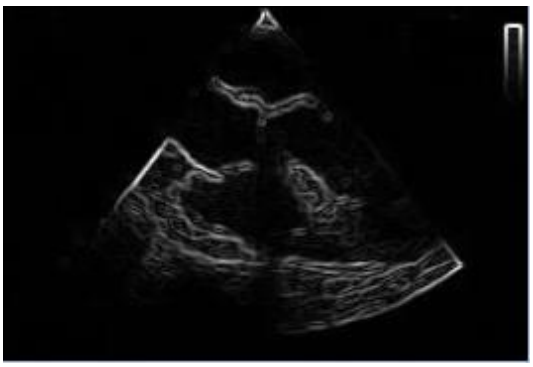


Fig. 5 Apply Sobel edge detection.

Figure 6 presents the final result after combining two images to provide a better illustrative view which could be more understanding.

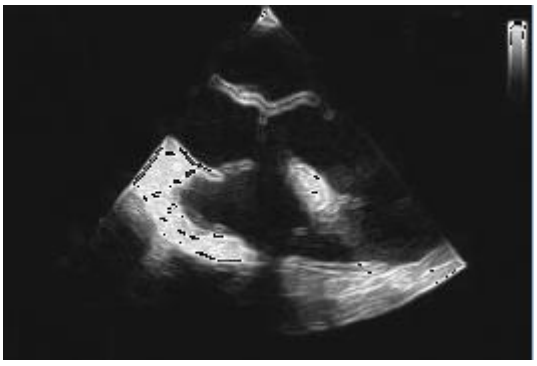


Fig. 6 Shows the final result.

The combination of several methods as a preprocessing stage in our method succeeds in enhancing the low contrast and heavy noise of echocardiographic image and present proper detection for various echocardiograph images.

V. Conclusion

In this work, we demonstrated a method to extract the heart boundaries in echocardiograph images where these images are famous with heavy noise and poor quality. Our method, based on improving the quality of these images before the detection operation, this improvement present great help to get accurate and clear contour extraction

Image smoothing using nonlinear filtering (Median filtering) and morphological operation followed by contrast enhancement results an effective enhancement for detection stage and the combination of two images from different stages produce an evident illustration that helps the specialist to diagnose cardiac disease, especially the diagnosis simplified because of the clear definition of heart structures.

The previous examples shown before have demonstrated that the illustration and motion detection for a sequence of frames is greatly influenced by the pre-processing. We can conclude that the proposed method accurately detects the heart motion in images with poor contrast and heavy noise. The application of the pre-processing techniques described here to improve the quality of the image and enable the original video to be treated in order to get a better view that is easier to diagnose.

References

[1] Boonchieng E., Boonchieng W., and Kanjanavanit R., "Edge-detection and Segmentation Methods", *IEEE, computer in cardiology 2004; 31:541-544, 2004.*

[2] Lacerda S. G., da Rocha A. F., Vasconcelos D. F., De Carvalho J. L. A., Iwens G. Sene Jr. and Camapum J. F., "LeftVentricle Segmentation in Echocardiography Using a Radial Search-Based Image Processing Algorithm". *Proc. of the30th Annual International IEEE EMBS Conference Vancouver*, British Columbia, Canada, August 20-24, 2008.

[3] Maria do Carmo dos Reis M., da Rocha AF., Vasconcelos DF., Espinoza BL., de O Nascimento FA., de Carvalho JL., Salomoni S., Camapum JF.," Semi-Automatic Detection of the Left Ventricular Border". *30th Annual International IEEE EMBS Conference*, Vancouver, British Columbia, Canada, August 20-24, 2008.

[4] Park H, Chin R T. "Decomposition of arbitrarily shaped morphological structuring elements". *IEEE Trans PAM I*, vol.17, no.1, pp.2215, 1995.

[5] Santos J.B., Celorico D., Varandas J., and Dias J., " The importance of the pre-processing on the echocardiographic images for the Left Ventricular contour extraction". *European journal of acardio-Thoracaic surgery*, Vol.18, Issue4, 1October 200, Pages 458-465.

[6] Ting L., Xiaogang L., Chenglin P., and Li W., "Improved Morphological Edge Detection Algorithm for Ultrasound Heart Ventricular Wall Image in the Motion of Its Rotation", *supported by National science Foundation of China*, IEEE, 2007.

Investigating the Role of Preprocessing and Attribute Selection Methods Towards the Performance of Classification Algorithms on News Dataset

Wateen A. Aliady

***Abstract*—** This paper uses two datasets Reuters and 20NewsGroup to analyze the impact of text preprocessing steps like tokenization, stemming and stopwords removal on classification results. In addition, it studies the effect of unigram, bigram and trigram attribute on classification results. Furthermore, it studies the impact of attribute selection methods on the generated number of attributes and classification accuracy. The paper analyzes the effect of the previous based on six classification algorithms. The results show that there is a positive impact of text preprocessing techniques on the used datasets on terms of classification performance accuracy. In addition, the unigram achieved the best results because there was an associated stop words removal list unlike the bigram and trigram. Furthermore, attribute selection methods can have positive impact on the performance of text classification algorithms but choosing the best attribute selection algorithm is dependent on the dataset used.

***Keywords*—** Correlation Based Feature Selection, Chi-squared, Naive Bayes, Support Vector Machine, Sequential minimal optimization, Decision Tree, and HyperPipes.

I. INTRODUCTION

One of the vital jobs done by data mining experts is classification. Classification is a process of grouping instances into a certain class for a general understanding [1]. Several classification algorithms have been proposed by researchers. It is a known fact that there is no classification algorithm that is suitable for all types of data [2].

Text classification is performed on the basis of several steps [3]. According to [3] the first step in text categorization is the collection of text documents in different formats. The following step is about conversion of these different text files (html, sgml, txt etc) into single acceptable format. Afterwards, these files are indexed into unified documents. After that, the selection of features is performed that has a great effect on the classification results. It should be noted that there are several feature selection algorithms or methods. After that, the classification algorithms are applied. The final stage is to evaluate the performance of algorithms using different evaluation measures.

The internet is the main source of data production and these data is needed in numerous businesses. For example, News website update their website with news on a minute basis. Some businesses need this information to be classified into different categories i.e. sports, entertainment or politics. Data produced on social networks like Facebook, Twitter, Instagram and Linkedin has a great importance. Reviews on products and comments on social media posts always carry significance for marketing and even political purpose. Text classification is

Wateen Aliady , Riyadh, SA (e-mail: wateen.aliady@gmail.com).

also used for email detection like if the email is spam or not. Human beings can understand linguistic structures and their meanings easily, but machines are not successful enough on natural language comprehension.

There are two types of text classifiers: the first type is the supervised classifier that splits the data into two set: training set and testing set. The second type is the unsupervised classifier that do not need any training data where there is no labeled data. Each of these need attributes or features on the basis of which they classify the text documents. These attributes have an impact on the classification results accuracy. Therefore, picking the best features, or attributes that provides more information is essential [4].

The motivation an aim for this research is to study:

- The impact of text preprocessing techniques on the classification algorithms performance in terms of accuracy.
- The impact of unigram, bigram and trigram attributes on the results of text classification algorithms.
- How far the attribute selection algorithms are useful in achieving high classification accuracy?

II. Dataset and Methods

A. *Dataset*

In this research two popular datasets are used: 20Newsgroups and Reuters 21578 for classification. The 20 Newsgroups was collected by Ken Lang [5]. The documents in the Reuters-21578 collection was presented by Reuters Ltd. in 1987[6]. In total more than 2000 text documents are used in these experiments, where each dataset contains more than 1000 text documents. Each dataset has been treated in 12 different ways which means 12 versions have been created for each dataset. In total 24 data versions have been used in this experiment. Table I and II presents the details of each dataset.

B. *Tool*

WEKA is a famous tool which has built in implementations of data mining and machine learning algorithms. It is one of the mostly used machine learning tool by researchers. Weka stands for Waikato Environment for knowledge analysis developed by University of Waikato, Newlands. [7]. Weka is free available tool for text classification and machine learning purpose. It can be used in two ways, command line and Graphical user interface.

C. *Text Preprocessing*

Different text preprocessing methods has been used to prepare that data sets. These preprocessing steps has great impact on the results of classifiers. Figure 1 presents the preprocessing steps applied in this work. It consists of several steps which help in purifying the data and get it ready to be used for classification.

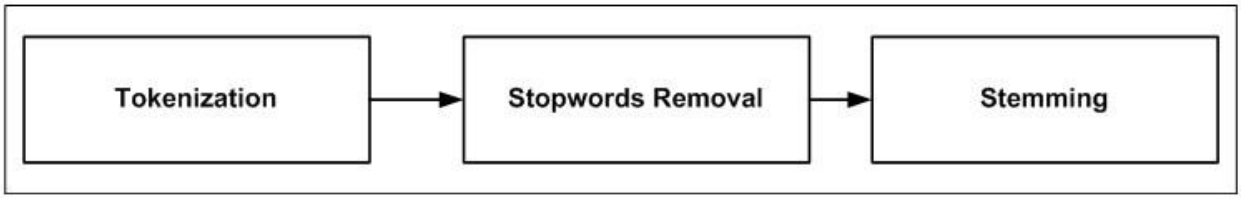


Figure 1: Text Preprocessing Steps

Tokenization is the process of splitting down the array of text in smaller chunks, i.e. words or phrases. The text in the used dataset for this work consists of news articles and they contain impurities that is the machine learning algorithms cannot understand the whole sentence like we human can understand. In order to make them ready to be used in machine learning algorithms, the sentences must be broken into single words or phrases [8].

As stated in the beginning that this paper studies the impact of different types of attributes, and therefore the data is treated within the following attribute types: the unigram type that is those attributes that consists of a single word or term. The only problem with unigram attributes is that when a document is divided in to single word it generates a huge vocabulary size. The bigram is those attributes which consists of two terms. Although it resolves the problem of the unigram attribute of huge vocabulary size and thus resolve the time and space complexity, it has its own problem. It is time consuming to develop a stop words as there is no stopwords list available for bigrams in Weka. The trigram is when the attribute contains three terms. The trigram further decreases the number of attributes but again we have no stopwords list for trigrams [9].

Stop words are those words which carry no special meaning by itself unless added to another word which provide sense to the sentence [8]. In order to understand their role, different datasets are created which is called data versions, where in some of these versions stops words were removed while in others were not. In this paper, no stopwords removal list is designed, but the build in list was used.

Stemming is another important preprocessing step in which the term is reduced to its root. This will also help in reducing the size of attributes. Different algorithms have been proposed by number of researches like Potter Stemmer and Lovin stemmer. In this study the Lovin stemmer is used [10].

Data version creation is an important step to understand the behavior of text preprocessing techniques in detail. The idea comes from the fact that it helps in observing the preprocessing impact on each data version separately. The total number of attributes are presented in the last two columns for the two-news dataset in table 1.

TABLE I
DATA VERSIONS AND NUMBER OF ATTRIBUTES

Dataset	Stemming	Stopwords removal	Tokenization	No: of attributes 20NesGroups	No: of attributes Reuters21578
D1	✓	✓	Unigram	1479	1692
D2	✓	✓	Bigram	1763	2071
D3	✓	✓	Trigram	2207	2361
D4	×	×	Unigram	1667	1854
D5	×	×	Bigram	1746	2077
D6	×	×	Trigram	2083	2184
D7	✓	×	Unigram	1565	1693
D8	✓	×	Bigram	1763	2071
D9	✓	×	Trigram	2207	2361
D10	×	✓	Unigram	1551	1689
D11	×	✓	Bigram	1746	2077
D12	×	✓	Trigram	2083	2184

D. *Attribute Selection Methods*

In this study, three attribute selection methods are used. This first attribute selection method**,** **Correlation Based Feature Selection (CFS)**, is based on feature correlation introduced by Hall [11]. This attribute selection algorithm selects those classification feature which have high correlation with the class while they have uncorrelated with each other. This algorithm is tested on different datasets and it shows that it eliminates irrelevant, redundant and noisy feature. It may degrade the performance of classifier when deleting the useful attribute. But in this study, it is explored to study its performance on text dataset.

The second attribute selection method is called the **Chi-squared**. It is a probabilistic model for selecting an appropriate set of features for classification purpose. It sees the relationship between the attribute and class. It is also called a statistical model. It was proposed by Liu and Setiono [12].

The third attribute selection method called **FilteredSubsetEval** is filter subset of attribute are evaluated. This algorithm is implemented in Weka and is used from there.

The impact of these attribute selection methods can be observed from the value presented in the Table II, presenting 20NewsGrousp and Reuters 21578 respectively. It is clear that the number of attributes are reduced dramatically when using these attribute selection methods.

TABLE II
NUMBER OF ATTRIBUTES AFTER APPLYING ATTRIBUTE SELECTION METHODS

Dataset Properties		**Total number of attributes after applying attribute selection methods (20 NewsGroups)**			**Total number of attributes after applying attribute selection methods (Reuters)**		
Dataset	**Total Attributes**	**Chisquared**	**CFS**	**FSE**	**Chisquared**	**CFS**	**FSE**
D1	1479	859	52	15	1177	45	22
D2	1763	1127	56	59	1725	54	33
D3	2207	1845	53	139	2156	58	40
D4	1667	1102	56	30	1300	46	43
D5	1746	1150	51	66	1765	50	39
D6	2083	1762	51	117	2002	56	62
D7	1565	940	53	15	1163	45	22
D8	1763	1127	56	59	1725	54	33
D9	2207	1845	53	139	2156	58	40
D10	1551	923	54	41	1249	51	10
D11	1746	1150	51	66	1765	50	39
D12	2083	1762	51	117	2002	56	62

E. *Classifiers*

The first classifier used is **Bayes net** provide a graphical structure, showing the dependencies among different variable. Each node in the graph present a variable while the connection/arcs shows the relationship among the variables. This classifier follows the probabilistic model that shows all the possible states of domain.

The second classifier used is **Naive Bayes** classifier that is one of the simple classifiers that belongs to Bayesian classifier family [13], based on Bays Theorem. This classifier represents the data as vector attribute values, whereas the labels of class are pinched from finite set of data. This method is used for labeling the dataset instances.

The third classifier used is **Support Vector Machine (SVM)** is among the most widely used algorithms for text classification and machine learning purpose. This categorization was introduced by two data scientists Vapnik [14] and Joachims [15]. For the first time it was used for text classification purpose.

The fourth classifier used is **Sequential minimal optimization (SMO)**, developed by John Platt [16] at Microsoft Research. This classifier was invented for improving the SVM classification algorithm. The implementation of this algorithm is in Libsvm and also used for svm training.

The fifth classifier used is **Decision Tree (J48)** classifier that is one of the most famous and effective decision tree classification algorithms. It was developed by Quinlan [17]. Furthermore, it works on information gain and the attribute with high information gain appears on the top of tree by recursively dividing the attributes into subsets by the normalized information gain.

The sixth classifier used is **HyperPipes** classifier is among the fastest and simplest classification algorithms [18]. It is a straightforward classifier and make it ensures consistency for each attribute. The HyperPipes also contain the bounds for the values of attribute [19].

III. Experiment And Results

There is a positive impact of text preprocessing techniques on the classification algorithms performance in terms of accuracy. The text preprocessing has positive effect in many cases while in some cases it has negative effect. The main reason for construction different versions of datasets is to know the impact of these text preprocessing techniques individually. Here in this study, three most popular techniques are used i.e. tokenization, stemming and stopwords removal

As for the tokenization process the results were that using stemming and stopwords removal on 20Newsgroups dataset has positive effect on the performance of classifiers in terms of accuracy. Most of the algorithms that is four out of six achieve high performance in terms of accuracy when treated with stemming and stopwords removal. On the other hand, it has been observed that the Reuters dataset, algorithms performs differently. The behavior shows that stopwords removal has negative impact on the results of classifiers whereas stemming proved itself positive when it comes to enhance the performance of categorization algorithms.

Figure 2 presents the difference between applying preprocessing technique and without applying preprocessing techniques on text. It shows the accuracy results of all six classification algorithms with total number of attributes, which mean that no attribute selection methods have been used. It can be seen from the below graph except J48 that all the algorithms perform well on unigram while trigram performs worst.

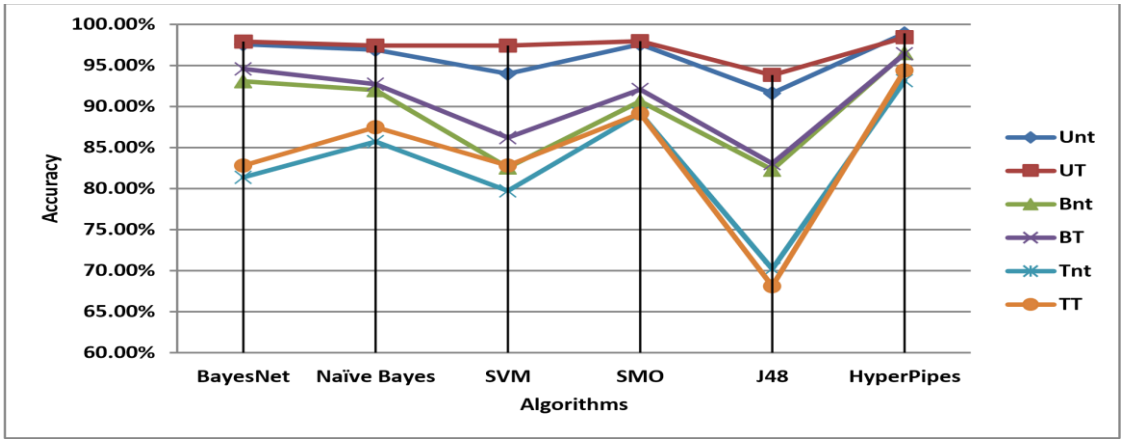


Figure 2: Difference between applying preprocessing methods and without using preprocessing methods on 20NewsGroups dataset

In the above graph on the y-axis accuracy is presented while the x-axis shows text classifiers. Other abbreviations in the graph are as in Table III.

TABLE III
SYMBOLS FOR RESULTS GRAPHS

Symbol	Stand For
Unt	Unigram attribute without text preprocessing
UT	Unigram attribute with text preprocessing
Bnt	Bigram attribute without text preprocessing
BT	Bigram attribute with text preprocessing
Tnt	Trigram attribute without text preprocessing
TT	Trigram attribute with text preprocessing

.

The impact of unigram, bigram and trigram attributes on the results of text classification algorithms can be shown in figure 2 above. Unigram attribute proved to be efficient because it increases the accuracy of classifiers. It also illustrates that unigram feature achieved high accuracy. One reason for this result is that there is no stopwords list for bigram and trigram. This can cause increase in the attribute list which according to some research degrade the performance of classifiers as shown in figure 3.

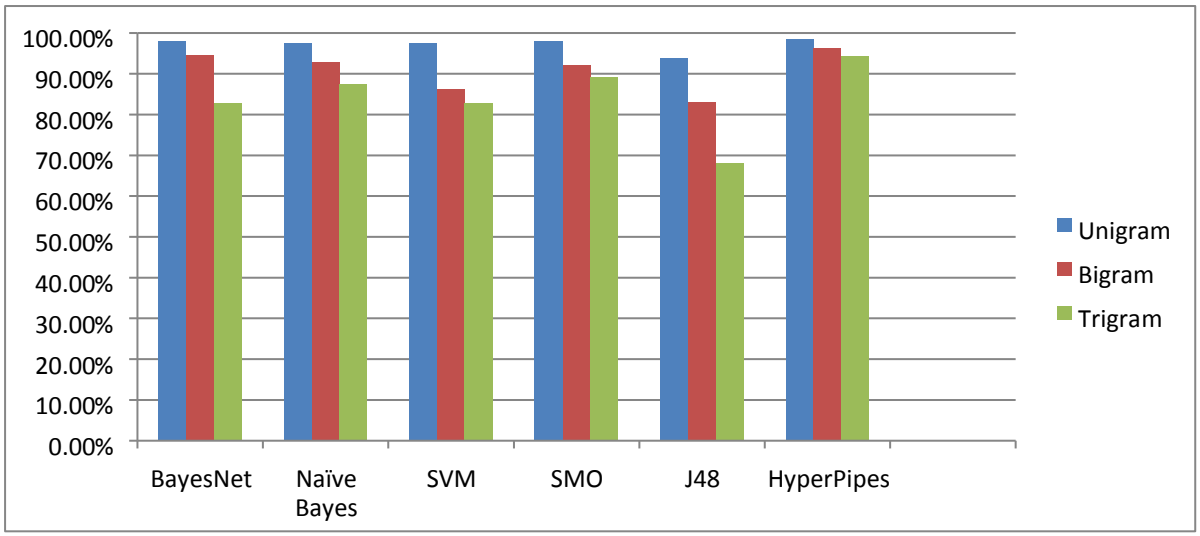


Figure 3: Classifiers accuracy level when using unigram, bigram and trigram attributes

The effect of attribute selection algorithms in achieving high classification accuracy could be splinted into two parts. The first is the impact of feature selection algorithms on the number of attributes presented in figure 4. The second part of this question is to investigate the impact on performance of classifiers, it can be shown in Table IV below.

Figure 4 below shows that attributes decrease dramatically when using CFS and FSE feature selection methods while Chi-square do not reduce too much attribute from the attribute list. The blue line presents the total number of attributes before applying the attribute selection methods.

Chi-squared enhances the performance of Baysnet, Naivebayes and HyperPipes performance for 20NewsGroups dataset and Chi-Squared also achieve high accuracy for HyperPipes using Reuters dataset. SVM and SMO perform well on total number of attribute when using 20Newsgroups dataset. CFS attribute selection has a positive impact on the performance of J48 decision tree, Baysnet, and SVM when using Reuters Dataset. FSE attribute selection method achieve high accuracy on Naive Bayes, and SMO when using Reuters dataset. Hence it proves that attribute selection methods can have positive impact on the performance of text categorization algorithms.

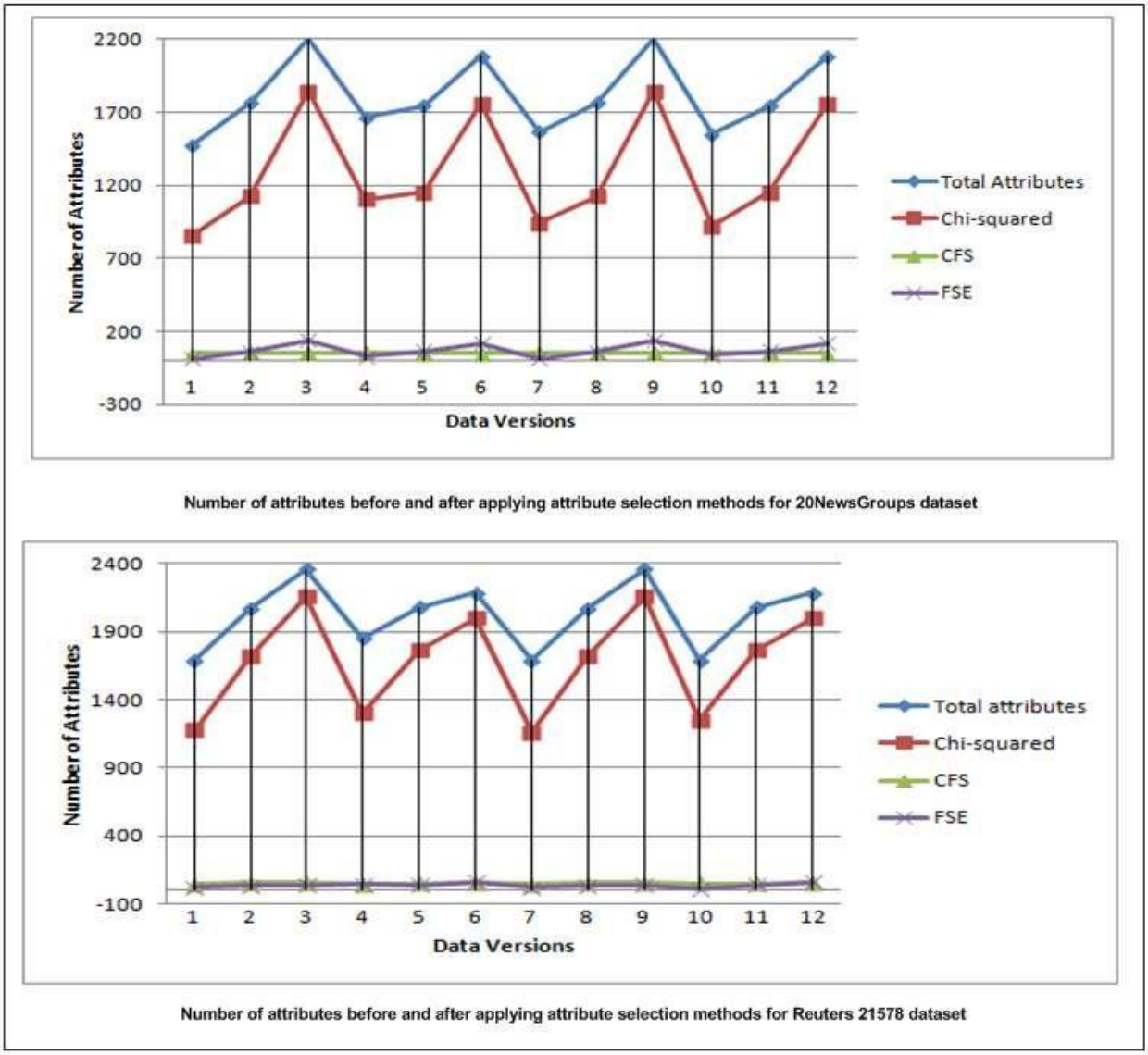


Figure 4: Different between number of attributes before and after applying attribute selection methods

TABLE IV
Data Versions and Attribute Selection Method having higher accuracy score

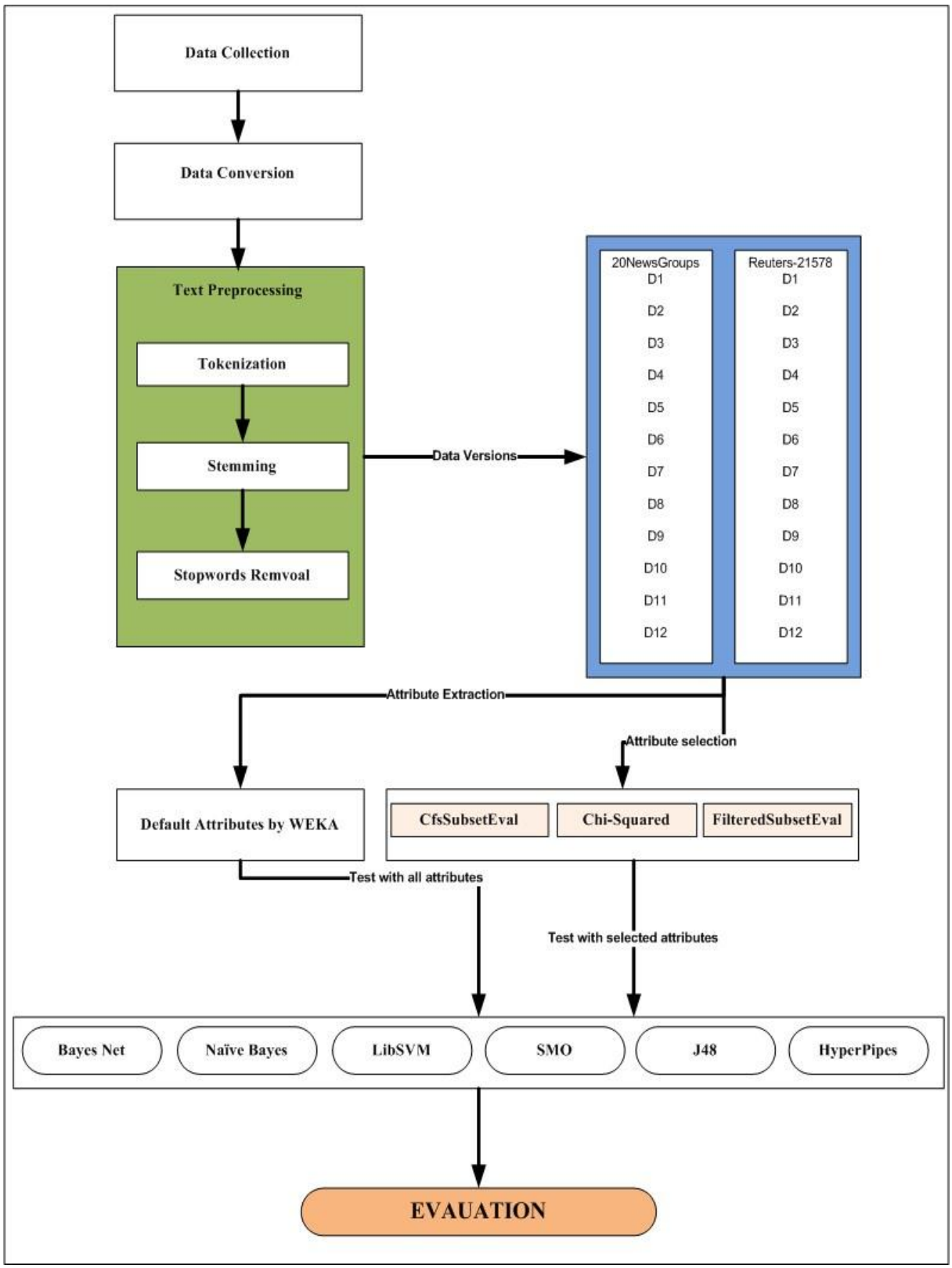
Classifier	20NewsGroups		Reuters21578	
	Attribute Selection Method	**Data Version**	**Attribute Selection Method**	**Data Version**
Baysnet	Chi-Squared	D1	CFS	D7
Naive Bayes	Chi-Squared	D1	FSE	D7
SVM	Total Attributes	D1	CFS	D7
SMO	Total Attributes	D7	FSE	D7
J48	CFS	D7	CFS	D7
HyperPipes	Chi-Squared	D4	Chi-Squared	D8

IV. CONCLUSION

This paper aims to study the impact of text preprocessing techniques on the performance of classification algorithms in terms of accuracy. Furthermore, it analyzes the impact of unigram, bigram and trigram attributes on the classification result values. Nonetheless, it studies the impact of attribute selection algorithms on classification accuracy. It uses two datasets for this purpose: 20Newsgroup and Reuters-21578 datasets. Figure 5 sums up all the work implemented work in Weka tool, where it presents the sequential order of this work as it goes through text preprocessing to generate 12 data versions for each data set. Then, attribute selection and classification are performed.

To conclude, there was a positive impact of text preprocessing techniques on the used datasets on terms of classification performance accuracy. In addition, the unigram achieved the best results because there was an associated stop words removal list unlike the bigram and trigram. Furthermore, attribute selection methods can have positive impact on the performance of text classification algorithms but choosing the best attribute selection algorithm is dependent on the dataset used.

Figure 5: Methodology



REFERENCES

[1] K. Nalini and L. Jaba Sheela, "Survey on Text Classification", International Journal of Innovative Research in Advanced Engineering, vol. 1, no. 6, 2014. [Accessed 10 March 2019].

[2] A. Gupte, S. Joshi, P. Gadgul and A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis", International Journal of Computer Science and Information Technologie, vol. 5, no. 5, 2019. [Accessed 10 March 2019].

[3] J. Mandowara, "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification," vol. 6, no. 2, pp. 126–129, 2016.

[4] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", International Journal of Advanced Research in Artificial Intelligence, vol. 2, no. 2, 2013. Available: 10.14569/ijarai.2013.020206.

[5] K. Lang, "Newsweeder: Learning to filter Netnews," in Proceedings of the 12th international conference on machine learning, 1995, pp. 331–339

[6] D. D. Lewis. Reuters-21578 text Categorization test collection. Distribution 1.0. README file (version 1.2). Manuscript, September 26, 1997

[7] WEKA, "WEKA: Data mining and Machine learning tool," 2013.

[8] V. Gurusamy, Vairaprakash & S.Kannan, Subbu. "Preprocessing Techniques for Text Mining", Proc. RTRICS,2014

[9] S. Hebbring, M. Rastegar-Mojarad, Z. Ye, J. Mayer, C. Jacobson and S. Lin, "Application of clinical text data for phenome-wide association studies (PheWASs)", Bioinformatics, vol. 31, no. 12, pp. 1981-1987, 2015. Available: 10.1093/bioinformatics/btv076.

[10] S. Vijayarani, M. Nithya and J. Ilamathi, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, vol. 5, no. 1, 2019. [Accessed 10 March 2019].

[11] M. a Hall, "Correlation-based Feature Selection for Machine Learning," Methodology, vol. 21i195-i20, no. April, pp. 1–5, 1999.

[12] R. Setiono, "Chi2: feature selection and discretization of numeric attributes," Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence. pp. 388–391, 1995.

[13] M. Friedman, Nir and Geiger, Dan and Goldszmidt, "Bayesian network classifiers," Mach. Learn., vol. 29, no. 29, pp. 131--163, 1997.

[14] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.

[15] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in Proceedings of the 10th European Conference on Machine Learning, 1998, pp. 137–142.

[16] J. C. Platt, "12 fast training of support vector machines using sequential minimal optimization," Adv. kernel methods, pp. 185--208, 1999.

[17] J. Quinlan, C4. 5: programs for machine learning, vol. 240. Elsevier, 1993.

[18] E. Witten, Ian H and Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[19] S. Alvestad, "Early warnings of critical diagnoses," Institutt for datateknikk og informasjonsvitenskap, 2009.

APPENDIX

A. *Bayes Net Classifier*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	Chi squared	CFS	FSE	Total Attributes	Chi squared	CFS	FSE
DV1	**97.91%**	**97.91%**	97.49%	95.82%	88.14%	88.41%	98.42%	97.86%
DV2	94.57%	94.57%	91.06%	92.23%	91.56%	91.56%	95.55%	94.53%
DV3	82.80%	82.80%	79.88%	83.55%	91.38%	91.38%	89.52%	88.41%
DV4	97.57%	97.57%	97.32%	95.99%	**95.18%**	**95.18%**	97.12%	**98.33%**
DV5	93.07%	93.01%	89.06%	91.90%	91.84%	91.84%	94.25%	93.32%
DV6	81.38%	81.38%	79.63%	81.96%	89.24%	89.24%	94.06%	94.71%
DV7	97.82%	97.82%	97.49%	95.82%	92.30%	92.30%	**98.70%**	98.23%
DV8	94.57%	94.57%	91.06%	92.23%	91.56%	91.56%	95.55%	94.53%
DV9	82.80%	82.80%	79.88%	83.05%	91.38%	91.38%	89.52%	88.41%
DV10	97.49%	97.74%	**97.57%**	**97.32%**	88.22%	88.22%	97.47%	94.06%
DV11	93.07%	93.07%	89.06%	91.90%	91.84%	91.84%	94.25%	93.32%
DV12	81.38%	81.38%	79.63%	81.96%	89.24%	89.24%	94.06%	94.71%

B. *Naive Bayes*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	Chi squared	CFS	FSE	Total Attributes	Chi squared	CFS	FSE
DV1	**97.41%**	**97.57%**	97.27%	95.82%	88.14%	88.78%	98.70%	98.51%
DV2	92.73%	94.57%	90.15%	91.73%	91.56%	92.21%	94.71%	94.43%
DV3	87.49%	88.48%	78.71%	82.22%	91.38%	**94.62%**	90.26%	88.04%
DV4	96.91%	97.16%	96.99%	95.90%	**95.18%**	95.73%	96.94%	97.96%
DV5	91.98%	93.82%	87.98%	90.98%	91.84%	92.02%	96.29%	93.24%
DV6	85.72%	87.06%	77.37%	79.54%	89.24%	94.06%	93.41%	93.97%
DV7	97.07%	97.41%	**97.32%**	95.82%	92.30%	93.04%	**98.88%**	**98.98%**
DV8	92.73%	94.57%	90.15%	91.73%	91.56%	92.21%	94.71%	94.43%
DV9	87.47%	88.48%	78.71%	82.22%	91.38%	**94.62%**	90.26%	88.04%
DV10	96.66%	97.07%	97.16%	**96.82%**	88.22%	89.06%	96.94%	94.62%
DV11	91.98%	93.82%	87.98%	90.98%	91.84%	92.02%	96.29%	93.23%
DV12	85.72%	87.06%	77.37%	79.54%	89.24%	94.06%	93.41%	93.97%

C. *SVM Classifier*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	Chi squared	CFS	FSE	Total Attributes	Chi squared	CFS	FSE
DV1	97.41%	97.32%	97.32%	95.74%	88.14%	98.79%	99.07%	98.51%
DV2	86.22%	88.06%	83.30%	86.14%	91.56%	91.10%	97.59%	94.43%
DV3	82.80%	83.72%	66.77%	72.53%	91.38%	70.62%	94.34%	88.04%
DV4	93.99%	94.90%	96.74%	94.99%	95.18%	96.94%	98.05%	97.96%
DV5	82.55%	83.97%	85.80%	86.97%	91.84%	91.47%	96.10%	93.24%
DV6	79.71%	77.79%	70.45%	70.78%	89.24%	96.23%	91.84%	93.97%
DV7	96.32%	96.74%	97.07%	95.74%	92.30%	98.98%	99.25%	98.88%
DV8	86.27%	88.06%	83.30%	86.14%	91.56%	91.10%	97.57%	94.43%
DV9	82.80%	83.72%	70.78%	72.53%	91.38%	70.62%	94.34%	88.04%
DV10	96.57%	96.91%	96.82%	96.07%	88.22%	98.23%	97.03%	94.62%
DV11	82.55%	83.97%	85.80%	86.97%	91.84%	91.47%	96.10%	93.23%
DV12	79.71%	77.79%	70.45%	70.78%	89.24%	69.23%	91.84%	93.97%

D. *SMO Classifier*

Dataset	**20NewsGroups Cross Validation @10 Fold Accuracy**				**Reuters-21578 Cross Validation @10 Fold Accuracy**			
	Total Attributes	**Chi squared**	**CFS**	**FSE**	**Total Attributes**	**Chi squared**	**CFS**	**FSE**
DV1	97.97%	97.41%	97.07%	95.65%	88.14%	99.35%	99.44%	99.19%
DV2	92.07%	94.49%	90.06%	90.98%	91.56%	98.79%	97.77%	91.75%
DV3	89.14%	91.37%	83.72%	90.06%	91.38%	95.05%	93.97%	91.75%
DV4	97.57%	97.41%	97.74%	95.74%	**95.18%**	99.16%	97.49%	95.60%
DV5	90.65%	92.15%	88.89%	91.40%	91.84%	98.60%	95.92%	96.20%
DV6	89.14%	90.31%	81.88%	87.72%	89.24%	97.40%	95.77%	93.60%
DV7	**97.99%**	**97.91%**	97.41%	95.65%	92.30%	**99.53%**	**99.44%**	99.25%
DV8	92.07%	94.49%	90.06%	90.98%	91.56%	98.79%	97.77%	**99.77%**
DV9	89.14%	91.73%	83.72%	90.06%	91.38%	97.33%	93.97%	91.75%
DV10	96.49%	97.24%	**97.82%**	**96.82%**	88.22%	99.07%	98.51%	94.71%
DV11	90.65%	92.15%	88.89%	91.40%	91.84%	98.60%	95.92%	96.20%
DV12	89.14%	90.31%	81.88%	87.72%	89.24%	97.40%	95.27%	93.60%

E. *J48*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	**Chi squared**	**CFS**	**FSE**	**Total Attributes**	**Chi squared**	**CFS**	**FSE**
DV1	**93.82%**	**95.15%**	94.74%	**94.65%**	88.14%	96.01%	96.84%	96.38%
DV2	83.05%	84.64%	77.71%	77.39%	91.56%	96.75%	95.55%	95.92%
DV3	68.11%	68.61%	78.71%	66.77%	**91.38%**	94.25%	84.61%	84.80%
DV4	91.62%	93.23%	93.07%	92.90%	95.18%	96.38%	95.27%	95.82%
DV5	82.30%	84.22%	79.38%	79.21%	91.84%	95.18%	94.06%	94.25%
DV6	70.20%	70.11%	65.77%	65.94%	89.24%	93.88%	87.95%	89.62%
DV7	**93.82%**	94.24%	**95.49%**	**94.65%**	92.30%	96.38%	**97.49%**	**97.03%**
DV8	83.05%	84.64%	77.71%	77.37%	91.56%	**96.75%**	95.55%	95.92%
DV9	68.11%	68.61%	66.77%	66.77%	91.38%	94.25%	84.61%	84.80%
DV10	92.57%	92.07%	93.65%	94.32%	88.22%	95.73%	96.20%	93.69%
DV11	82.30%	84.22%	79.38%	79.21%	91.84%	95.18%	94.06%	94.25%
DV12	70.20%	70.11%	65.77%	65.94%	89.24%	93.88%	87.95%	89.62%

F. *Hyperpipes*

Dataset	20NewsGroups Cross Validation @10 Fold Accuracy				Reuters-21578 Cross Validation @10 Fold Accuracy			
	Total Attributes	**Chi squared**	**CFS**	**FSE**	**Total Attributes**	**Chi squared**	**CFS**	**FSE**
DV1	98.41%	98.49%	94.24%	86.89%	88.14%	96.10%	76.27%	75.16%
DV2	96.41%	96.57%	86.47%	89.64%	91.56%	98.70%	90.36%	84.98%
DV3	94.40%	94.40%	81.55%	90.40%	91.38%	97.77%	89.24%	84.70%
DV4	**98.91%**	**98.99%**	**95.40%**	92.32%	**95.18%**	98.33%	87.95%	**89.06%**
DV5	96.49%	96.66%	86.64%	89.64%	91.84%	96.94%	87.58%	82.11%
DV6	93.15%	93.23%	79.46%	87.81%	89.24%	96.20%	86.56%	87.85%
DV7	98.33%	98.41%	94.65%	86.89%	92.30%	96.38%	78.03%	76.16%
DV8	96.41%	96.57%	86.47%	89.64%	91.56%	**98.70%**	**90.36%**	84.98%
DV9	94.40%	94.40%	81.55%	90.64%	91.38%	97.77%	89.24%	84.70%
DV10	98.66%	98.74%	95.07%	**94.49%**	88.22%	98.14%	80.74%	50.13%
DV11	96.40%	96.66%	86.64%	89.64%	91.84%	96.94%	87.58%	82.11%
DV12	93.15%	93.23%	79.46%	87.81%	89.24%	96.20%	86.56%	87.85%

.

Comparison of Query Performance Between MySQL and MongoDB Database

Roman Ceresnak, Olga Chovancova

***Abstract*—** Nowadays exists enormously variety of databases. They are essential for storing and managing information using operations in the query. In the article, MySQL and MongoDB query performance was compared. The time duration of operations for body mass index (BMI), was taken into account. Experiments were executed on generated sample data of 100, 1~000, 10~000 items size. We conducted that relational database is more suitable for a straightly defined data structure from the perspective of time. Our results showed, which type of database is faster for the selection of data for watching BMI index of individual groups.

***Keywords*—** MySQL, MongoDB, DBMS, NoSQL, BMI index.

I. Introduction

A database is a combination of information that is organized so that it can efficiently be obtained, managed, and updated. Among operations belongs these operations: create, read, update and delete (CRUD). CRUD operations are used for managing the primary function of persistent storage [1].

Database Management Systems (DMSs) are higher-level software programs that work with lower-level application programming interfaces that take care of CRUD operations.

New kinds of DBMSs, like relational and non-relational (NoSQL) databases, have been developed to help to solve different kinds of problems. Applications programs like MySQL, MongoDB, PostgreSQL were developed to implement these kinds of DBMSs. Most of them, are using for accessing databases, the most common standardized language a SQL. The SQL is a structured query language.

The most popular relational SQL database among open source community is MySQL. It is mostly used for web sites, which are running on open source systems. The most popular non-relational SQL refers to MongoDB. It allows for exploring new ways of storing information. It becomes popular mostly due to their scalability and flexibility for Cloud Computing and Big Data. This article focuses on a comparison of query performance between MySQL and MongoDB database.

Section II. contains background for this topic. Experiments and results are described in Section III. In Section IV., results are being summarized, and potential future work will be discussed. Section V. is oriented for an overall conclusion.

II. Background

There are many databases commonly, relational and non-relational (NoSQL) databases. The concept of a relational database is that, a data structure that allows linking information from different tables, or different types of data buckets. A non-relational database only stores data without explicit and structured mechanisms to link data from different buckets to one another. Relational databases usually work with structured data, and non-relational databases are work with semi-structured data.

Truica [2] examined CRUD operations for MongoDB and MySQL databases. Asynchronous replication, which is necessary for a scalable and flexible system, was examined in the paper of [3].

Roman Ceresnak, Faculty of Management Science and Informatics, University of Zilina, Slovakia, (email: roman.ceresnak@fri.uniza.sk)
Olga Chovancova, Faculty of Management Science and Informatics, University of Zilina, Slovakia, (email: olga.chovancova @fri.uniza.sk)

They used for testing, the execution time for CRUD operations for a single database instance and a distributed environment.

In the research [4], a comparative study of non-relational databases and relational databases was presented. Their primary focus was on the comparison of MongoDB to MySQL. Their results stated that MongoDB is more efficient than MySQL. They used a no-relational database for integration in a forum in the field of personal and professional development. They also presented a framework for database integration. Performance evaluation of MySQL and MongoDB was also performed in paper [5].

A comprehensive comparison of SQL and MongoDB databases for various CRUD operations and large datasets was performed in work of [6].

Continuation of Gyordodi work [7] was a comparative study between the usage abilities of MongoDB, and MySQL, as a back-end for an online platform. They presented advantages for using MongoDB compared to a MySQL. They integrated their results on an online platform for publishing articles, books, and so on, with the possibility of sharing them with other users. The primary outcome of their work is highlighting differences between MySQL and Mongo for executed operations in a parallel system.

In research [8], the main concepts of NoSQL databases were compared to four selected products databases (Riak, MongoDB, Cassandra, Neo4J) according to their capabilities concerning consistency, availability, and partition tolerance, as well as performance, were presented.

III. Databases

In following section, relational database MySQL and non-relational MongoDB will be shortly described.

A. MySQL Database

An ACID is an abbreviation regarding atomicity, consistency, isolation, and durability. Those features are all useful in a database system and connected to the understanding of a transaction. Atomic units of operation are called transactions, and they can be rolled back or committed. When a transaction saves changes to the database, either all the changes are, or they are rolled back. A consistent state is persisted in a database for all times. It is also after each commit or rollback, and during progressing of transactions. If data are updated across various tables, then queries returns old values or new values. Old and new values are not combined.

Indexes are significant features of query performance. Many applications require fast lookups in a query. It is necessary to design better tables, queries, and indexes for better performance. The typical database design applies a covering index wherever possible. The query results are determined totally of the index, without viewing the original table data. Respectively foreign key constraint additionally demands an index, to efficiently check if values exist either in parent and child tables.

A query is an operation, which reads information from a table. It could be one or more tables. Optimization by index depends on the parameters and structure of data. Join is a query if multiple tables are included. Example of MySQL query is shown in Fig. 1.

```
select name, surname
   from user
       join height using (op)
       join weight using (op)
   where current_year = 2019
       and (weight.current_weight
           /(height.current_height
           * height.current_height*)
        ) < 19,9;
```

Fig. 1. Example of query code for MySQL database.

B. MongoDB Database

MongoDB is a non-relational database. It is also open-source and document-based. High performance, automatic, and high availability provides [9].

The term "MongoDB" originates from the word "humongous." That is mainly because of databases ability to scale up with ease, and it allows containing enormous amounts of data. Documents are stored in collections within databases [10].

MongoDB performs requests to read data from the database. MongoDB uses a JSON-like query language. Its language includes a variety of query operators which begins with character *$*. In the mongo shell, can be called query using the commands for methods like *db.collection.find(), db.collection.findOne()*.

In Fig. 2., example of MongoDB query is shown. The query creates a collection of users by aggregating BMI index which is computed by dividing the current weight of user against the power of two of current height. If values are less than 19, then they are selected, which results in selecting all underweight users from 10 000 data sample.

```
db.getCollection('User3')
   .aggregate([
    { "$project": {
      "total": {
        "$divide": [
          "user.current_weight"
          { "$multiply": [
            { "$multiply":
               [
               "user.current_height",
                user.current_height"
               ]
            },
            10000
          ]},
          :{ $lte: 19.9}
      ]}}}
])
```

Fig. 2. Example of query code for MongoDB database.

IV. EXPERIMENTS

Experiments were performed on the machine with the following configuration:

- **Computer**: MacBook Pro (Retina, Early 2015),
- **CPU**: 2,9 GHz Intel Core i5,
- **RAM**: 8 GB 1867 MHz DDR3,
- **Hard Disk**: 500 GB SSD,
- **Operating system**: macOS Mojave.

Sample data were generated for conducting experiments. These data will be used for calculation of body mass index, shortly BMI. Users were divided, according to the results of BMI, into five groups, which are shown in Table 1.

Users were created based on the mentioned groups. The main idea of experiments is to compare the speed of gaining individual data about BMI index from users' weight and height. Records in the constructed table for experiments contained 100, 1 000, and 1~000 items, which represented users.

The objective of the experiments is to make a query, which shows the count of users with the same BMI index. BMI is measured by is Eq. 1:

$$BMI = {}^{m}/_{h^2} \tag{1}$$

where *m* is body mass in kilograms, and *h* is body height in meters.

In Table 1, the groups of BMI based on health risk are shown. This table is referential for

evaluating category in our experiments.

TABLE I
BMI INDEXES ACCORDING TO THEIR HEALTH AND WEIGHT CATEGORY

BMI	Health risk	Weight
0 - 19.9	middle	underweight
20 - 24.9	low	normal weight
25 - 29.9	middle	overweight
30 - 39.9	high	obesity
> 40	very high	extreme obesity

Attributes such as name, surname, height, and weight are needed for selecting the BMI group. The first step is creating the data, and the next step is indexing up the data, so it could be used for observing the changes in the gaining of values for each BMI group.

V. RESULT

In this section are described results from experiments of query performance between MySQL relational database a MongoDB non-relational database.

A. MySQL results

Results of first experiment are shown in Fig. 3. Results of the experiment are measured in milliseconds by retrieving each user (records) with related BMI index.

For finding out the value of people we will need attributes such as name, surname, height, and weight and so, in the next step we will create the index up to the data and we will also watch the changes of the times needed for the gaining of values for each BMI group.

Fig. 4 shows times measured for the gaining of all people with using the index up to the chosen data with the same BMI for the chosen number of lines in the MySQL database.

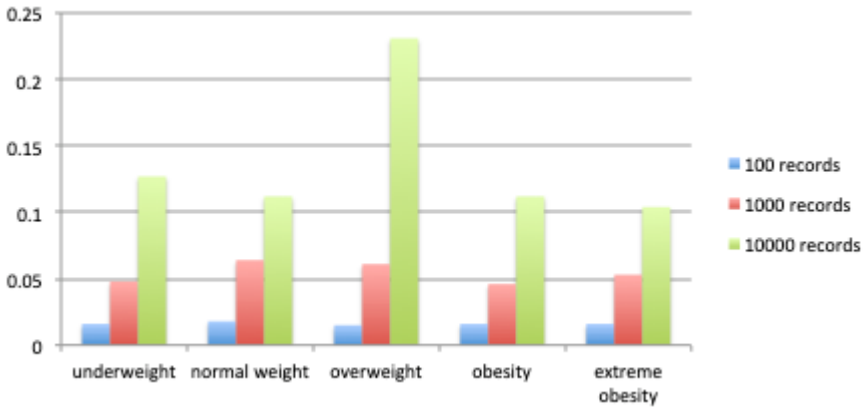


Fig. 3. Results of Experiment 1 on MySQL database

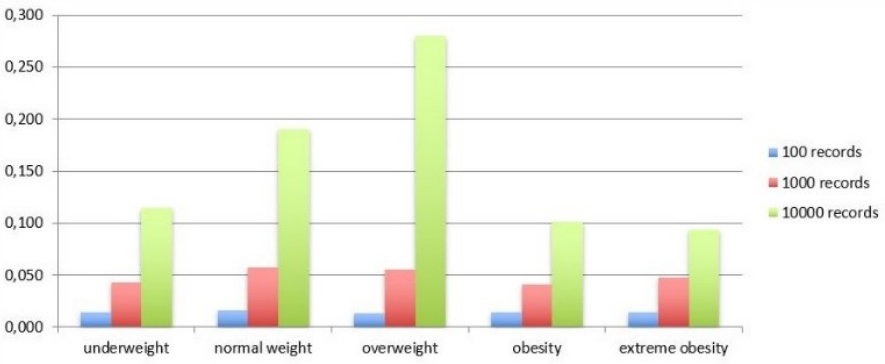


Fig. 4. Results of indexing up on MySQL database

B. MongoDB results

The experiment with the speed of gaining the data was also implied in non-relational database MongoDB, and up to the created attributes we created the index too. Differences between the speed of data are possible to compare in Fig. 5 and. Fig. 6. presents times measured for the gaining of every person with the same BMI for the chosen number of lines in the MongoDB database. Fig. 6. shows times measured for the gaining of all people with using the index up to the chosen data with the same BMI for the chosen number of lines in the MongoDB database.

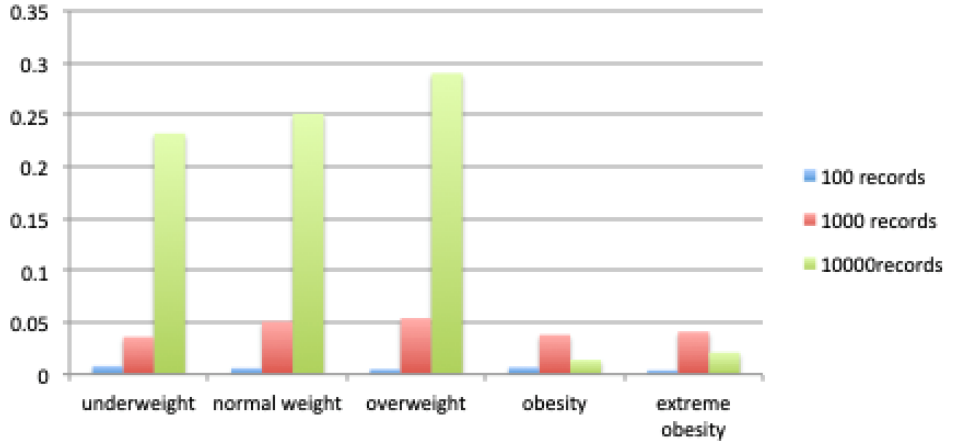


Fig. 5. Results of Experiment 1 on MongoDB database

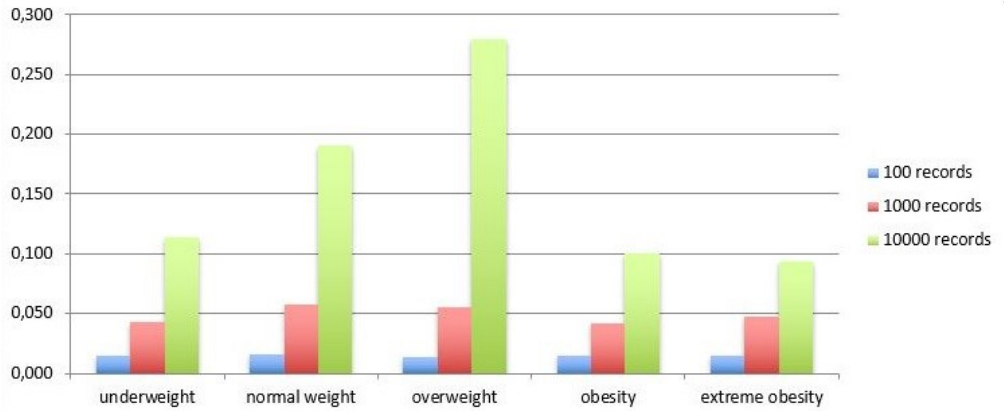


Fig. 6. Results of indexing up on MongoDB database

C. Experiment with indexes

For the comparing reasons we chose two types of data, there are visible in Fig. 7 and 8. For the first reason, we chose the table with line value 10 000 and watched the difference of times with and without using. Subsequently, we did this experiment for the non-relational database too.

Fig. 7. presents comparation of difference in using respectively no-using of the index in relational database MySQL on the sample of 10 000 data. Fig. 8. displays comparation of difference in using respectively no-using of the index in non-relational database MongoDB on the sample of 10 000 data.

As we can see in Fig. 7. and 8., in the case of optimization the speed for the gaining of data amount, it is suitable to use the index, which will make the demand up to the individual data faster. In increasing the data amount, the speed of choosing wanted data in non-relational database MongoDB is decreasing, and the effectiveness of the relational database is increasing.

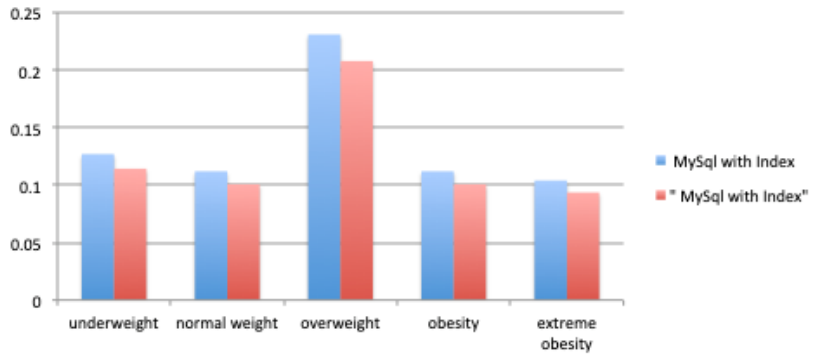


Fig. 7. Comparation of MySQL on the sample of 10 000 data

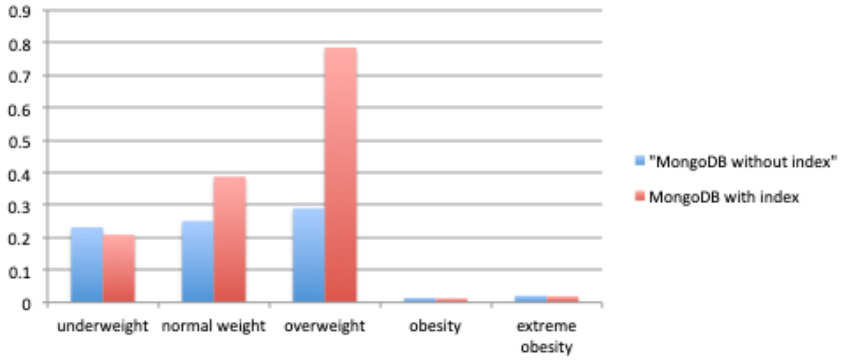


Fig. 7. Comparation of Mongo on the sample of 10 000 data

VI. Discussion

We did the experiments, which shows us that with the straightly defined data structure in the number of data 10 000 and more is from the time perspective more suitable to use a relational database. Then we examined how the choice of data is affected by the creation of the current index up to the given data. This also made the selecting of data in watching of body mass index of individual groups faster.

In the case of optimization, the speed for the gaining of data amount, it is suitable to use the index, which will make the demand up to the individual data faster. In increasing the data amount, the speed of choosing wanted data in non-relational database MongoDB is decreasing, and the effectiveness of the relational database is increasing.

VII. Conclusion

In this article, we focused on query performance between most used relational MySQL database and most used a non-relational MongoDB database.

For a demonstration of experiments, we created a simple table of users and filled up with random data. Later, was created a query which selected BMI category based on weight and height. We used a small data size because we were observing the query performance on small data.

Our finding was conducted for a most used type of query, which is select, and better suited is relational database MySQL instead of MongoDB database.

It gives a better and faster result for 10 000 data sample size from the viewpoint of a time scale. For 10,000 records, we found out that in the most common select operation, the relational database will be suited for faster results.

For future work, we plan to conduct more experiments, and with more significant and more relationships between tables by comparing more types of databases.

References

[1] M. G. Lindquist, “Managing the data-base environment,” Inf. Process. Manag., 2002.
[2] C.-O. Truica, A. Boicea, and I. Trifan, “CRUD Operations in MongoDB,” 2013.
[3] C. O. Truica, F. Radulescu, A. Boicea, and I. Bucur, “Performance evaluation for CRUD operations in asynchronously replicated document oriented database,” in Proceedings - 2015 20th International Conference on Control Systems and Computer Science, CSCS 2015, 2015.
[4] C. Gyorodi, R. Gyorodi, G. Pecherle, and A. Olah, “A comparative study: MongoDB vs. MySQL,” in 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), 2015, pp. 1–6.
[5] D. Damodaran B, S. Salim, and S. M. Vargese, “Performance Evaluation of MySQL and MongoDB Databases,” Int. J. Cybern. Informatics, 2016.
[6] R. Aghi, S. Mehta, R. Chauhan, S. Chudhary, and N. Bohra, “A comprehensive comparison of SQL and MongoDB databases,” Int. J. Sci. Res. Publ., 2015.
[7] C. Gyorödi, R. Gyorödi, I. Andrada, and L. Bandici, “A Comparative Study Between the Capabilities of MySQl Vs. MongoDB as a Back-End for an Online Platform,” Int. J. Adv. Comput. Sci. Appl., 2016.
[8] T. Wiktorski, “NOSQL Databases,” in Advanced Information and Knowledge Processing, 2019, pp. 77–84.
[9] D. Hows, P. Membrey, E. Plugge, and T. Hawkins, “Introduction to MongoDB,” in The Definitive Guide to MongoDB, Berkeley, CA: Apress, 2015, pp. 1–16.
[10] MongoDB Inc., “The MongoDB 3.2 Manual,” MongoDB Manual 3.2. 2016.

Development of foreign economic activity at the regional level: impact factors modeling

Roman V. Fedorenko, Tamas Czegledy, Nadezda A. Zaichikova

***Abstract*—** The article is devoted to the study of factors influencing the process of integration of chosen regions of the Russian Federation into the system of world economic relations. The relevance of the research is due to the increasing importance of foreign trade activities for chosen regions of the country. The dynamic development of international economic integration with a combination of globalization and localization processes characterize the current level of economic growth. These trends contribute to a significant increase in the level of foreign trade activity of particular regions, allowing increasing the competitiveness of the state in the international arena. The purpose of the article is to create an indicative model for assessing the factors that have the most substantial impact on the development of export-import operations at the regional level.

***Keywords*—** region, modeling, world economic relations, foreign economic activity, export, import.

I. Introduction

In the modern world, the success of a state depends on its ability to integrate into the international economic system. Effective integration gives opportunities to accelerate national economy. The urgency of researching the problem of integrating particular Russian regions into the system of world economic relations is due to the need to find new growth points to ensure successful economic development both at the regional and national levels.

Characteristics of the current Russian national economy condition leads to the conclusion that it is integrated into the world and is part of it, but the role of Russia is reduced mainly to the trade in resources [12]. Most of the particular regions of the country are successfully engaged in foreign economic activity. At the same time, it is necessary to identify the problem of the extremely uneven contribution of different regions of the country to the development of international trade operations. The need to focus on foreign markets is due to the continuous stagnation of the domestic market and the lack of positive expectations regarding the increase in the purchasing power of the local population.

Currently, the Federal center in Russia sets social and economic targets for each region. Responsibility for their achievement rests with the Governor and local authorities. The ability to successfully attract external financial flows and fulfill the requirements of the federal center to ensure the achievement of target indicators of economic development becomes extremely important for the Russian regions. Successful entry into the international market of regional enterprises allows to solve this problem and creates conditions for ensuring financial stability and building the basis for further economic growth.

A large number of researchers of foreign economic activity (FEA), proceed from the statement that it is an effective tool for economic development and a key factor in the formation of territory competitiveness.

A. Molchan et al. claim that "participation in foreign economic relations traditionally provides for the simultaneous implementation and increase of the resource potential of the territory, in some cases becoming the dominant factor of socio-economic development" [11].

M. Partridge et al. notes the importance of the impact of foreign economic activity on the development of the regional labor market [10].

R. V. Fedorenko, Samara State University of Economics, Samara, Russia (e-mail: fedorenko083@yandex.ru).
T. Czegledy, University of Sopron, Sopron, Hungary (e-mail: czegledy.tamas@uni-sopron.hu).
N. A. Zaichikova, Samara State University of Economics, Samara, Russia (e-mail: zajna@yandex.ru).

A significantly smaller number of researchers studied the factors influencing the development of the region's foreign economic activity. For example, Nazarczuk et al. considered the impact of specialization on the activity of foreign economic activity of the region [9]. N. Dritsakis analyzed the relationship between exports, investments and economic development in Bulgaria and Romania using a multivariate autoregressive VAR mode [14].

In this article, the authors consider the problem of enhancing the foreign economic activity of country's particular regions. The purpose of the article is to create an indicative model for assessing the factors that have the most substantial impact on the development of export-import operations at the regional level.

II. METHODS

The information and empirical base of the research was compiled by statistical data of the Federal State Statistics Service (http://www.gks.ru/) and the Federal Customs Service of the Russian Federation (http://www.customs.ru). Models of factors interdependence were built using the methods of mathematical statistics and econometrics. Analysis and processing of statistical information were carried out using the software packages Statistica and Microsoft Excel.

III. HISTORY OF MODELING FEA DEVELOPMENT

The history of individual elements of international trade, and thus essentially the foreign economic activity of independent state entities, has a number of millennia. In the course of economic thought development appeared a huge number of theoretical and applied models of foreign economic activity.

Mathematical models are a powerful device for research and prediction of various phenomena. The use of mathematical models in economic research allows to:

- Highlight and formally describe the most significant, essential links of economic variables and objects;
- From well-formulated initial data and correlations by mathematical methods, obtain conclusions corresponding to the object being studied to the same extent as the assumptions;
- Mathematical methods (especially statistical) allow to obtain new knowledge about the object, to evaluate the form and parameters of the dependencies of its variables, which are most relevant to the existing observations;
- Accurately and compactly set out the provisions of economic theory; formulate its concepts and conclusions.

As applied to such an object of modeling as foreign economic activity, economic and mathematical models can be defined as mathematical images of various areas and forms of foreign economic activity that are intended to imitate them, serve as confirmation of theories, or as a tool for analysis, forecasting, management.

It should be noted that there is no category of "economic and mathematical models of foreign economic activity" in the economic literature. This concept is rather collective, while its constituent objects are models of certain areas and forms of foreign economic activity, or applied models that imitate the activities of various individualized subjects of foreign economic activity.

The term "model of foreign trade" is most often used in the scientific literature. We can divide entire set of developed models of foreign trade into 3 significant groups depending on the belonging to one or another concept of foreign trade: classical models of foreign trade, neoclassical models of foreign trade and modern models international trade.

Traditionally, the classic models of foreign trade include the model of absolute advantages of

A. Smith [1], the model of comparative advantages of D. Ricardo [2], the model of factor proportions of Heckscher-Ohlin [3]. Many researchers have used the idea of comparative advantage in the study of problems of economic development of regions [6, 7, 8].

Taken together, the classical models of international trade through graphic and arithmetic tools demonstrate theoretical ideas about the causality of the current structure and directions of international trade.

These models are basic in the assessment of factors of foreign economic activity development. They create prerequisites for the disclosure of the mutual influence mechanism of foreign economic activity and the socio-economic state of the macroeconomic objects.

Neoclassical models of international trade or the "standard model of international trade" associated with the theory of general equilibrium in international trade. The general equilibrium model and the partial equilibrium model are referred to the neoclassical models of foreign trade [4].

Taken together, the neoclassical models of international trade have a weak mathematical apparatus used in the analysis. They mainly use geometric and arithmetic tools. Moreover, the models have numerous assumptions (for example, absolute freedom of competition), which reduces their practical applicability in real conditions. Nevertheless, the theoretical conclusions presented in neoclassical models reveal the interrelation of foreign trade and economy; illustrate the mutual impact of changes in the conditions of foreign economic activity and the level of the macroeconomic system well-being. Theoretical postulates of neoclassical models create a significant prerequisite for building multi-factor models of the mutual influence of socio-economic development and foreign economic activity factors.

Current models of international trade relate to the so-called “new trade theory”. The main foreign trade modern models feature is the reverse paradigm of international economic activity study. Classical and neoclassical models proceed from the existence of a homogeneous firm, which represents the entire country or industry. They describe the international “top-down” exchange of goods, i.e. focused on macroeconomic systems. With the emergence of the “heterogeneity of firms” theory and due to the development of statistics, the mathematical and econometric apparatus, modern models of foreign trade are associated with the use of a fundamentally inverse mechanism of bottom-up research. While traditional trade theory focused on the country, the newest theory emphasizes the role of firms and firm heterogeneity in international trade [5].

Thus, by economic and mathematical models of foreign economic activity, we understand mathematical models that imitate various manifestations of foreign economic activity at macroeconomic and microeconomic levels of economic activity. The scope of these models is to identify and study the most significant characteristics and trends, as well as the assessment of the foreign economic activity development prospects. The application of economic and mathematical modeling to foreign economic activity is aimed at improving the efficiency of this activity and identifying the key factors of its development.

In the present article, the main object of study is the foreign economic activity of the regions. The authors studied the main factors that can influence the increase of the export potential of the regions and the activation of foreign trade.

Foreign economic activity of the regions is an integral part of the production process. It has a significant impact on the state of the whole country economic system . The development of foreign economic activity at the regional level requires continuous government management, including planning and monitoring through an indicative assessment of key impact factors.

IV. IDENTIFICATION OF FACTORS AFFECTING FOREIGN ECONOMIC ACTIVITY

Foreign economic activity for Russia plays a priority role, being an important factor of financial stability. Revenues from foreign economic activity (customs duties, excise taxes, non-tax revenues) form a significant share of the country's budget.

Traditionally, foreign economic activity is understood as various types of business activities related to interaction with partners from abroad. The main subject of foreign economic activity is a company. Based on this position, "foreign economic activity" is an activity of the organization associated with entering the foreign market.

In relation to the national and/or regional economy (macro- and meso-level), foreign economic activity should be understood as the full range of international business relations of firms (micro-level), as well as public entities with foreign partners. These relations are aimed at integrating macro/meso-subject (state, region) to the world economic space. The interrelation of influencing factors at the micro-, macro- and meso- levels is presented in fig. 1.

The basis for the development of foreign economic activity is the successful entry of enterprises to foreign markets. Introduction of modern management technologies, reduction of costs, improvement of product quality, search for unoccupied niches in foreign markets - this is an incomplete list of conditions allowing companies to increase their foreign trade activity.

In the course of development of foreign economic activity and increasing the volume of export-import operations companies become more and more dependent on the state of region economic development. The state of the local infrastructure, the current tax regimes, the availability of financing, the human resource potential of the regions, the capacity of local markets and other indicators refer to the regional level factors of foreign economic activity development.

Conditions at the regional level are directly dependent on national economic factors. The key ones are the dynamics of GDP, the interest rate of the Central Bank, the current tax and customs legislation.

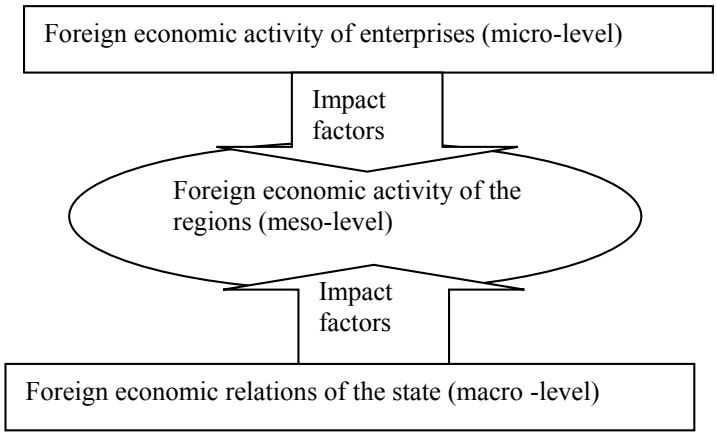


Figure 1 – Interrelation of factors influencing foreign economic activity

In addition to these factors, operating on a country-by-country basis, international factors can also be identified. The most significant factor for the model under consideration is the presence of sanctions and trade restrictions, the activities of international organizations, the dynamics of world prices for the main traded resources.

The main problem of empirical evaluation of factors influencing the development of foreign trade is the difficulty of quantifying their effect. In addition, it is mathematically difficult to quantify the favorable and adverse changes resulting from the impact of certain factors.

A. Factors Influencing Foreign Economic Activity at the National Level

The volume of foreign trade turnover largely depends on the level of economic development of the country and its regions. The state of the world economy as a whole, the level of development of relations with developed countries, the cost of oil and gas resources in the international market and the exchange rate of the national currency have a great influence. Figure 2 shows the dynamics of export-import operations of the Russian regions in comparison with the indicators of economic development.

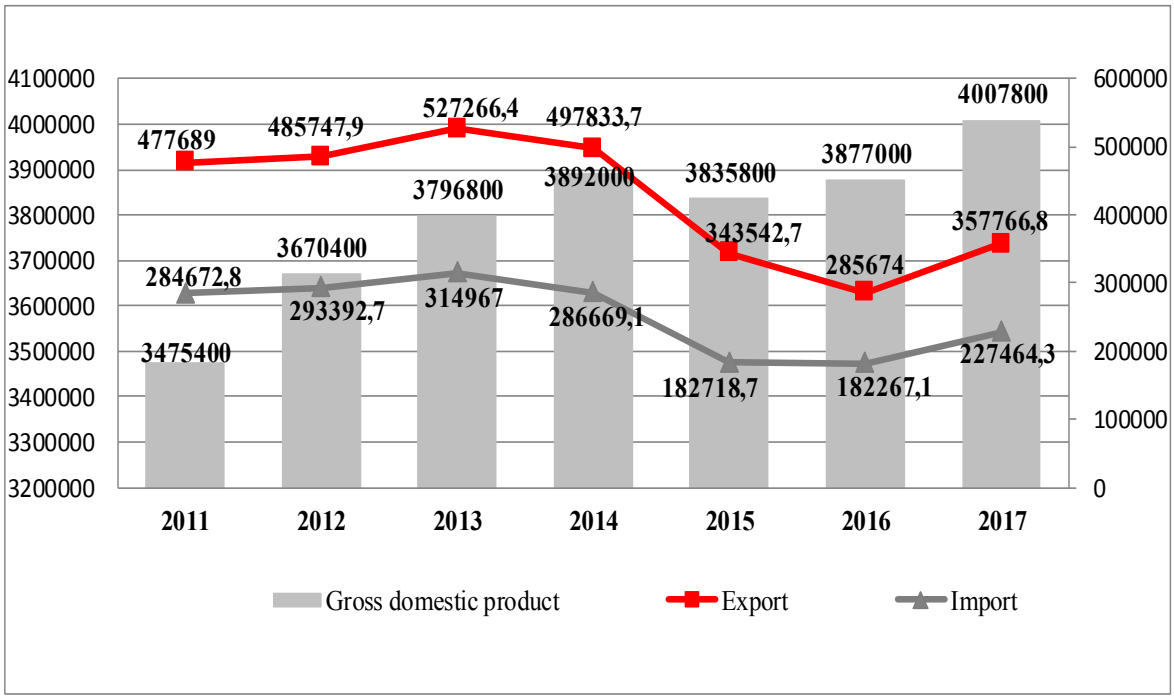


Figure 2 – Dynamics of export-import operations of Russian regions in comparison with indicators of economic development, million dollars USA

In 2011-2014, GDP of Russia grew at a rapid pace. The foreign trade turnover showed a similar dynamics in 2011-2013. The key reasons for the sharp decline in foreign trade turnover in 2014-2016 were the introduction of sanctions and anti-sanctions, which resulted in a significant reduction in export-import operations between Russia and Western countries. The sanctions had a negative impact on the level of Russia's GDP, however, the impact on foreign trade turnover was faster and stronger.

Following the results of 2017 and 2018, we can see the positive dynamics of the foreign trade turnover of the Russian Federation. However, for most of Russia's trading partners, even at the end of 2018, the pre-crisis volume of foreign trade turnover was not achieved. Thus, the foreign trade turnover between the Eurasian Economic Union and the EU in 2018 amounted to only 78% of the 2014 figure. One of the few exceptions to this trend is the turnover between the EU and China, which in 2018 amounted to 116% of the turnover in 2014. In General, we can talk about the presence of serious potential for increasing the volume of foreign trade turnover of Russia, provided that the status quo with Western countries and/or expanding cooperation with Asia-Pacific countries.

B. Factors Influencing Foreign Economic Activity at the Regional Level

In recent studies, it is noted that Russia is characterized by significant differences between individual parts of its space, both in terms of volume and structure of participation in foreign economic relations [4]. Significant disproportions are observed according to the data of 2018.

At the end of 2018, only 19 regions of the Russian Federation have a share in the total foreign trade turnover of the country more than 1%. Only seven regions have this share higher than 2% - Moscow (43.7%), St. Petersburg (7.1%), Moscow region (5.1%), Khanty-Mansi Autonomous district (3.0%), the Republic of Tatarstan (2.8%), Kemerovo region (2.5%),

Sakhalin region (2.3%). There are 85 regions in Russia. Seven listed regions provide about 67% of the total foreign trade turnover of the country. Moscow's leading role is due not only to physical volume of cargo concentration, but also to the fact of state registration of a large number of foreign trade companies in the capital of Russia. This explains the fact that the share of Moscow is about the same for both total exports and total imports.

In other leading regions, high rates of foreign trade turnover are achieved in different ways. In the regions, there is a significant discrepancy in the volumes of exports and imports. For example, in Khanty-Mansi Autonomous district foreign trade turnover is 20.6 billion dollars. And of these, 20.1 billion accounted for export operations. For the Sakhalin and Kemerovo regions, a similar situation is typical: exports account for more than 95% of the total foreign trade turnover. These regions are united by almost exclusively raw export orientation.

On average, the excess of exports over imports characterizes most regions of Russia. Of the 7 leading regions only in the Moscow region, the indicator of imports exceeds exports.

V. MUTUAL INFLUENCE OF THE FACTORS IN DEVELOPMENT OF REGIONAL FOREIGN ECONOMIC ACTIVITY: CONCEPTUAL MODEL

Factors affecting the level of foreign trade of individual regions can be divided into several levels:

- Micro-level. Factors prevailing at the individual enterprise level.
- Meso-level. Factors prevailing at the regional level.
- Macro-level. Factors prevailing at the state level
- Mega-level. Factors emerging in the international market.

Conceptually, the model of mutual influence of factors has the following form:

$$\Phi(\vec{X}) = F(\vec{X}; G(\vec{X}; H(\vec{X}; I(\vec{X})))), \quad (1)$$

where $\Phi(\vec{X})$ – the superposition of the four functions,

$I(\vec{X})$ - the functions of the micro-level,

$H(\vec{X}; I)$ – the functions of the meso-level;

$G(\vec{X}; H)$ – the functions of the macro-level,

$F(\vec{X}; G)$ – the functions of the mega-level.

For successful forecasting of foreign trade development prospects it is necessary to consider all levels of factors. The authors chose the foreign economic activity of the regions as the object of research. In this article modeling of interaction of the separate factors operating at the regional level is carried out.

A. Main Stages of Modelling

Modeling is carried out in several stages. At the first stage, an array of statistical data reflecting the state of the factors acting on the foreign trade of the region is formed. Depending on the purpose of the assessment, a set of statistical data describing foreign economic activity is formed both across the country and across the regions.

At the second stage, it is necessary to find the relationship between the variables under study. Theoretically established relationship between factors and FEA of the region can often not be traced in practice. In this regard, before proceeding to the direct determination of the most important factors, it is necessary to establish the fact of the relationship between the variables studied, and if it exists, to describe this relationship mathematically.

The third stage is scenario forecasting of the impact of certain factors on the level of foreign trade development. This stage is of the greatest practical interest for managers of local, regional

and national level, as it allows to calculate the effectiveness of investments in the development of various parameters of foreign economic activity.

B. Modeling the Factors of Foreign Economic Avtivity Development on the Example of Russian Regions

It is quite difficult to recognize and identify changes in the indicators of export-import activity caused by changes in socio-economic factors. In addition, it is mathematically difficult to quantify the favorable and adverse changes resulting from the impact of certain economic indicators.

Based on the essence of the proposed concept of the mutual influence of foreign trade indicators and the state of macro-and meso - systems, we propose to use absolute changes in the parameters of export or import, explained by the direct impact of indicators of socio-economic development.

We will use indicators of foreign trade activity as indicators of the region's foreign economic activity. Quantitative indicators of foreign trade activity are the volume of exports (EX) and the volume of imports (IM).

Let us consider the problem of constructing spatial econometric models, for which statistical data for 2017 for 85 regions of Russia were collected [13]. We present a scheme of econometric research on the example of building a model for export. For simulation the quality factor variables influencing the export, EX (US$ million) and imports, IM (US$ million), socio-economic indicators were selected (table I).

TABLE I.

SYSTEM OF INDICATORS OF SOCIO-ECONOMIC DEVELOPMENT OF REGIONS OF THE RUSSIAN FEDERATION

№	Name of indicator	Unit of measurement	Reference designation
1	Mid-annual number of people employed	thousand people	L1
2	Per capita cash income (per month)	roubles	PLL1
3	Average monthly nominal accrued wages of employees of organizations	roubles	L2
4	Fixed assets in the economy	billions of roubles	PR1
5	Output of agriculture	billions of roubles	PR2
6	Commissioning of the total area of residential premises	thousands m^2	INFR1
7	Retail trade turnover	billions of roubles	TR
8	Net financial result (profit minus loss) in the economy	billions of roubles	PR3
9	Consumer price index (December to December of the previous year)	%	PLL2
10	Investments in fixed capital	billions of roubles	I
11	Density of public roads with hard surface at the end of the year	km of tracks per 1000 km^2	INFR2

Since the methodology focuses on both national and regional levels, the table provides an indicative list of numerical indicators. Their set, mostly, is determined by the availability of statistical data, so for the Russian Federation as a whole and its regions indicators are somewhat different.

At the output, after the first stage of modeling, we obtain two numerical matrices of the same time period, the elements of which are the parameters of socio-economic development (SED) and foreign economic activity (FEA):

$$^{1}SED_{s,p} = \{sed_{ij}\}s,p \qquad ^{1}FEA_{m.p} = \{fed_{ij}\}m,p \qquad (2)$$

where p is the time period,

s, m – number of indicators of socio-economic development and foreign economic activity. For foreign economic activity, there will be no more than two of them: EX, IM; for social and economic development – at least six, namely indicators of blocks L, PLL, PR, INFR, TR, I).

The studied values have a number of features. First, the indicators of foreign economic activity are largely due to internal factors of the national economy, i.e. the level of socio-economic development of the territory. For example, the degree of development of production, of course, affects the volume of foreign economic activity and investment attractiveness for foreign investors.

Secondly, the indicators of socio-economic development are mutually influential. For example, the level of development of production, financial performance of organizations due to the degree of development of the labor market, production and finance affect the standard of living, influences budget revenues.

Third, each parameter of the socio-economic development of the macroeconomic system has a basic part, formed by root causes and subject to insignificant influence of foreign economic activity.

Thus, conditionally indicators of living standards of the population are formed under the influence of internal factors of economic development and social policy of the state, production indicators vary depending on the level of business activity, the phase of economic growth, economic policy.

The third stage of modeling is to conduct a preliminary analysis of the data in order to identify dependencies, determine their quality, and select the most relevant indicators. At this stage, the generated numerical matrices will be investigated by means of correlation analysis tools. Table II presents a matrix of paired correlation coefficients for export (EX).

TABLE II.

MMATRIX OF PAIR COEFFICIENTS OF CORRELATION FOR EXPORT

Показатели	**EX**	**L1**	**PLL1**	**L2**	**PR1**	**PR2**	**INFR1**	**TR**	**PR3**	**PLL2**	**I**	**INFR2**
EX	1	0,800	0,478	0,395	0,813	-0,051	0,305	0,860	0,948	0,152	0,598	0,640
L1	0,800	1	0,400	0,233	0,803	0,338	0,719	0,974	0,852	0,180	0,686	0,602
PLL1	0,478	0,400	1	0,909	0,579	-0,091	0,275	0,454	0,532	0,270	0,596	0,179
L2	0,395	0,233	0,909	1	0,524	-0,281	0,093	0,280	0,456	0,241	0,548	0,021
PR1	0,813	0,803	0,579	0,524	1	0,032	0,460	0,819	0,905	0,324	0,931	0,446
PR2	-0,051	0,338	-0,091	-0,281	0,032	1	0,472	0,247	0,003	-0,033	0,116	0,015
INFR1	0,305	0,719	0,275	0,093	0,460	0,472	1	0,706	0,429	0,139	0,497	0,389
TR	0,860	0,974	0,454	0,280	0,819	0,247	0,706	1	0,894	0,194	0,674	0,649
PR3	0,948	0,852	0,532	0,456	0,905	0,003	0,429	0,894	1	0,192	0,764	0,641
PLL2	0,152	0,180	0,270	0,241	0,324	-0,033	0,139	0,194	0,192	1	0,301	0,065
I	0,598	0,686	0,596	0,548	0,931	0,116	0,497	0,674	0,764	0,301	1	0,313
INFR2	0,640	0,602	0,179	0,021	0,446	0,015	0,389	0,649	0,641	0,065	0,313	1

Source: Authors

As a result of the analysis of the constructed matrix of pair correlation coefficients of the factors selected above, the following conclusions are made: the export is most influenced by such indicators as:

- net financial result (profit minus loss) in the economy;
- retail trade turnover;
- fixed assets in the economy,
- average number of employed,

- density of paved public roads.

Factors in descending order of influence are arranged as follows: PR3, TR, PR1, L1, INFR2, I, PLL1, L2, INFR1, PLL2, PR2. Socio-economic indicators in the regions are correlated and, accordingly, cannot be included in one regression model. To eliminate multicollinearity, factors that mutually affect each other are not used in models together.

After analysis various variants of multifactor models were constructed by the least squares method. Step-by-step methods of factor selection in the model were used to build the best statistical quality and explanatory capacity of the model variants. Robust estimates of standard errors (corrected for heteroscedasticity) were calculated for the models. The regression equation for the export model is:

$$EX = -5425{,}31 + 2{,}64339PR1 + 14{,}0235INFR2,$$

the regression statistics are presented in summary table IV.

The constructed model shows that with an increase in fixed assets by 1 billion rubles, exports will increase by an average of 2,643 million dollars. with a constant density of paved public roads; with an increase in the density of paved public roads by 1 m2, exports will increase by an average of $ 14.024 million at a fixed value of fixed assets.

Table III presents a similar table for import indicators.

TABLE III.

MATRIX OF PAIR COEFFICIENTS OF CORRELATION FOR IMPORT

Показатели	**IM**	**L1**	**PLL1**	**L2**	**PR1**	**PR2**	**INFR1**	**TR**	**PR3**	**PLL2**	**I**	**INFR2**
IM	1	0,844	0,457	0,356	0,907	-0,052	0,455	0,911	0,964	0,152	0,765	0,725
L1	0,844	1	0,387	0,210	0,851	0,338	0,713	0,975	0,850	0,155	0,794	0,621
PLL1	0,457	0,387	1	0,908	0,589	-0,095	0,258	0,446	0,518	0,248	0,670	0,193
L2	0,356	0,210	0,908	1	0,491	-0,292	0,063	0,267	0,429	0,205	0,551	0,038
PR1	0,907	0,851	0,589	0,491	1	0,025	0,447	0,890	0,938	0,261	0,926	0,557
PR2	-0,052	0,338	-0,095	-0,292	0,025	1	0,475	0,247	-0,002	-0,039	0,135	0,017
INFR1	0,455	0,713	0,258	0,063	0,447	0,475	1	0,702	0,410	0,111	0,531	0,407
TR	0,911	0,975	0,446	0,267	0,890	0,247	0,702	1	0,899	0,179	0,810	0,662
PR3	0,964	0,850	0,518	0,429	0,938	-0,002	0,410	0,899	1	0,153	0,839	0,677
PLL2	0,152	0,155	0,248	0,205	0,261	-0,039	0,111	0,179	0,153	1	0,224	0,083
I	0,765	0,794	0,670	0,551	0,926	0,135	0,531	0,810	0,839	0,224	1	0,480
INFR2	0,725	0,621	0,193	0,038	0,557	0,017	0,407	0,662	0,677	0,083	0,480	1

Source: Authors

After analysis of the constructed matrix of pair correlation coefficients of the factors selected above, it is concluded the indicators that affect the import the most are:

- net financial result (profit minus loss) in the economy;
- retail trade turnover;
- fixed assets in the economy,
- average number of people employed,
- investments in fixed capital.

Factors in descending order of influence are arranged as follows: PR3, TR, PR1, L1, I, INFR2, PLL1, INFR1, L2, PLL2, PR2.

If we compare tables II and III, we can conclude that the most important factors affecting exports and imports are the same. Only on the fifth largest coefficient of linear correlation of indicators, there are differences. In addition, a comparison of the two matrices suggests that

socio-economic development indicators have a greater impact on imports than on exports. Thus, the correlation coefficient for imports for only one indicator (PR 3) showed a very high correlation, exceeding the level of 0.9. For the import there are 3 such indicators: PR3, TR, PR1.

After analysis of the matrix of pair coefficients for import various variants of multifactor models were constructed by the least squares method. Step-by-step methods of factor selection in the model were used to build the best statistical quality and explanatory capacity of the model variants. Robust estimates of standard errors (corrected for heteroscedasticity) were calculated for the models. According to the first model for import, the regression equation has the following form:

IM=−3533,28+2,29619R1+8,84305INFR2−0,0116066 PR2,

It can be said that with an increase in fixed assets by 1 billion rubles, imports will increase by an average of 2,296 million dollars with the remaining factors unchanged. With an increase in the density of public paved roads by 1 m2, imports will increase by an average of $ 8.843 million with the remaining factors unchanged. With an increase in agricultural production by 1 million rubles, imports will decrease by an average of 0.012 million dollars with the remaining factors unchanged. According to the second model for import:

IM=−1595,69+0,023226PR3+3,70419INFR2,

With the increase in the net financial result by 1 million rubles, import will increase on average by 0,023 million dollars with the remaining factors unchanged. The increase in the density of public paved roads by 1 m2 will lead to an increase in imports by an average of $ 3.704 million dollars with the the remaining factors unchanged.

Below is a summary table of regression indicators for the constructed models (table IV).

TABLE IV.

REGRESSION STATISTICS OF MODEL PERFORMANCE

Models	**EX**	**IM1**	**IM2**
Variable	Coefficient (Std. Error)	Coefficient (Std. Error)	Coefficient (Std. Error)
PR1	2,643394*** (0,244551699)	2,29619*** (0,41619)	-
INFR2	14,023479*** (2,468243)	8,84305*** (2,67075)	3,70419 * (1,90225)
INFR1	-	-	-
PR2	-	−0,0116066** (0,00504801)	-
PR3	-	-	0,023226*** (0,00188712)
Constant	-5425,310*** (1117,894944)	−3533,28*** (750,192)	−1595,69*** (478,779)
R^2	0,756312	0,898289	0,939635
R^2_{adj}	0,75036817	0,894475	0,938145
F-statistic (Probability)	127,247775 (7,24796E-26)	30,67505 (2,67e-13)	166,3580 (2,08e-29)

Source: Authors

Note:

* corresponds to the significance of the coefficient estimate at the significance level of 10%;

** corresponds to the significance of the coefficient estimate at significance level 5%;

*** corresponds to the significance of the coefficient estimate at the significance level of 1%.

Table IV shows that all models meet the requirements of good statistical quality and explanatory capacity. The models are statistically significant in general at the significance level of 1% (Probability<0.01), all models have a very high explanatory ability (determination

coefficient >75%).

In the export model and the first import model, all estimates of the regression coefficients are statistically significant at a significance level of 1%. In the second model for import, the statistical significance of the regression coefficient estimate for the variable INFR2 is statistically significant at the significance level of 10%. However, the corrected coefficient of determination in this model is higher (0.938145>0.894475), which indicates a better fit of the equation for empirical data.

Thus, if the high explanatory power of the model is a priority, then the IM2 model should be preferred, if the statistical significance of regression coefficient estimates is high, and, as a consequence, the reliability of forecasts is higher, then IM1 model should be preferred. Also, the choice of a model from the two presented can be carried out according to the purpose of the study, based on what factors, from those listed in table IV, are the most suitable for its implementation

VI. CONCLUSION

Thus, according to the results of the study, it can be concluded that the following factors have the greatest impact on the development of foreign economic activity of the regions: the balanced financial result in the economy, retail trade turnover, fixed assets in the economy and the average annual number of employees.

The compiled matrices of paired correlation coefficients for export and import have minor differences, and demonstrate the stability of estimates of regression coefficients, which is a prerequisite for the study of the considered indicators in dynamics.

Conceptual model to determine the influence of individual factors in the development of foreign economic activity of the region, mathematical and methodological support of their calculation can be used in practical activities of public authorities in the formation and implementation of regional and national economic development strategies. This model allows us to conclude which indicators currently have the greatest impact on the development of export-import activities in the region.

ACKNOWLEDGMENT

The reported study was funded by RFBR and FRLC according to the research project № 19-510-23001.

REFERENCES

[1] A. Smith, “An Inquiry Into the Nature and Causes of the Wealth of Nations”. Glasgow: T. Nelson and Son, 1776.

[2] D. Ricardo, “On the Principles of Political Economy and Taxation”. London: G. Bell and sons. Retrieved on April 16, 2019 from http://www.econlib.org/library/Ricardo/ricP.html

[3] E. Trifonova and E. Bezglasnaya, “The development of International trade based on the theory of Heckscher-Ohlin. Practical use of the theory,” in *The legacy of the Nobel laureates in economics*, pp. 228-231, 2016.

[4] D. Trefler and S. C. Zhu, “Trade and Inequality in Developing Countries: A General Equilibrium Analysis,” in *Journal of International Economics*, 65(1), pp. 21–48, 2005.

[5] D. Ciuriak, B. Lapham, R. Wolfe, T. Collins-Williams and J. Curtis, “Firms in International Trade: Trade Policy Implications of the New New Trade Theory,” in *Global Policy*, 6(2), pp. 130-140, 2015.

[6] J. Nazarczuk and S. Umiński, “The geography of openness to foreign trade in Poland: the role of special economic zones and foreign-owned entities,” in *Bulletin of Geography. Socio-Economic Series*, 39, 2018. https://doi.org/10.2478/bog-2018-0007

[7] A. Cassey, “State Foreign Export Patterns,” in *Southern Economic Journal*, 78(2), pp. 308-329, 2011.

[8] K. Behrens and J.-F. Thisse, “Regional economics: A new economic geography perspective.” in *Regional Science and Urban Economics*, 37(4), pp. 457-465, 2007.

[9] J. Nazarczuk, S. Umiński and K. Gawlikowska-Hueckel, “The Role of Specialization in the Export Success of Polish Counties in 2004-2015,” in *Entrepreneurial Business and Economics Review*, 6(2), pp. 91-109, 2018. https://doi.org/10.15678/EBER.2018.060205

[10] M. Partridge, D. Rickman, M. Olfert and Y. Tan, “International trade and local labor markets:Do foreign and domestic shocks affect regionsdifferently?” in *Journal of Economic Geography*, 17, pp. 375–409, 2017. https://doi:10.1093/jeg/lbw006

[11] A. Molchan, “ Foreign economic activity as a factor of sustainable development of the regional economy ” *Polythematic network electronic scientific journal of the Kuban State Agrarian University”*, 3(97), pp. 1-11, 2014.

[12] R. Laptev, “ The national economy of Russia in the system of world economic relations "in Society: politics, economics, law”, 8, pp. 25-27, 2018.

[13] Russia in numbers: Summary of Statistics / Moscow, Rosstat, 2017.

[14] N. Dritsakis, “Exports, Investments and Economic Development of PreAccession Countries of the European Union: An Empirical Investigation of Bulgaria and Romania,” *Applied Econom. Vosow, 2017ics*, 36 (16), pp. 1831-1838, 2004.

Analysis of Data from the Social Media

Ladislav Burita, Taha Nejad Falatouri Moghaddam

***Abstract*—**Paper deals with analysis of data from the social media. It starts with explanation of used terms and specification of the social media importance in various areas of business and social life. Authors present two experiments: 1) Sentiment analysis of the Instagram data, 2) Content analysis of the Facebook data. The detail results of the both experiments with comments are included.

***Keywords*—**Sentiment and content analysis, social media, data mining; categorization; statistics

I. Introduction

Predicting of success, the future activities has been always a debating for all businesses [1]. Getting access to first hand data for these prognostications is not an easy process where most of the customers are not available after purchase.

While the social media (SM) could create a new area of investigation on this issue [2]. These days' the SM is an inseparable part of human life's this is more radical for the youngster who already lives on SM [3].

Facebook by having more than 1.79 billion users is the most popular social network, following by Instagram by having more than 500 million daily active users where more than 95 million of photos and videos share daily [4].

The amount of data daily published on social media made a potential opportunity for most business to attend. Where most of the Fortune 500 member have been established their own SM analytics system. It is estimated that the total market of SM from 1.6 billion dollars in 2015 will get 5.4 billion dollars in 2020 [5].

TABLE I
SOURCES FOR THE SOCIAL MEDIA ANALYSIS

Source for SMA	%
Microblogging	46
Review	17
Blogpost	10
Internet forum	9
Review	7
Q&A	6
Tag	5

The use of social media analysis (SMA) is versatile; it has been used in banking, education, child welfare, tourism, marketing, entertainment, government, food industry, clothing, etc. [6].

SM is any web based service with these aspects: First all the members have to sign up as a profile; Second they can make links with each other, and finally the user can share original content, or re-share second hand content there [7]. In the same paper is mentioned the percentage of using sources for the SMA, see Tab. I.

Dell Company proposes the most comprehensive definition for the SMA: "An evolving business discipline that aggregate and analyzes online conversation (industry, competitive,

L. Burita, University of Defence, Brno Czechia (e-mail: ladislav.burita@unob.cz).
T. Nejad Falatouri Moghaddam, Tomas Bata University in Zlin, Zlin Czechia (e-mail: falatouri_moghaddam@utb.cz).

prospect, and customer) and social activity generated by brands across social channels. SMA enable organizations to act on the derived intelligence for business results, improving brand and reputation marketing and sales effectiveness and customer satisfaction" [8].

Authors present two experiments in the paper:

1. SMA of Instagram with the goal to discover research results of the sentiment analysis of the supermarket chain customers.
2. SMA of the Facebook with the goal to discover research results of the content analysis with the key word "military".

II. Sentiment Analysis Experiment

A. Literature Review

By Advent of Web 2.0 technologies, the content providing in the web has been changed from publisher oriented to the user oriented, where programing abilities is not needed to broadcast a content. The personal background develops data in the SM and daily activates [8]. The user generated content is one of the important resource of SM and the valuable content on it [9].

Sentiment analyzes is a called opinion mining and it is based on the feeling of the customers. SMA is widely used in finding out the link and connections. Text mining is a method for extraction information from unstructured data. In the paper [10] is proposed clustering that combines sentiment tone, relevance, keyword analysis, intensity and alert analysis.

Personal comments show the influence of cooperating attendance in the SM. Some research results have been done on the correlation of comments and online shopping, while for the offline shops this correlation has not been investigated by the researchers [11]. Especially relating to lack of access to consumer after purchase experience directly. To overcome this shortage, the researchers tried to utilize SM comments as the useful source of consumer experience [12]. Attending in the SM where the buyers could contact the companies with no mediator, aware the companies of consumer sentiment, and improve brand reputation by gaining follower [13].

B. Research Questions

As was mentioned in the literature review, companies use the SM posts to attract customers. In this study, we are implementing data mining (DM) methods to find out how to improve influence of Instagram reputation of Ofogh-Koroush (OK). To obtain this, we analyze the last 100 post of OK Instagram page to find out a reliable rule, see Fig. 1.

Fig. 1 Ofogh-koorosh Instagram page

Investigating Instagram posts of OK Chain stores, we came across with six types of post sharing (see Fig. 2):

1) **Events:** These posts are related to some special days such as Mother's day. Sometimes the page owner shares some related content for celebrating the event or gathering attention.

These posts are mainly finished by related Hashtags.

2) **Voting:** In these posts, a question is asked about the performance and satisfaction level of the customers about a specific situation.
3) **Competition Result:** The page owner uses this media to announce the winners of any competition and campaigns in the physical stores.
4) **Self-Advertisement:** These are advertisement of OK's services and staffs' performance.
5) **Sale:** Daily sales and discounts are announced via these posts.
6) **Product Advertisement:** OK's third party cooperation's products are advertised in these posts, too.

Fig. 2 Sample of the six types of OK's Instagram posts page

To find out which type of posts are more influential for increasing OK chain stores reputation, we have to answer three main questions.

Q1: Which type of contents could bring more comments? The company has allocate some money for content providing by recognizing the most attractive type of content the money flow could invest in efficient way.

Q2: What is the best time to share a post on Instagram? It is mentioned in the literature that the number of daily posts in Instagram reach 100 million daily while most of the user follow different pages from different countries by choosing the right timing of content sharing the users could not lost your posts.

Q3: Which type of posts bring sense that is more positive? Positive comments could attract more users to follow and improve reputation of the brands. In this case, we need to understand the sentiment of the company user for each category, to emphasis on it.

C. Methodology and Analysis

Majority of comments in this page were in Persian language, therefore we faced several difficulties and challenges and had to set few changes to have a clean usable data set. The Iranian language has official speaking, which is used in the News, is a spoken language. More than 80% of the content was written in spoken language, which may be processed wrongly by the data mining software. In order to increase the accuracy, we set a dictionary to change the spoken verbs to official verbs.

The Uses of emoji in many comments make some barriers for the application to analyze the sentiment. To overcome this problem, we used some queries to change emoji to a word or

sentence based on Iranian culture; for instance, 🙌 to thank you or 😍 to like.

Persian language comment with English characters are rare and hard to be understand by programming but we used Google application (https://www.google.com/intl/fa/inputtools/try/) to overcome it. In some cases, and events such as football match prediction competition the user have to comment their guess of result, which is considered unrecognizable for most of the algorithms. We decided to change the real answer to positive answer and not related comment as a negative answer.

Near 15% of the comments were related to the admin of Instagram page and we eliminate them. In some events, the owner set competitions for mentioning other people. These posts could not be a part of our investigation relatively near to 300 comments was related to mentioning people. It was not recognizable for us to find out if it is a positive reaction or negative. By removing the duplicated comments, the final data set include 3263 comment of latest 100 post of OK Instagram. Example of the data set is at the Fig. 3.

ExportComments.com			
Source URL	https://www.instagram.com/p/BvHBptwAjos/		
Name (click to view profile)	Date	Likes	Comment
mr.ballon24	17/03/19 14:32:32	0	دوستان از پیج من هم دیدن و در صو
bijan_h.ma	17/03/19 14:44:35	1	تخفیف ها رو ول کنید به مشتریان اون
llvllr_r	17/03/19 15:05:04	0	@bijan_h.ma عید شد خبری نشد
hashem.shirazi	17/03/19 15:36:59	0	مفتشم گرونه
damnoshsaraneshaa	17/03/19 15:58:21	0	سلام ی مرغ و گوشت وبرنج وحبوبات
beti6672	17/03/19 17:04:18	0	👏👏👏
3427_maryam	17/03/19 18:38:26	0	بله خریدم از عرقیات عالی بود
hello_hazarat	17/03/19 18:52:29	0	ببخشید این امتیازات توی نرم افزار باش
nova_concept1	17/03/19 22:52:21	0	عرقیجات
rozhman_esmaili	18/03/19 16:02:31	0	یک مینو سفارش میدم
nahal.shafaatian	18/03/19 21:58:01	0	کالاهای اساسی تخفیف بزنید سرکه و
zima_gate	20/03/19 05:04:34	0	خوشم اومد😊👍

Fig. 3 Sample of downloaded comments using exportcomments.com

D. Answering Question

Answering the first question, we count the number of comments of each type of posts; the result is in the Tab. II. It is worth to mention that the company has to allocate some amount of money on content providing. According to the result, it is revealed that the Event's posts are the best for comment gathering, it could mainly be related to the trends and hashtags for that special day.

TABLE II
AVERAGE NUMBER OF COMMENTS FOR EACH POST TYPE

Post Type	Average number of Comments
Event	82.40
Voting	68.60
Competition Result	58.11
Self-Advertisement	51.57
Sale	36.29
Product Advertisement	24.50

The next we investigated how the time of posting comments to find out; what is the best time for publishing a post. The expected sense in this affair was that sharing at night while people are at home is the best choice, though as it is shown in Fig. 4, we find out that most comments have been sent between 2 pm to 7 pm while people are free of work to do shopping. It means

that best time to share a post is by the noon and two or three hours sooner than 2 pm.

The last and most important part of our research is the sentiment analysis of the comments in the six mentioned categories. For this, we used Rapid Miner studio and Rosette text Analytics extension (https://www.rosette.com/rapidminer/) that can support Persian language.

The process of sentiment analysis is shown at the Fig. 5. The results were eliminated to only Positive, Negative and Neutral comments, see the Fig. 6; the most positive comments are received on the events posts.

It shows that investment in these rather activates could worth and the problem is in sale category where most of the customers send their complaint. The least important category is the product advertisement, although the company could benefit from product advertisement by asking for the advertisement fee it is not affect its reputation online.

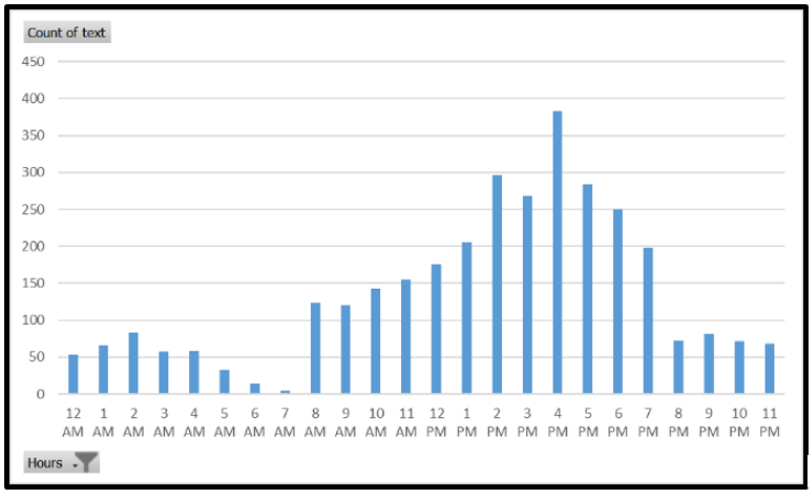


Fig. 4 Number of comments on daily time

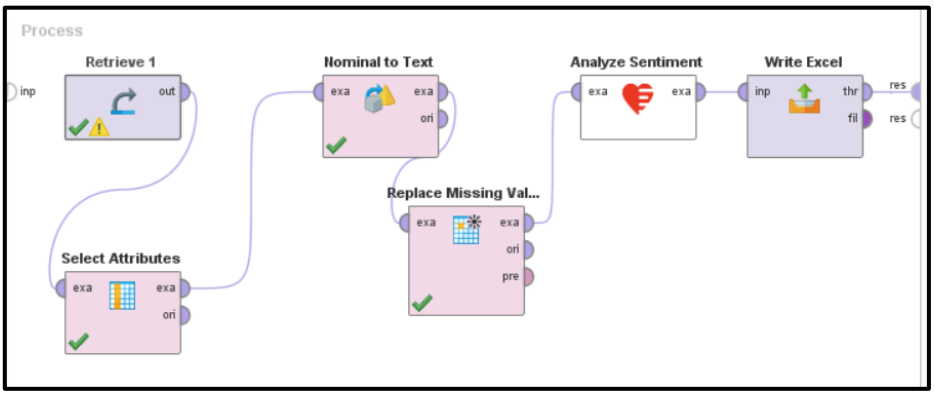


Fig. 5 Process of the sentiment analysis

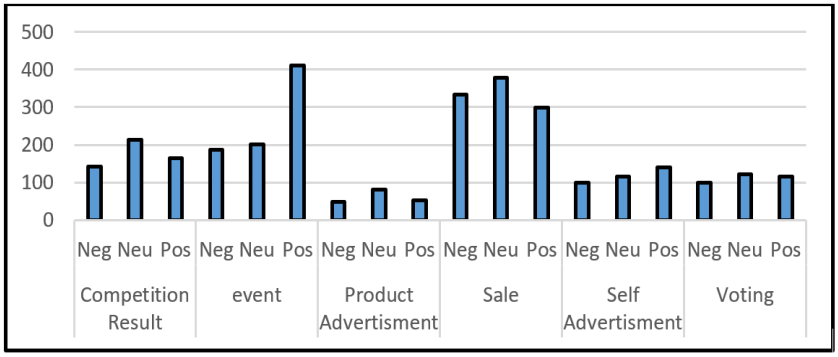


Fig. 6 Process of the sentiment analysis

III. Content Analysis Experiment

A. The Literature Review, Research Question and Hypothesis Development

The literature review was oriented to find the effective tools for data extraction from the Facebook. From the set of tools in [15] was selected only those that can be used without programming. Some tools are using conditional participation in the development team, which is also not appropriate for the authors.

The research experiment was prepared with data, obtained from Facebook, using tool Netvizz [16]. The Netvizz application provides "raw" data for both personal networks and pages, but provides data perspectives not available in other tools, e.g. comment text extraction; it also provides data for groups, a third functional space on Facebook. Running as a Web application, Netvizz does not require the use of Microsoft Excel on Windows like NodeXL and thereby further lowers the threshold to engagement with Facebook's rich data pools [17]

The aim of research [18].was to analyze the content posted by Municipal and State Tourism Organizations (DMO) of the twelve headquarters cities and States of the FIFA 2014 World Cup in their fan pages on Facebook. In the first stage, the official Facebook fan pages were identified, then posts published between June 1st and July 31st of 2013, period from pre to post-event FIFA Confederations Cup Brazil 2013 were collected. The data analysis method employed was content analysis from the perspective of Bardin (2011), which is divided into: i) pre-analysis using dedicated SW, phase ii) material exploration and iii) treatment of results, inference and interpretation. It was observed that the DMOs analyzed publish diversified information to users, including actions addressed to the abovementioned event.

An academic group and discussion forum were established on Facebook for a cohort of postgraduate students studying the concepts and principles of eLearning. The Forum had a constructivist, student-centric ethos, in which students initiated topics for discussion, while the course leader and administrator facilitated. Previous research has been conducted, involving content analysis of the topics and academic discourse, but the present study focuses on social aspects, investigating social-and study-related pursuits and determining whether synergy can exist between them. A literature review shows how social networking by students, initially social, began to overlap with academia, leading to the use of groups for academic purposes and forums for subject-related discussions. In the present study, data was triangulated and two methods of data analysis were used [19].

The research question: Find objects and subjects, services or activities, connected with the key word "military".

The working hypothesis:

H1: Data does not obtain any specific military activity, connected with warfare. Subjects and objects are not in detail described from the military organization point of view.

H2: The most records offers any services for military support or offer sale of any products from the military environment.

B. Data Acquisition and Research Results

In the basic form, 100 records can be obtained in tabular form, the structure of which consists of fields: identifier, name, check-ins, description, cover-picture, link to Facebook, and link to website. After the initial examination were removed duplicated records (4) and records that do not match the keyword query "military" (8). The remaining 88 records were subject to farther analysis. The filled values in records were incomplete, for example description (28%), cover-picture (45%), and link to website (59%). The detailed statistical analysis includes:

- Records by country of origin (Fig. 7).

- Analysis of records by category (Tab III).
- Arrangement by contributor's area of interest (Tab IV).

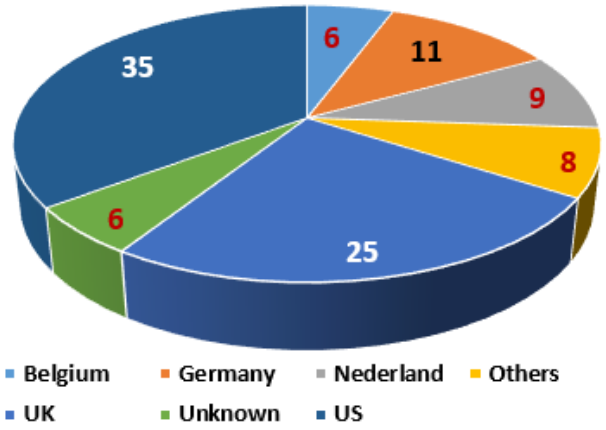


Fig. 7. Records by country of origin (%)

TABLE III
RECORDS BY CATEGORY

Category	No	%
Military shop	26	30
Military service	12	14
Military Academy or College	9	10
Military base or camp	8	9
Military Recruiting Office	6	7
Military cemetry	5	6
Military museum	5	6
Military restaurant	4	5
Military community	3	3
Military charity	2	2
Military press	2	2
Military road	2	2
Military Intelligence	1	1
Military Order	1	1
Military Police	1	1
Military racing	1	1

TABLE IV
RECORDS BY AREA OF INTEREST

Area of interest	No	%
sale	21	24
advertisement	11	13
education	10	11
history	10	11
recruitment	8	9
social	8	9
military installation	5	6
sport	5	6
housing	2	2
topography	2	2
association	1	1
honour	1	1
journal	1	1
military tatoo	1	1
psychology	1	1
publisher	1	1

Records by country of origin, see Fig. 7, is the statistics of number records that were included in any country. The most records comes from US (35), UK (25), and Germany (11); the other countries means France (2), Canada (1), Pakistan (1), Philippines (1), Russia (1), and Spain (1).

Result of records analysis by category, see Tab. IV, shows the military topic and goal of the including information.

C. Hypothesis Verification

The source for evaluation hypothesis H1 was detailed inspection of data. There is only one record about military unit ”The 40th Military Police Battalion” and NO records about military activities, connected with warfare. **The hypothesis H1 is true.**

The source for evaluation hypothesis H2 is Tab. III and for its verification is Tab. IV. Military service, support and sale in the Tab. III includes military shop, service, Recruiting Office, community, and charity, together 56%. Military service, support and sale in the Tab. IV

includes sale, advertisement, recruitment, social, housing, association, and psychology, together 59%. The result is more than 50%. **The hypothesis H2 is slightly true.**

IV. Discussion and Conclusion

The study of sentiment analysis identified the most attractive Instagram context of OK page based on text mining; the sentiment of user comments was analyzed in six different post categories (Events, Voting, Competition Result, Self-Advertisement, Sale, and Product Advertisement) to establish a reliable role for the future patch. The most attractive type of post form are Events and Sale where the event's post received more positive comment and sale post received more negative and neutral comments. According to the research on the time of leaving a comment, the best time to share a post would be between 2-4 pm.

The research results of the data content analysis from Facebook are quite different from the sentiment analysis in the first experiment. The goal of the study is searching objects, subjects, and activities, connected with selected key word. In our case, it was "military". This is useful in inspection of SM participants' interests. The surprising finding was relatively large representation of the military history (museum, cemetery, and order) at the Facebook.

The comparison both experiments summarizes following facts: The data acquisition, data content and format is nearly the same (table). The quality of data was similar, about 10% records was excluded (off topic, duplicity). In the first experiment was used for SMA a data-mining tool; in the second experiment categorization and statistics. The results of the first experiment are useful in marketing and customer satisfaction, and of the second experiment for insight into the SM data.

Acknowledgment

Visegrad Fund

This publication is a result of the project implementation: Exhibition and Special Discussion Section on Info and Digital Technologies; reg. no. 21830315. The project is co-financed by the Governments of Czech Republic, Hungary, Poland and Slovakia through Visegrad Grants from International Visegrad Fund. The mission of the fund is to advance ideas for sustainable regional cooperation in Central Europe.

References

[1] G. Chen, et al., "NPP: A neural popularity prediction model for social media content," Neurocomputing, 2019, p. 221-230.

[2] D. Baum, et al., "The impact of social media campaigns on the success of new product introductions" Journal of Retailing and Consumer Services, 2018.

[3] T.P.S. Humaniora, "83 Percent of Teenagers Inseparable from Social Media," 01/06/2016; Available at: http://news.unair.ac.id/en/ 2016/06/01/83-percent-of-teenagers-inseparable-from-social-media/.

[4] M. Ahlgren, "Top 28 Instagram Statistics & Facts For 2019," Available at: https://www.websitehostingrating.com/instagram-statistics/.

[5] I. Lee, "Social media analytics for enterprises: Typology, methods, and processes," Business Horizons, 2018. 61(2): p. 199-210.

[6] N. Misirlis, and I.M. Vlachopoulou, "Social media metrics and analytics in marketing–S3M: A mapping literature review," 38(1), 2018, p. 270-276.

[7] N.A. Ghani, et al., "Social media big data analytics: A survey," 2018.

[8] Association, I.R.M., "Social media and networking: Concepts, methodologies, tools, and applications," 2015: IGI Global.

[9] X. Xu, et al., "Business intelligence in online customer textual reviews: Understanding consumer perceptions and influential factors," 2017. 37(6): p. 673-683.

[10] I.J.B.H Lee, "Social media analytics for enterprises: Typology, methods, and processes," 2018. 61(2): p. 199-210.

[11] X. Li, C. Wu, and F. Mai, "The effect of online reviews on product sales: A joint sentiment-topic analysis," Information & Management, 2019. 56(2): p. 172-184.

[12] A. Lawani, et al., "Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston," Regional Science and Urban Economics, 2018.

[13] M. Schaarschmidt, and G. Walsh, "Social media-driven antecedents and consequences of employees' awareness of their impact on corporate reputation," Journal of Business Research, 2018.

[14] A. Erz, B. Marder, and E. Osadchaya, "Hashtags: Motivational drivers, their use, and differences between influencers and followers," Computers in Human Behavior, 2018. 89: p. 48-60.

[15] D. Freelon, "Social media data collection tools," Available at http://dfreelon.org | @dfreelon

[16] Netvizz v1.6 – tools to analyze the Facebook platform. Available at https://apps.facebook.com/107036545989762/

[17] B. Rieder, "Studying Facebook via Data Extraction: The Netvizz Application". Available at http://thepoliticsofsystems.net/permafiles/ rieder_websci.pdf

[18] A.A. Biz, C.K. Santos, E.M. Bettoni, et al." Analysis of content conveyed by the tourism departments of cities and states the headquarters of the World Cup 2014 on your Facebook pages," PASOS-REVISTA DE TURISMO Y PATRIMONIO CULTURAL, 14, 2, 2016, pp. 543-559.

[19] R. de Villiers, M.C. Pretorius, "Academic Group and Forum on Facebook: Social, Serious Studies or Synergy?" PROCEEDINGS OF THE 6TH EUROPEAN CONFERENCE ON INFORMATION MANAGEMENT AND EVALUATION, 2012, p. 63-73.

Application of fuzzy filtering for thermal infrared satellite data resolution enhancement

Elena Zaitseva, Mykola Lubskyi, Ján Rabčan

***Abstract* —** proposed technique for satellite thermal infrared imagery spatial resolution enhancement involves frequency domain imagery processing using Fast Fourier Transform. This paper consider approach of frequency components separation, that are appropriate for existing spatial resolution enhancement technique, which require the pair of images of the same area with subpixel shift. This method allow enhance detalization and informativity of existing long-term data of Landsat legacy satellite data, that exists since 1984 and will be useful for data received by future satellites, like Landsat-9.

***Keywords* —** fast Fourier transform, frequency domain processing, fuzzy filtering, longwave infrared imagery, remote sensing satellite data, resolution enhancement.

I. INTRODUCTION

The Landsat program, as the product of collaboration between United States Geological Survey (USGS) and National Aeronautics and Space Administration (NASA) has provided accurate measurements of Earth's land cover since 1972. Since 1984, with launch of Landsat-4 satellite, this program started providing longwave infrared data in the wavelength range 10.3-12.5 μm that corresponds to low-energy heat radiance. Along all this period Landsat satellites remains the only satellite program, which promptly provides longwave infrared imagery with moderate spatial resolution (< 100 m). This data help to resolve many tasks, which consider land or subterranean heat radiance: volcanoes activity [1], sea surface temperature changes [2], definition of surface CO_2 expiration [3], urban heat island research [4], etc. Due to rapid urbanization process problem of urbanized area heat emission became extremely especially important. In addition, urbanization with cities population increasing is also closely related to industrial areas expansion. All these factors leads to expansion of surfaces with capability of heat accumulation: asphalts, concrete, roofing rubber, metal roofs etc. which emits heat backwards to atmosphere during night period, keeping urbanized area's air overheated. This heat-transferring regime affects population health pernicious.

Longwave infrared data from Landsat satellites became the basis for surface temperature mapping. Many researches of the urbanized area heat islands using Landsat-8 data are already conducted [5, 6]. Landsat data also provides the possibility of long-term time series processing for regression analysis and microclimate condition prediction [7]. However, relatively low spatial resolution of raw Landsat longwave infrared data remains the problem, which do not allow detailed mapping of territories with high landscape and surface heterogeneous, like cities.

Different approaches for resulting temperature images spatial resolution enhancement technique are exist. Fast Fourier Transform (FFT) allows expanding of imagery processing possibilities. It transforms images from spatial domain into frequency domain and represent it like set of frequency components. It permits frequency filtration for many purposes: noise attenuation [8], windowed filtration [9], etc. FFT proved to be very flexible and efficient technique, and it is allow separate frequency component separation for purposes of spatial resolution enhancement.

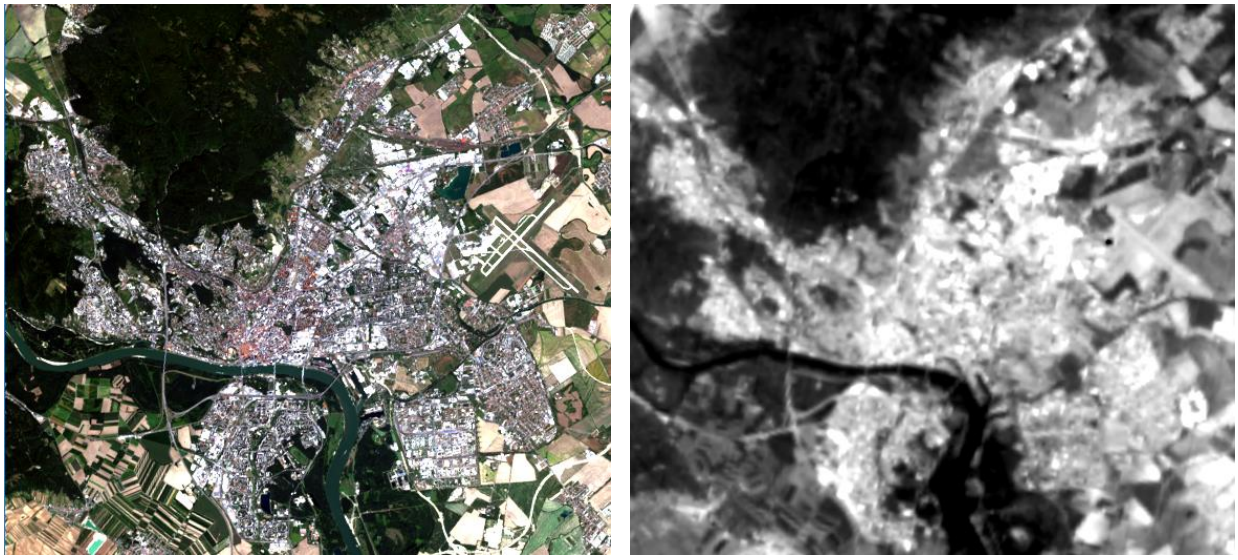
E. Zaitseva, Faculty management science and informatics, University of Zilina, Zilina, Slovakia, (email: elena.zaitseva@fri.uniza.sk).
M. Lubskyi, National Academy of Sciences of Ukraine, Kiev, Ukraine, (email: N.Lubsky@nas.gov.ua)
J. Rabcan, Faculty management science and informatics, University of Zilina, Zilina, Slovakia, (email: jan.rabcan@fri.uniza.sk)

II. Dataset and Methods

A. *Landsat data receiving*

Landsat data also as many other types of remote sensing data are provided by USGS web-service EarthExplorer (https://earthexplorer.usgs.gov/). All presented data is in free access. The EarthExplorer user interface is an online search, discovery, and ordering tool, with powerful filtering tools and area search interface, based on GoogleMaps. Landsat archive contains images in all spectral bands, including longwave infrared data, in GeoTIFF format.

For this research a pair of Landsat-8 imagery stack were acquired. Study area – capital of Slovakia Bratislava. Dates of imagery acquisition 12 August, 2018 and 28 August, 2018. As the typical city Bratislava contain a lot of different landscapes and surfaces: parks, roads, metal and rubber roofs with physical dimensions much more lover than spatial resolution not only TIRS sensor but also OLI. This makes superresolution task especially important for urbanized area imagery, that contains surface temperature. Fig. 1 demonstrates raw Landsat-8 data of the data from 12 August, 2018.



a *b*

Fig. 1 Landsat-8 imagery of Bratislava city (12 August, 2018): *a)* visible data, provided by OLI sensor (combination of the 2, 3 and 4 band), *b)* longwave infrared data provided by TIRS sensor (10 band)

B. *Temperature and emissivity estimation*

Temperature distribution image is the main result of the long-wavelength infrared data processing of the radiance data, which is derived with taking into account the influence of atmospheric effects. For precise determination of the land surface radiance it is necessary to eliminate the influence of the atmosphere caused by the presence of atmospheric trace gases, water vapor that absorb, reflect and scatter infrared radiation [10]:

$$L_\lambda = \varepsilon_\lambda \tau_\lambda \int L_{b\lambda} S_\lambda + L_\lambda^{\uparrow} + \left(1-\varepsilon_\lambda\right)\tau_\lambda L_\lambda^{\downarrow}, \quad (1)$$

where L_λ – spectral radiance in spectral range λ; ε_λ – spectral emissivity in spectral range λ; τ_λ – spectral atmospheric transmittance; $L_{b\lambda}$ – spectral emittance of the blackbody; S_λ – sensor's normalized spectral response $c_1 = 2hc^2 = 1{,}191 \cdot 10^{-16}$ W·m^2 and $c_2 = 1{,}439 \cdot 10^{-2}$ m·K – first and second Planck's constant; λ – radiance wavelength; $L_\lambda^{\uparrow}$ and $L_\lambda^{\downarrow}$ – upwelling and downwelling irradiance.

Planck's law is the equation for temperature determination:

$$T = \frac{c_2}{\lambda \ln\left(\frac{\varepsilon_\lambda c_1}{\lambda^5 L_s} + 1\right)} \quad (2)$$

where Ls – spectral radiance from the Earth's surface; $c_1 = 2hc^2 = 1{,}191 \cdot 10\text{-}16$ W·m2 and $c_2 = 1{,}439 \cdot 10^{-2}$ m·K – first and second Planck's constant.

The emissivity can be estimated on the basis of the visible and near infrared (NIR) data processing. Determination of the Earth's surfaces emissivity distribution using remote sensing data is performed by processing images of the visible and near-infrared range, in particular by establishing of the relationship between emissivity the normalized difference vegetation index (NDVI) distribution [11]. The emissivity is a rather inert surface feature, and for its determination it is possible to involve data obtained with some time interval in comparison with the data of the long-wavelength range. The determination emissivity and the NDVI index relationship for the surfaces covered with vegetation and bare soil is established separately from other types of surfaces, including artificial ones.

This relationship for nonvegetation covers is estimated on the basis of the regressive dependence between artificial surfaces spectra taken from ASTER Spectral Library (http://speclib.jpl.nasa.gov) and NDVI index beyond its range that corresponds to the vegetation cover. Derived quasi-optimal spline-approximation of the dependence is performed through the obtained averaged point (Fig. 2).

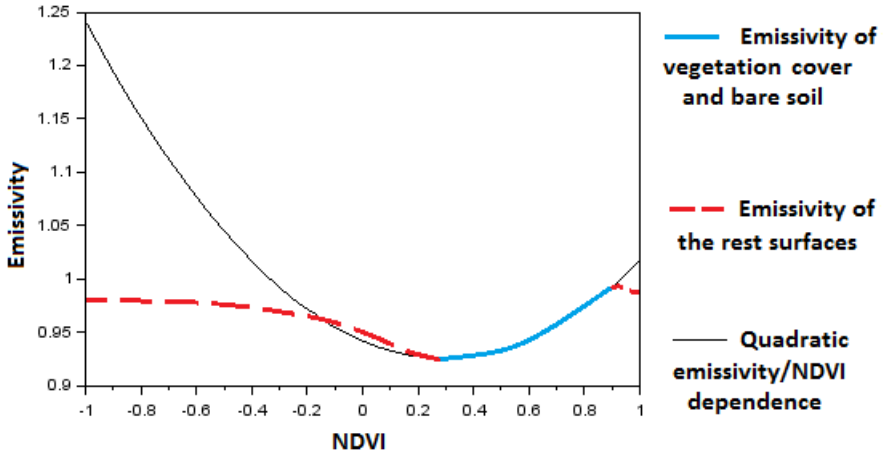


Fig. 2 Distribution of the relationship between Earth's surfaces emissivity and NDVI vegetation index

This approach of emissivity estimation allows obtaining sufficiently detailed data on the thermal characteristics of the presented landscapes and significantly improving the informativity of the resulting surface temperature distribution relatively to the raw Landsat longwave infrared data (Fig. 3).

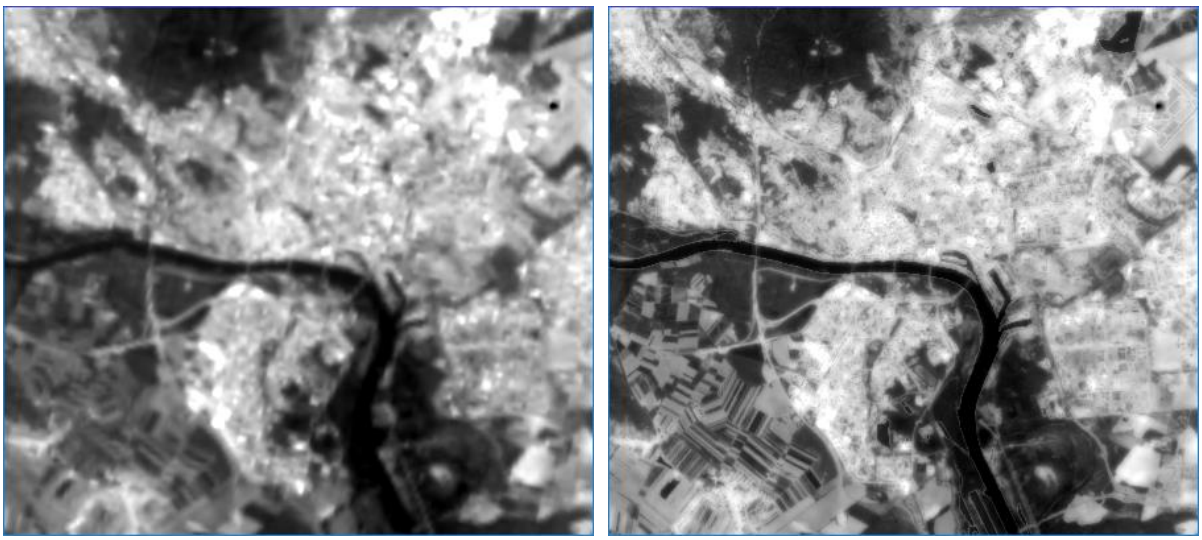


Fig. 3 Visual comparison between detalization of raw longwave data from Landsat-8 (left) and processed surface temperature image (right)

Detailed temperature end emissivity estimation procedure presented in [12].

C. Basics of subpixel spatial resolution enhancement technique

Subpixel superreosution technique [13] is based on specific processing of minimum two images of the same territory with subpixel shift, with distribution of the same feature and radiometrically equivalent. Subpixel offsets are stochastic linear deviations of pixel nets per pixel share that occur when rescannig the same area. In case of using temperature data, those images must be taken in the same time, otherwise temperature difference would make distortions in resulting enhanced image. This technique extracts features from both of images, what makes spatial resolution enhancement possible.

Obtaining an image of enhanced spatial resolution from a pair of images of low spatial resolution is achieved by realization of the next actions:

- subpixel shift estimation;
- estimation of the joint noise image as the difference between input images;
- merging of low-resolution input images into a common resampled image by interlaced scan into a grid of high resolution with replacing two pixels diagonally with pixels of input images, taking into account the subpixel offset and noise matrix;
- estimation of the inverse operator matrix estimation, for the rest of the pixels restoration;
- enhanced image restoration;
- iterative image reconstruction for irregularities and suppress noise elimination.

The corresponding software has been developed to perform the procedure for restoring the high definition image [13].

Emissivity for a period of Landsat revisit time (up to 16 days) used to be an appropriate data for sub-pixel processing. However, the proposed improvement for this technique allow adopting it for temperature images processing, despite the significant temperature difference. Frequency domain processing permit extracting particular frequency components, which can be utilized instead of emissivity images.

D. Fuzzy imagery filtering in frequency domain

FFT allows transferring data from spatial or time domain into frequency domain. Amplitude spectrum as the main output data for gives detailed information about frequency components, contained in image and its' direction across the image (Fig. 4).

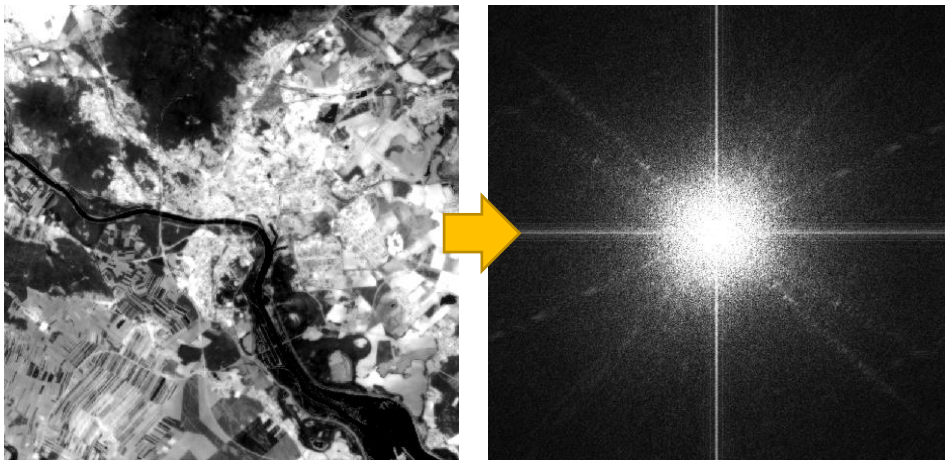
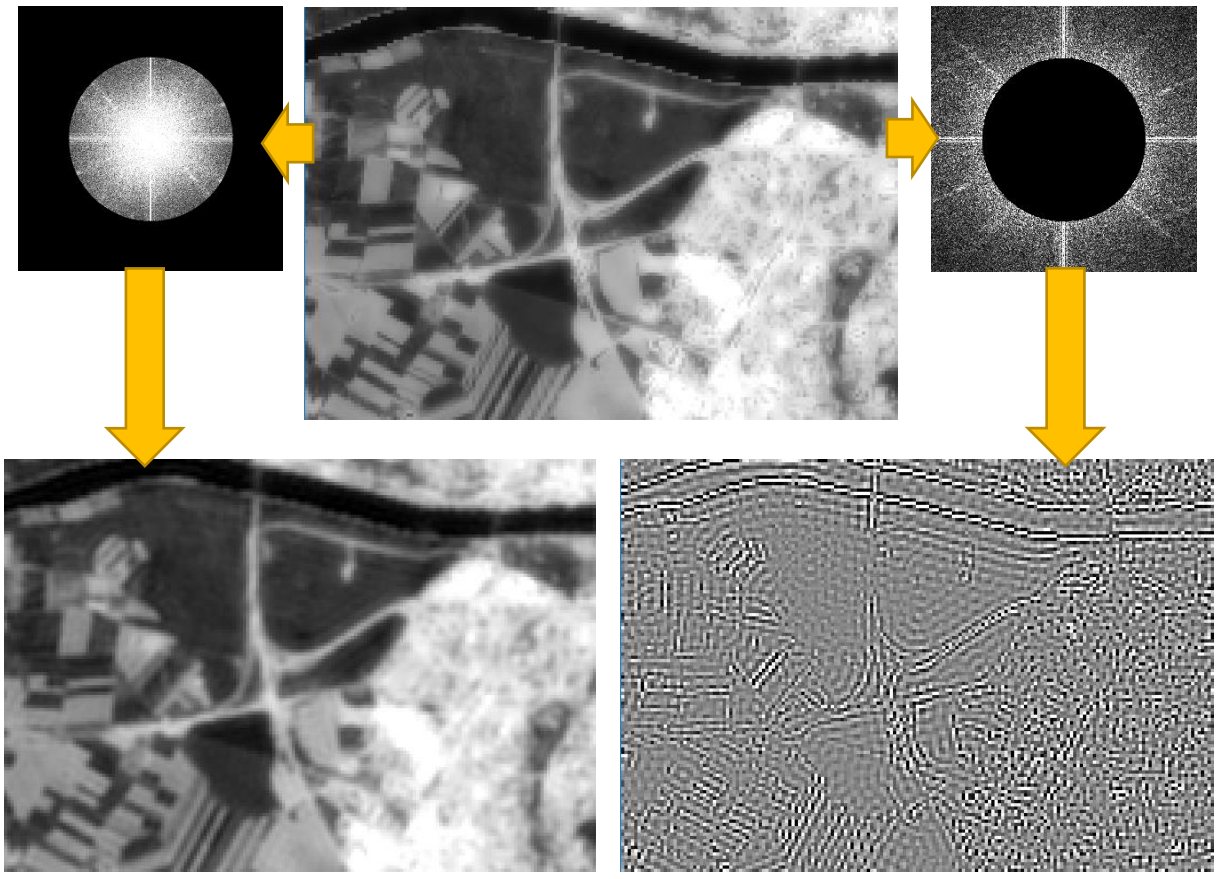


Fig. 4 Temperature image and it's Fourier Amplitude spectrum

Surface temperature imagery, retrieved using Landsat-8 longwave infrared data in frequency domain consist from different spatial frequency components. Longwave components are concentrated in central part of Amplitude spectrum and corresponds to extensive homogenous fields, like lakes, agriculture fields, forests, rivers, etc. Raise of the spatial frequency and its transition closer to borders of amplitude spectrum corresponds to decrement of the distance between surfaces with different temperature. The highest frequencies corresponds to edges between different surfaces represented by low frequency component.



Low frequency component High frequency component

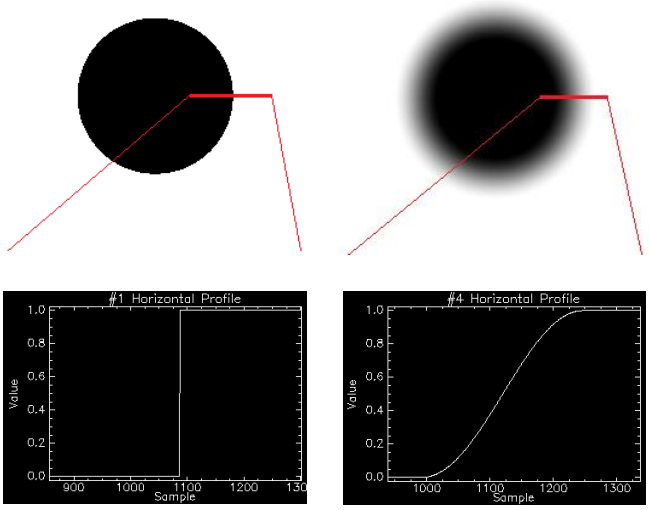
Fig. 5 Extraction of low and high frequency components

Using the suggestion, that contrast edges between different landscape objects remains unchangeable for a long period, imagery pairs of high frequency components distribution becomes the appropriate data for superresolution technique, because its equivalency despite time difference. After estimation of high frequency components distribution data, it merges with low-frequency components, and in result we have enhanced surface temperature data. Each amplitude spectrum value is represented as absolute FFT roots value:

$$A(u,v) = \sqrt{= R_{u,v}^2 + I_{u,v}^2} \tag{3}$$

where $A(u,v)$ – amplitude value; $R_{u,v}^2$ – real part of FFT root; $I_{u,v}^2$ – imaginary part of FFT root. As real matrix can not be presented in complex values, Fourier spectrum must be decomposed into matrices of real and imaginary values.

Next step is building of Fourier spectrum filter, which separates low and high frequency components. For both imaginary and real part of spectrum for both of images high-frequency components are need to be extracted for superresolution performance. For this purpose a fuzzy relationship matrices have been built, which allow smooth separation of low and high frequencies. This approach allow selecting the most appropriate and effective separation method.



Threshold low-pass filter Fuzzy low-pass filter

Fig. 6 Threshold and fuzzy filters

Each relationship matrices must meet next requirement:

$$M_{Lo} + M_{Hi} = M_1 \tag{4}$$

where M_{Lo}, and M_{Hi} are low and high frequency relationship matrices respectively, M_1 is all-ones matrix.

Each spectrum multiplies by fuzzy relation matrix with data in the range 0..1, that represents fuzzy relation degree. As the maximum frequency represented by frequency $\frac{1}{2}N$ m^{-1}, where N – images' spatial resolution, the frequency, with 0.5 relation degree to each of components represented by frequency $\frac{1}{4}N$ m^{-1}. Separated high-frequency component were utilized for spatial resolution enhancement technique, and then extracted from resulting images Fourier

spectra were merged with low-frequency components of each of the images. Inversed Fast Fourier Transform performs transition from frequency domain into spatial domain.

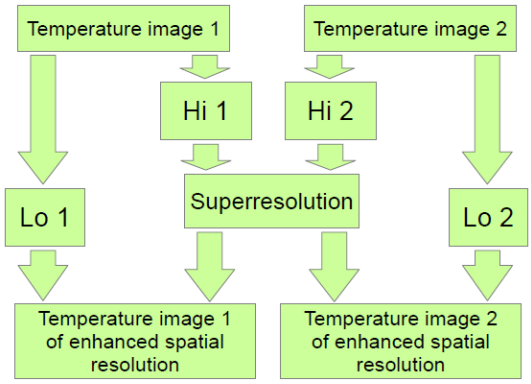


Fig. 7 General flowchart of frequency component merging for imagery spatial resolution enhancement

III. Results and Discussion

Estimation of the spatial resolution enhancement conducted using modulation transfer function (MTF) of input images and resulting image and its comparison. This technique also help to compare the efficiency of both approaches: threshold separation and fuzzy separation. MTF analysis demonstrated that threshold technique gives nearly 74 % of spatial resolution enhancement [14], relatively to input images. Fuzzy approach 12 % appeared to be more effective, than threshold, which utilizes binary separation logic. Next figure demonstrates visual difference temperature images before and after superresolution processing.

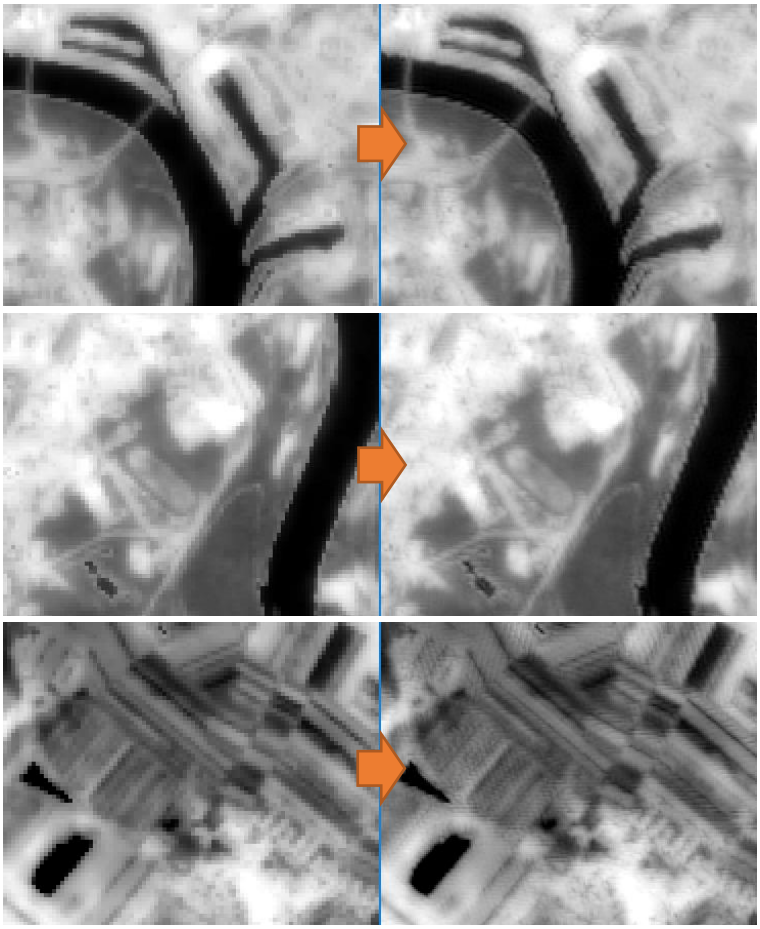


Fig. 8 Comparison between images before (left) and after (right) superresolution processing

According to existing factsheet (https://www.usgs.gov/land-resources/nli/landsat/landsat-9), next Landsat satellite, Landsat-9, would have longwave infrared sensor (TIRS-2) with the same spatial resolution as TIRS, mounted on Landsat-8, and data informativity will remains the same. Proposed improved technique for thermal imagery spatial resolution enhancement is an effective way of processing of available variety of collected Landsat data, starting from 1984. This amount of big data is an appropriate information for detailed temperature imagery long-term time series processing. Results of this processing will help in long-term urbanized areas development and expanding analysis, its effect on public health and environment: ecosystems, natural landscapes and water bodies. In addition, it will be useful for further development planning and expanding.

ACKNOWLEDGMENT

This research was supported Faculty of Management Science and Informatics of University of Žilina and by the National Scholarship Programme of the Slovak Republic for the support of mobility of students, PhD students, university teachers, researchers and artists and.

REFERENCES

[1] C. González, M. Inostroza, F. Aguilera R. González, J. Viramonte and A. Menzies, "Heat and mass flux measurements using Landsat images from the 2000–2004 period, Lascar volcano, northern Chile," *Journal of Volcanology and Geothermal Research*, vol. 301, Aug. 2015, pp. 277–292.

[2] H. Ding and A. J. Elmore, "Spatio-temporal patterns in water surface temperature from Landsat time series data in the Chesapeake Bay, U.S.A.," *Remote Sensing of Environment*, vol. 168, Oct. 2015, pp 335–348.

[3] R. A. Crabbe, D. Janouš, E. Dařenová and M. Pavelka, "Exploring the potential of LANDSAT-8 for estimation of forest soil CO_2 efflux," *International Journal of Applied Earth Observation and Geoinformation*, vol. 77, May 2019, pp 42–52.

[4] L. Sheng, X. Tang, H. You, Q. Gu and H. Hu, "Comparison of the urban heat island intensity quantified by using air temperature and Landsat land surface temperature in Hangzhou, China," *Ecological Indicators*, vol. 72, Jan. 2017, pp. 738–746.

[5] G. Grigorașa and B. Urițescu, "Land Use/Land Cover changes dynamics and their effects on Surface Urban Heat Island in Bucharest, Romania," *International Journal of Applied Earth Observation and Geoinformation*, vol. 80, Aug. 2019, pp. 115–126.

[6] Roy, M.A. Wulder, T.R. Loveland, et al., "Landsat-8: science and product vision for terrestrial global change research," *Remote Sensing of Environment*, vol. 145, Apr. 2014, pp. 154–172.

[7] V. I. Gornyy, V. I. Lyalko, S. G. Kritsuk, I. Sh. Latypov, A. A. Tronin, et al., "Forecast of Saint-Petersburg and Kiev thermal replies on climate change (on the basis of EOS and Landsat satellite imagery)," *Current problems in remote sensing of the Earth from space*, vol. 2(13), 2016 pp. 176–191.

[8] J. Zhou, W. Lu, J. He, B. Liu and T. Ren, "A data-dependent Fourier filter based on image segmentation for random seismic noise attenuation," *Journal of Applied Geophysics*, vol. 114, Mar. 2015, pp. 224–231.

[9] R. Zhao, X. Li and P. Sun, "An improved windowed Fourier transform filter algorithm," *Optics & Laser Technology*, vol. 74, Nov. 2015, pp. 103–107.

[10] X. Yu, X. Guo and Z. Wu, "Land surface temperature retrieval from Landsat-8 TIRS - comparison between radiative transfer equation-based method, split window algorithm and single channel method," *Remote Sensing*, 2014, vol. 6(10), pp. 9829–9852.

[11] J. A. Sobrino, J. C. Jiménez-Muñoz and L. Paolini, "Land surface temperature retrieval from LANDSAT TM 5," Remote Sensing of Environment, vol. 90, 2004, pp. 434–440.

[12] I. Piestova, M. Lubskyi, M. Svideniuk, S. Golubov and P. Sedlacek "Satellite Imagery Resolution Enhancement for Urban Area Thermal Micromapping," *Central European Researchers Journal*, 2018, vol. 4(1), pp. 35–39.

[13] S. A. Stankevich, S. V. Shklyar and M. S. Lubskyi, "Thermal infrared aerial imagery spatial resolution enhancement using sub-pixel registration," *Proceedings of the State Scientific-Research Institute of Aviation*, 2013, vol. 9(16), pp. 110–117.

[14] S. A. Stankevich, M. S. Lubskyi and A. Forgac, "Thermal infrared satellite imagery resolution enhancement with fuzzy logic bandpass filtering," in *Proc. of the International Conference on Information and Digital Technologies 2019*, Žilina, 2019, pp. 446–450.